SCHOLARLY PUBLICATIONS

A CURRENT AWARENESS BULLETIN
OF RESEARCH OUTPUT

@DTU

(97th Edition)

JANUARY 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 **Ninety Seventh** 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 **January, 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 Contemporary Survey of Multimodal Presentation Attack Detection Techniques: Challenges and Opportunities, ***6.Kavita***, Gurjit Singh Walia and ***3.Rajesh Rohilla***, Electronics, DTU

2. A Contemporary Survey of Unimodal Liveness Detection Techniques: Challenges & Opportunities, ***6.Kavita***, Gurjit Singh Walia and ***3.Rajesh Rohilla,*** Electronics, DTU

3. A hitchhiker's guide to single-cell transcriptomics and data analysis pipelines, ***6.Richa Nayak*** and ***3.Yasha Hasija***, Biotechnology, DTU

4. A Hybrid Grey Wolf-Whale Optimization Algorithm for Optimizing SVM in Breast Cancer Diagnosis, ***3.Indu Singh***, Raghav Bansal, ***8.Anchal Gupta*** and ***8.Ayush Singh***, Computer, DTU

5. A Literature Review On Sentiment Analysis Techniques Involving Social Media Platforms, ***8.Samarth Garg***, ***8.Divyansh Singh Panwar***, ***8.Aakansha Gupta*** and ***3.Rahul Katarya***, Computer, DTU

6. A New Approach to Design and Optimize Sizing of Hybrid Microgrid in Deregulated Electricity Environment, ***6.Astitva Kumar***, ***3.M. Rizwan*** and ***3.Uma Nangia***, Electrical, DTU

7. A New Reconstruction of Numerical Fluxes for Conservation Laws using Fuzzy Operators, ***chab*** and ***3.Vivek Kumar***, Applied Mathematics, DTU

8. A Review of AI for Urban Planning: Towards Building Sustainable Smart Cities, ***8.Avinash Kumar Jha***, ***8.Awishkar Ghimire***, ***8.Surendrabikram Thapa***, ***8.Aryan Mani Jha*** and ***3.Ritu Raj***, Civil, Computer and Mechanical, DTU

9. A Theoretical Comparison between CH3NH3PbI3 and CH3NH3SnI3 based Solar Cells, Gagandeep, ***3.Mukhtiyar Singh***, Ramesh Kumar, Vinamrita Singh and Sunita Srivastava, Physics, DTU

10. A unified approach for detection of Clickbait videos on YouTube using cognitive evidences, ***6.Deepika Varshney*** and ***3.Dinesh Kumar Vishwakarma***, IT, DTU

11. Advancements in RNA Secondary Structure Prediction using Machine Learning Methods, ***7.Shubham Mittal*** and ***3.Yasha Hasija***, Biotechnology, DTU

12. Affective state recognition from hand gestures and facial expressions using Grassmann manifolds, ***3.Bindu Verma*** and Ayesha Choudhary, DTU.

13. Alignment-Free Cancelable Biometric: A Contemporary Survey, Opportunities & Challenges, ***6.Pratima Sharma***, Gurjit Singh Walia and ***3.Rajesh Rohilla***, ECE, DTU

14. All dielectric metasurface based tunable optical modulator: Design and analysis, Keshav Samrat Modi, Satya Pratap Singh, Jasleen Kaur, Umesh Tiwari and ***3.Ravindra Kumar Sinha***, Physics, DTU

15. An experimental assessment of bio-mixed-phase change material in a lightweight compartment and its potential as latent energy storage in north-east climate of India, Shubham Kumar Jaiswal and ***6.1.Devesh Kumar***, Mechanical, DTU

16. Analog and RF Performance Evaluation of Junctionless Accumulation Mode (JAM) Gate Stack Gate All Around (GS-GAA) FinFET, ***6.Bhavya Kumar*** and ***3.Rishu Chaujar***, Physics, DTU

17. Application of Industry 4.0 technologies for effective coordination in humanitarian supply chains: a strategic approach, ***3.Pravin Kumar*** and Rajesh Kr Singh, Mechanical, DTU

18. Band gap and gate metal engineering of novel heteromaterial InAs/GaAs-based JLTFET for improved wireless applications, ***6.Samriti Sharma*** and ***3.Rishu Chaujar***, Applied Physics, DTU

19. Big data analytics application for sustainable manufacturing operations: analysis of strategic factors, ***3.Narender Kumar***, ***3.Girish Kumar*** and Rajesh Kumar Singh, Mechanical, DTU

20. Cascade-IλDμN controller design for AGC of thermal and hydrothermal power systems integrated with renewable energy sources, Yogendra Arya, Nishant Kumar, ***3.Pankaj Dahiya***, Gulshan Sharma, Emre Çelik, Sandeep Dhundhara and Mandeep Sharma, Electronics, DTU

21. COMPAC: Compressed Time-Domain, Pooling-Aware Convolution CNN Engine With Reduced Data Movement for Energy-Efficient AI Computing, ***8.1.Aseem Sayal***, Shirin Fathima, SS Teja Nibhanupudi and Jaydeep P. Kulkarni, Electronics, DTU

22. Context- and Sequence-Aware Convolutional Recurrent Encoder for Neural Machine Translation, ***8.1.Ritam Mallick***, ***3.Seba Susan***, ***8.Vaibhaw Agrawal***, ***8.1.Rizul Garg*** and ***8.1.Prateek Rawal***, IT, DTU

23. Cyclone Frame Prediction by Gaussian Modeling of Most Recent Spatio-Temporal Data, ***3.Seba Susan***, ***8.1.Achin Saxena***, ***8.1.Anuvart Badhwar***, ***8.1.Akshay Takhi*** and ***8.1.Abhishek Varshney***, Computer , DTU

24. Design and Analysis of a Three-Phase Grid- Connected Solar PV Inverter with Proportional- Integral Controller, ***7.Kemoh Lumei***, Electrical, DTU

25. Designing a Compact Microstrip Antenna Using the Machine Learning Approach, Kanhaiya Sharma and ***7.1.Ganga Prasad Pandey***, Electronics, DTU

26. Detection of Lung Cancer using Machine Learning Techniques Based on Routine Blood Indices, ***7.Puneet*** and ***3.Anamika Chauhan***, IT, DTU

27. Dog Breed Classifier for Facial Recognition using Convolutional Neural Networks, ***8.Bickey Kumar shah***, ***8.Aman Kumar*** and ***8.Amrit Kumar***, CSE, DTU

28. Dual Metal Gate Dielectric Engineered Dopant Segregated Schottky Barrier MOSFET With Reduction in Ambipolar Current, ***3.Sumit Kale*** and Madduri Sai Chandu, Electronics, DTU

29. Effect of SNF-Based Superplasticizer on Physical, Mechanical and Thermal Properties of the Geopolymer Concrete, ***6.Manvendra Verma*** and ***3.Nirendra Dev***, Civil, DTU

30. Electronic Voting Machine Using a Microcontroller, ***8.Payal Yadav*** and ***8.Ansh Gupta***, Electrical, DTU

31. Emergence of Carbon Nanodots as a Probe for Super-Resolution Microscopy, Navneet C. Verma, Aditya Yadav, Chethana Rao, ***7.1.Pushpendra M. Mishra*** and Chayan K. Nandi, Biotechnology, DTU

32. Estimation and Experimental Validation of Mean-Field Homogenised Effective Properties of Composite, ***6.A. Sharma***, ***3.V. Rastogi*** and ***3.A.K. Agrawal***, Mechanical, Production, Industrial and Automobile, DTU

33. Evaluation of water quality and potential metal contamination in ecologically important Bhindawas bird sanctuary, India, ***6.Vandana Shan***, ***3.S. K. Singh*** and ***3.A. K. Haritash***, Environmental, DTU

34. Exploring Semi-Supervised Learning for Predicting Listener Backchannels, Vidit Jain, ***8.Maitree Leekha***, Rajiv Ratn Shah and Jainendra Shukla, CSE, DTU

35. Fabrication of GaN Nano-Towers based Self-Powered UV Photodetector, ***Lalit Goswami***, Neha Aggarwal, Pargam Vashishtha, Shubhendra K. Jain, Shruti Nirantar, Jahangeer Ahmed, M. A. Majeed Khan, ***3.Rajeshwari Pandey*** and Govind Gupta, Electronics, DTU

36. FLUORIDE REMOVAL FROM DRINKING WATER BY PRISTINE PUMICE AS ADSORBENT, ***3.S. K. Singh***, ***8.S. Shweta***, ***8.Mansi Shrivastava*** and ***8.Shivankar Aggarwal***, Environmental, DTU

37. Footer Voltage Controlled Dual Keeper Domino Logic with Static Switching Approach, ***8.Chirag PARASHAR***, *Avijeet Kumar TRIVEDI*, ***8.Aman AGARWAL*** and ***3.Neeta PANDEY***, Electronics, DTU

38. Framework for agricultural e-trading platform adoption using neural networks, ***6.Sanjay Chaudhary*** and ***3.P. K. Suri***, DSM, DTU

39. Geotechnical parameter assessment of sediment deposit: A case study in Pasakha, Bhutan, Karma Tempa, ***6.Nimesh Chettri***, ***3.Raju Sarkar***, Sunil Saha, Lily Gurung, Tshering Dendup and Bhawani Shankar Nirola, Civil, DTU

40. hGWO-PS optimized dissimilar fractional order 2DOF-PID (FO-2DOF-PID) controllers for performance analysis of interconnected thermal power system, Shailaja Yogesh Kanawade and ***6.Vikas Soni***, DTU

41. How COVID-19 Brought Indian Wedding Industry to a Halt?, ***8.1.Ravi Kumar Arya***, Ritik Ranjan Gupta, Jatin Kumar, Tanay Shubham and ***8.1.Ravi Dugh***, ECE, DTU

42. Improvement in photovoltaic response of bismuth ferrite by tuning its ferroelectric and bandgap properties, Arti, Reema Gupta, ***3.Renuka Bokolia*** and Vivek Verma, Applied Physics, DTU

43. In silico analysis of plumbagin against cyclin-dependent kinases receptor, Arpita Roy and ***8.Karanpreet Singh Bhatia***, Biotechnology, DTU

44. Increasing Accuracy of Software Defect Prediction using 1-dimensional CNN with SVM, ***7.Hitendra Singh Yadav***, CSE, DTU

45. Influence of cross-linkers on the properties of cotton grafted poly (acrylamide-co-acrylic acid) hydrogel composite: swelling and drug release kinetics, ***6.Preeti Gupta*** and ***3.Roli Purwar***, Chemistry, DTU

46. Intrusion Detection System a Need, ***7.Raj Kishore*** and ***3.Anamika Chauhan***, IT, DTU

47. Is Ex-situ Crop Residue Management a Scalable Solution to Stubble Burning?, L. S. Kurinji and ***8.Sankalp Kumar***, Environmental, DTU

48. Leveraging crowd knowledge to curate documentation for agile software industry using deep learning and expert ranking, ***3.Akshi Kumar***, CSE, DTU

49. Low Fidelity VTOL UAV Design Optimization Using an Open Source Framework, ***7.Nikhil Sethi***, ***7.Saurav Ahlawat*** and ***3.Prof. N. S. Raghava***, Mechanical, Physics and ECE, DTU

50. Lyapunov Stability Theory based Non Linear Controller Design for a Standalone PV System, ***6.Aman Sharma*** and ***3.Narendra Kumar***, Electrical, DTU

51. Mechanical and wear performance of Al/SiC surface composite prepared through friction stir processing, ***3.Ravi Butola***, ***8.Lakshay Tyagi***, ***3.Ranganath M Singari***, ***3.Qasim Murtaza***, Harish Kumar and Dhiraj Nayak, Mechanical, DTU

52. MPPT Based Grid Connected Photovoltaic System Using Flyback Converter, ***6.MD. Modassir Masoom***, ***3.DR. Narendra Kumar*** & ***3.Ashish Rajeshwar Kulkarni***, Electrical, DTU

53. Multilevel Inverter Based Power Quality Enhancement Using Improved Immune Control Algorithm, ***6.Praveen Bansal*** and ***3.Alka Singh***, Electrical, DTU

54. Numerical Simulation of n-TiO2/p-CIGS Solar Cell, Gagandeep, ***3.Mukhtiyar Singh***, Ramesh Kumar and Sonia Kumari, Physics, DTU

55. Numerical Simulation of Single and Double Slope Solar Still for Different Performance Affecting Variables, ***3.Rohit Kumar*** and ***8.Rahul Kumar Meena***, Mechanical, DTU

56. On A Generalized Briot-Bouquet Differential Subordination, ***3.S. Sivaprasad Kumar*** And ***6.Priyanka Goel***, Mathematics, DTU

57. Optimum Parametric Analysis Based On Thermodynamic Modeling Of A Compression Absorption Cascade Refrigeration System, ***3.Naushad Ahmad Ansari***, ***3.Akhilesh Arora***, ***3.Samsher*** and ***3.Manjunath K***, Mechanical, DTU

58. PLC and SCADA based Industrial Automated System, ***7.Basant Tomar*** and ***3.Narendra Kumar***, Electrical, DTU

59. Polyacrylonitrile/clay nanofibrous nanocomposites for efficient adsorption of Cr (VI) ions, ***6.Priya Bansal*** and ***3.Roli Purwar***, Chemistry, DTU

60. Quadrature Sinusoidal Oscillators Using CDBAs: New Realizations, ***3.Ram Bhagat***, ***3.D. R. Bhaskar*** and ***3.Pragati Kumar***, Electrical, DTU

61. Rapid mapping of landslides in the Western Ghats (India) triggered by 2018 extreme monsoon rainfall using a deep learning approach, Sansar Raj Meena, Omid Ghorbanzadeh, Cees J. van Westen, Thimmaiah Gudiyangada Nachappa, Thomas Blaschke, Ramesh P. Singh and ***3.Raju Sarkar***, Civil, DTU

62. Recent Development in Green Fuel Vehicles and their Future Advancements, ***7.Amit Sachdeva*** and Dadaso D Mohite, DSM, DTU

63. Robust and Adaptive Traffic Signal Control for Unstructured Driving Scenarios in the Developing World, ***8.1.Ishan Maheshwari***, Ujwal Padam Tewari and Chetan Arora, CSE, DTU

64. Sensorless control of PMSM using Direct and Model References Adaptive system estimation techniques and observing the effect of using PI Fuzzy logic controller in MRAS Model, Mohamed ***7.Abdallah Anwar Said***, ***3.Mini Sreejeth***, and ***3.A.B.K.Bhattacharya***, DTU

65. Sentiment Analysis of Emoticon Based Neuro Fuzzy System, ***3.Ritu Agarwal*** and ***7.Sourabh Shreshth***, IT, DTU

66. Single Dual-X Current Conveyor based BASK/BPSK Modulators, ***3.Deva Nand***, Electronics, DTU

67. Teleportation Criteria Based on Maximum Eigenvalue of the Shared d ⊗d DimensionalMixed State: Beyond Singlet Fraction, ***7.Anuma Garg*** and ***3.Satyabrata Adhikari***, DTU

68. Universal Biquad Fractional order Filter Using Single LT1228 IC, ***7.Gopal Singh*** and ***3.Garima***, Electrical, DTU

69. Unveiling contemporary dimensions in the internationalisation of family firms through bibliometric analysis and thematic analysis, ***3.Sonal Thukral*** and ***6.Apoorva Jain***, DSM, DTU

70. VDDDA Based Multifunction Voltage-Mode MISO Filter, ***7.Dhruv Goel*** and ***3.Deva Nand***, Electronics, DTU

71. VDTA based Fractional Order Universal Filter, ***7.Jyoti Srivastava***, Electrical, DTU

1. *Chancellor*

2. *Pro Vice Chancellor*

2.1. *Ex Pro Vice Chancellor*

3. *Faculty*

3.1. *Ex Faculty*

4. *Teaching-cum-Research Fellow*

4.1. *Asst. Librarian*

5. *Alumni*

6. *Research Scholar*

6.1. *Ex Research Scholar*

7. *PG Scholar*

7.1. *Ex PG Scholar*

8. *Undergraduate Student*

8.1. *Ex Undergraduate Student*

SN Computer Science (2021) 2:49
https://doi.org/10.1007/s42979-020-00425-3

SURVEY ARTICLE

A Contemporary Survey of Multimodal Presentation Attack Detection Techniques: Challenges and Opportunities

Kavita[1] · Gurjit Singh Walia[2] · Rajesh Rohilla[1]

Received: 22 November 2020 / Accepted: 10 December 2020
© The Author(s), under exclusive licence to Springer Nature Singapore Pte Ltd. part of Springer Nature 2021

Abstract
Biometric recognition is a broad and dynamic field of research but the main problem of this field is spoofing attack or presentation attack "use of fake biometric in place of the real biometric sample from original user". Liveness detection is the prime countermeasure to spoofing attacks, which is based on the principle that some additional information can be obtained to verify that the produced data is genuine or not by a standard verification system. It utilized anatomical signs of life, such as facial expression, blinking of eyes, movement of the head, etc. This paper presents a comprehensive review of various liveness detection techniques based on a multimodal biometric system, in which physiological and behavioral properties are used to differentiate between genuine and fake biometric traits. Multimodal systems utilize two or more biometric traits which makes them more secure as compared to unimodal systems. These systems overcome the limitations of the unimodal system such as spoof attack, noisy data, non-universality, distinctiveness, and intra-class variations, etc. Hence to make the biometric systems more secure and robust, multimodal techniques are used. In this paper, we categorized and discuss the various multimodal biometric techniques proposed by various researchers in the last decade, and a new classification is also developed for the same. This paper covers theories, methodology, evaluation datasets, and aims at future work in this field of research.

Keywords Spoofing-attacks · Liveness detection · Unimodal biometric · Multimodal biometric

Introduction

Biometrics is taken from the Greek word in which "bio" refers to the "life" and "metric" refers to the "measure", which tries to identify or to verify individuals based on physiological characteristics (e.g., face, fingerprint, iris) or behavioral characteristics (e.g., signature, voice, gait).

Presently, the biometric recognition system is widely used in many security systems but presentation attack or spoofing attack "use of fake biometric in place of the real biometric sample from original user" is still the main issue for these systems. However various methods exist to find out that the person available in front of the biometric sensor is an artifact or live but the problem of spoofing attacks is remained [1]. There are mainly two types of attacks that take place in a biometric recognition system, one is, direct attack and another indirect attack. Direct attacks take place at the biometric sensor level (Fig. 1, Point1) and these attacks are made on the biometric system by representing the human characteristics or an artifact in front of the biometric capture subsystem. The attacks, which are performed inside the system by invaders, are known as indirect attacks. These attacks are made by escaping the feature extractor (Fig. 1, point 3) or matcher (Fig. 1, point 5), in the database by manipulating templates (Fig. 1, point 6), and in the communication channels by exploiting possible weak points (Fig. 1, points 2, 4, 7, and 8) [2]. The figure shows the possible attack point's in the general biometric system.

This article is part of the topical collection "Advances in Computational Approaches for Artificial Intelligence, Image Processing, IoT and Cloud Applications" guest edited by Bhanu Prakash K N and M. Shivakumar.

✉ Kavita
kavbit.207@gmail.com

Gurjit Singh Walia
gurjit.walia@gmail.com

Rajesh Rohilla
rajesh@dce.ac.in

[1] Delhi Technological University, Delhi, India

[2] SAG, DRDO, New Delhi, India

Published online: 20 January 2021

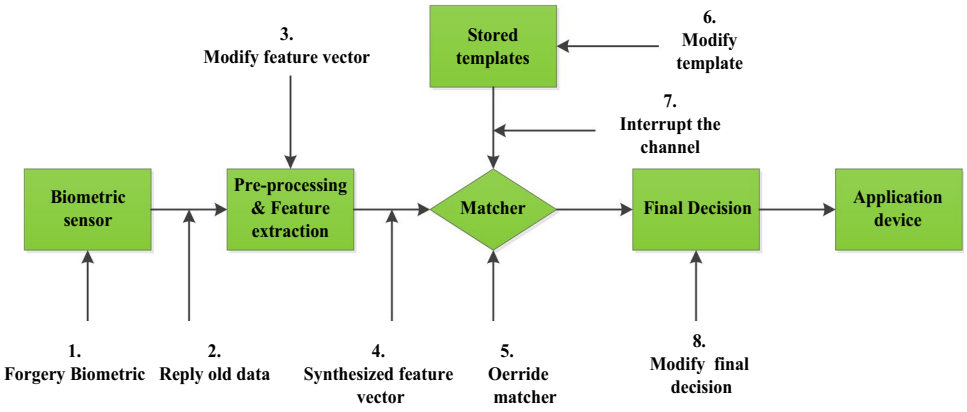


Fig. 1 Attack points in the general biometric system [2]

Liveness detection is also known as Presentation attack detection, is the prime countermeasure to spoofing attacks, which is based on the principle that some additional information can be obtained to verify that the produced data is genuine or not by a standard verification system. Some examples of liveness detection are facial expressions, blinking of eyes, movement of the head, blood pressure, and electrical heart signals (ECG), brain wave (EEG), etc. [3]. In this review, we discuss the various liveness detection techniques based on a multimodal biometric system. In a unimodal system, only a single trait (e.g. face, iris, etc.) is used which can be spoofed easily and therefore it has the following limitations: distinctiveness, spoof attack, noisy data, intra-class variations, non-universality, etc. The multimodal system utilized two or more biometric traits (such as face and iris, face and fingerprint, face, fingerprint, and iris, of a person) which make it more secure and robust as compared to the unimodal system against the spoofing attack. Also, these systems overcome the limitations of the unimodal system.

In this survey paper, various liveness detection techniques based on the multimodal biometric system are discussed, for user authentication. This survey is organized as follows: "Introduction" gives the formal introduction to liveness detection techniques followed by the proposed classification in "Proposed classification of presentation attack detection techniques based on multimodal biometric system", the different methods proposed are presented as: multiple traits in "Multiple traits", multiple sensors in "Multiple sensors", multiple representations in "Multiple representations", behavioral techniques in "Behavioral techniques", and image quality based techniques in "Image quality based techniques". Lastly, concluding remarks are mentioned in "Conclusion".

Proposed Classification of Presentation Attack Detection Techniques Based on Multimodal Biometric System

Nowadays to increase the performance of biometric systems and to make them more secure against spoofing or presentation attacks, multimodal systems are considered better than the unimodal biometric system (Fig. 2).

The main limitations of the unimodal system are spoof attacks, noisy data, distinctiveness, intra-class variations, and non-universality which can be overcome by a multimodal system. Also, apart from the liveness detection, these systems provide better security against spoof attacks as compared to the unimodal system. In [4] a prototype design was proposed for a multimodal biometrics system in which ring fingerprints and left/right index, left/right near-infra-red dorsal hand vein patterns, etc. were taken. The main advantage of adding these modalities was in liveness detection. In [5] two novel approaches were proposed, in which one was the extension of likelihood ratio-based fusion and the second was the use of fuzzy logic against spoofing attacks. In this work impact of spoofing attacks was analyzed on the multimodal biometric systems and it was observed that this scheme was more robust against the spoofing attacks as compared to the likelihood ratio and weighted sum. Jiang et al. [6] have been proposed a video-based multimodal biometric approach that utilized face and speech fusion in the Laplacian subspace for speaker recognition. Fusion using the presented approach achieved better accuracy as compared to the single face or audio modality. Barrero et al. [7] presented the software-based

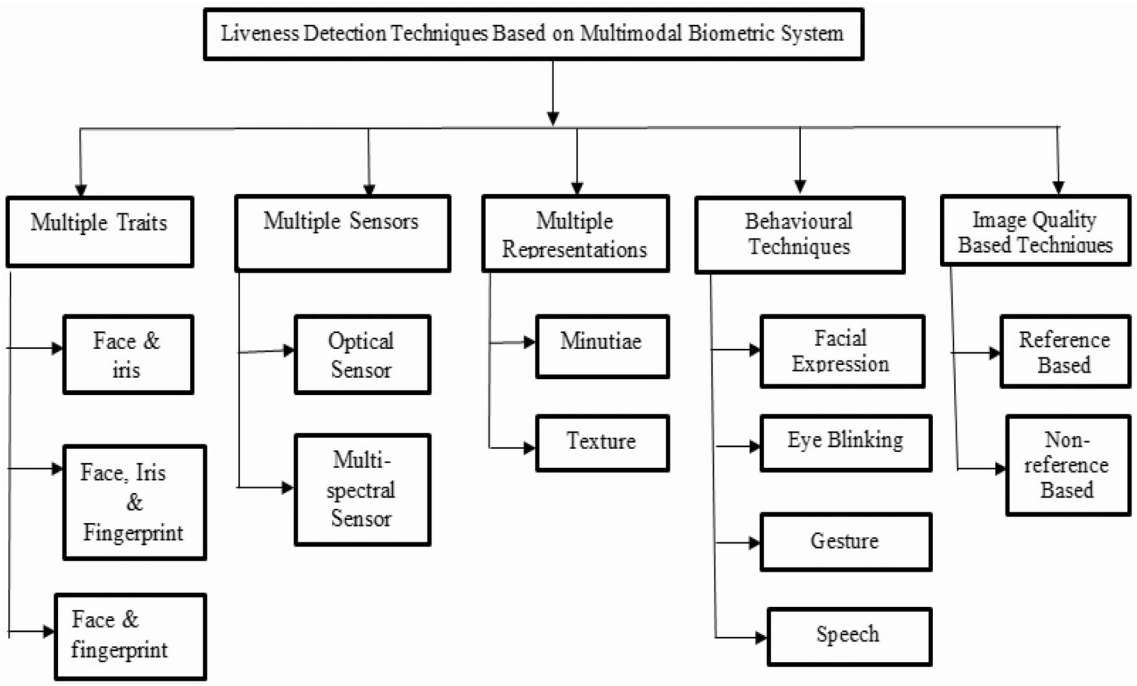


Fig. 2 Proposed classification for presentation attack detection techniques based on the multimodal biometric system

attack against multimodal systems in which face and iris were used to check the performance. Das et al. [8] proposed a structure for software-based liveness detection in multimodal ocular biometrics. The proposed scheme is utilized for direct attack detection. The authors also include class level liveness detection in their work. Kavitha et al. [9] proposed a multimodal biometric framework that utilized feature level fusion to fuse the extracted features and support vector machine (SVM) classifier to detect the fake face.

Multiple Traits

In this method, two or more biometric characteristics such as face and iris, face, fingerprint, and iris, etc. are used to know about liveness. To make the system more robust multiple traits are used, as a single trait can be spoofed easily. In [10] a multimodal system that overcomes the limitations of a unimodal system was proposed. The authors utilized fingerprint, face, and speech for the identification of the user. For the fingerprint and speech verification, they used the minutiae matching algorithm and left to right hidden Markov model respectively. To recognize the face the author's utilized used Eigen face-based method. Komeili et al. [11] proposed a framework that utilized ECG and fingerprint for user authentication as well as liveness detection. Also, a criterion was proposed based on the averaging and correlation that evaluate heartbeat consistency. In [12] a multimodal biometric system was presented based on optimal score level fusion that can differentiate between real and fake subjects. To optimize the performance of the system backtracking search optimization algorithm was used. Chetty and Wagner [13] proposed a multilevel liveness verification framework based on multimodal fusion and feature extraction for a secure face-voice biometric authentication system. In this framework, face information was obtained from speaking faces. Akhtar et al. [14] have been proposed a mobile biometric liveness detection (MoBio LivDet) method for analyzing the structures of the face, fingerprint, and iris images by utilizing decision level fusion and feature descriptors. In [15] optimized score level fusion was presented that utilized a grasshopper optimization algorithm to enhance the performance of the multimodal biometric system. The proposed system shows high performance and reliability, and robustness against the dynamic environment.

Multiple Sensors

In these various sensors are used on one biometric trait to know the liveness of the system. Wei Bao et al. [16] have proposed an optical flow field method to recognize and find out the liveness of the face.The main limitation of this method was the assumption that fake face will be on a plane. Sahidullah et al. [17] proposed a throat microphone

and body-conducted sensor for automatic speaker verification, liveness detection, and enhance security against a replay attack. In [18] an RF sensor-based liveness detection approach was proposed for the capacitive fingerprint system. In [19] a study was presented to assess the liveness detection of 3D cameras. The authors presented a system that utilized Kinect sensor for testing and comparing the effectiveness of depth data in spoof detection. Also, to recognize the faces that appear on the screen Haar Cascade algorithm was used. Wang et al. [20] presented a theoretical model to present the relationship between the oral airflow pressure and the characters in users' speech. The airflow sensor is used to measure the consistency between the actual and estimated pressure signals to determine whether the given command is live or artificial (Table 1).

Multiple Representations

It utilized a representation of minutiae points and image texture to detect the liveness. In [21], the authors presented an approach based on texture analysis for iris recognition in which spatial filters are utilized for differentiating the texture features. In [22] a method based on the extraction of multiple features from multispectral images was presented for iris liveness detection. This method provides high accuracy of classification between fake iris and live iris. Parveen et al. [23] have been proposed a dynamic local ternary pattern that removes the manual threshold setting in local Ternary pattern by utilizing Weber's law. Also, the comparison between the dynamic local ternary pattern and the local ternary pattern was performed. Boulkenafet et al. [24] utilize the color texture analysis approach for face spoofing detection. The authors utilized joint color texture information obtained from the chrominance and luminance channels. The information is obtained from different color spaces by extracting complementary low-level feature descriptors. Agarwal et al. [25] proposed a method for liveness detection in which spatial analysis of the fingerprint pattern and statistical texture features were used to differentiate between fake and real images. For the effectiveness of the method, the authors also used the fusion of fingerprint and texture features of iris. Dempster–Shafer approach was used for the fusion at the decision level.

Behavioral Techniques

In a multimodal system, behavior includes the movement of the face, hand, eyes blinking, voice, etc. and by making the use of these traits liveness is identified. Sun et al. [26] have been proposed a Conditional Random Fields (CRFs) method to detect the liveness of the face for which they utilized blinking of eyes. They also compare the CRFs model with the cascaded Adaboost and Hidden Markov (HM) model, which provides better results as compared to these models. Zhao et al. [27] proposed an approach to recognize the dynamic texture. To deal with the dynamic events, for example, facial expression, a block-based method was developed. Also, the authors have been developed a volume local binary pattern (VLBP) to combine motion and appearance. In [28] a face live detection method that utilized physiological motion was proposed. The authors utilized a conventional active shape structure to recognize the local appearance around each landmark. In [29] a fuzzy fusion approach was presented for liveness detection that was based on mutual dependency models. These models extract the Spatio-temporal correlation between face and voice. Komogortsev et al. [30] proposed an oculomotor plant characteristics (OPC) based liveness detection method which utilized eye movement signals for the identification of people. The live data is taken from 32 individuals. In [31] liveness detection measure was proposed to recognize and find the liveness of the face that was based on the challenge and response method. The author's utilized eye and mouth movement to generate random challenges and observe the response of users. Haar classifier was used to identify eye and mouth movements. Somasundaram et al. [32] have been proposed a Spatio-temporal feature detector based on the sparse representation length of the Spatio-temporal patches measured by residual error. The authors suggested a classification framework based on a bag of features that provide a better result on KTH and UCF sports action datasets. Nagrani et al. [33] proposed a convolution neural network architecture for cross-modal matching between voices and faces, which was used for both binary and multi-way cross-modal face and audio matching. In [34] face liveness detection method was presented, in which deep features are extracted through convolution neural network and to extract the color features rotation invariant local binary patterns were utilized. To differentiate between genuine and fake faces support vector machine was used. Schardosim et al. [35] have been proposed a method that integrates the imaging features with liveness features and used to distinguish between real access and spoofing attacks. These are based on the models learned by an artificial neural network.

Image Quality Based Techniques

Many techniques have been developed to differentiate between fake and genuine face, image quality assessment is one of them. Properties of texture or image quality are used in the image quality-based method. These are categorized into two main parts reference-based and non-reference-based. In the reference-based method, an undistorted

Table 1 Liveness detection techniques used in the multimodal biometric system

References	Features	Attack	Database	Description
Li Ma et al. [21]	Iris	Spoof attack	CASIA Iris	Proposed a method based on texture analysis for iris recognition
Sun et al. [26]	Face	Presentation attack	Self-constructed	Proposed a conditional random fields (CRFs) method
Shahin et al. [4]	Hand vein and Fingerprint	Spoof attack	Self-constructed	Proposed a new prototype design for multimodal biometrics
Rodrigues et al. [5]	Face, Fingerprint	Presentation attack	FVC2004DB1, FERET-b series	Proposed two novel schemes, the extension of the likelihood ratio based fusion and fuzzy logic
Liting et al. [28]	Face	Presentation attack	Self-constructed	Presented a face live detection method which utilized physiological motion
Bao et al. [16]	Face	Photo-attack	Self-constructed	Proposed optical flow field method to recognize the liveness of face
Chetty [29]	Face and voice	Spoof attack	VidTIMIT, DaFeX Corpora	Proposed a new fuzzy fusion technique for liveness detection
Komogortsev et al. [30]	Eye	Spoof attack	Self-constructed	Proposed Oculomotor Plant Characteristics (OPC) based liveness detection method
Barrero et al. [7]	Face, Iris	Indirect Attack	BioSecure	Presented and evaluated the first software-based attack
Singh et al. [31]	Face	Spoofing attack	Self-constructed	Proposed a liveness detection measure to recognize and find the liveness of face, based on challenge and response method
Akhtar et al. [14]	Face, iris fingerprint,	Spoofing attack	ATVS-Flr, ATVS-FFp	Proposed mobile biometric liveness detection (MoBio LivDet) method
Wild et al. [42]	Face and Fingerprint	Direct attack	LivDet 2013, CASIA FASD	Proposed 1-median filtering as a spoofing resistant
Das et al. [8]	sclera and iris	Direct attack	Self-constructed	Proposed a framework for software-based liveness detection
Boulkenafet et al. [24]	Face	Presentation attack	MSU-MFSD,CASIA-FASD	Proposed an approach based on color texture analysis for face spoofing detection
Lee et al. [37]	Finger-vein	Spoof attack	Self-constructed	Proposed a finger-vein biometric recognition system based on image quality assessment
Parveen et al. [23]	Face	Presentation attack	UPM, CASIA FASD, NUAA	A dynamic local ternary pattern has been proposed which utilizes Weber's law
Sollinger et al. [40]	Iris, face fingerprint, and finger vein	Presentation attack	ATVS-Flr, ATVS-FFp, IDIAP	Proposed non-reference image quality measures (IQM) to differentiate between fake and real data
Kavitha et al. [9]	Face	Fake face attack	CASIA FASD, MSU MFSD	Proposed multimodal biometric framework to detect face spoofing
Sahidullah et al. [17]	Voice	Replay attack	Self-constructed	Proposed a body conducted sensor and throat microphone for automatic speaker verification

Table 1 (continued)

References	Features	Attack	Database	Description
Komeili et al. [11]	ECG and fingerprint	Presentation attack	LivDet2015	Proposed a framework in which fingerprint and ECG are combined for user authentication and liveness detection
Schardosim et al. [35]	Face	Presentation attack	NUAA, CASIA	Proposed a method based on the models learned by an artificial neural network
Wang et al. [20]	Voice	Presentation attack	Self-constructed	Proposed a theoretical model to show the relationship between the characters in the user's speech and oral airflow pressure
Agarwal et al. [25]	Fingerprint and iris	Presentation attack	LivDet2011, ATVS-FFp, ATVS Fir	Proposed a liveness detection method which utilized fusion of fingerprint and iris. Dempster–Shafer (D–S) approach is used for the fusion at the decision level

reference image is utilized to estimate the quality of the test image while in the non-reference-based method pre-trained statistical models are utilized to estimate the quality of the test image [36]. In [22] the authors presented a method for iris liveness detection by utilizing multispectral images which can differentiate between real and fake images. Lee et al. [37] have been proposed a finger-vein biometric recognition system based on image quality assessment in which edge detection algorithm and two-dimensional mask-based entropy algorithm were used. In this work, a near-infrared (NIR) image quality assessment module was implemented and the proposed model was a prototype on an embedded field-programmable gate array prototyping (FPGA) platform. Akhtar et al. [38] proposed a face spoof recognition algorithm based on discriminative image patches in which frames were selected randomly from a given face video. The author's utilized seven new methods to obtain discriminative patches from a face image, and then features were selected and fed to a classifier for the final classification of real and spoof faces. Xia et al. [39] proposed a feature extraction method that utilized circularly symmetric.

Gabor feature and weber local binary pattern for liveness detection. Fingerprint images were analyzed in the spatial domain and frequency domain and final features were decided by the co-occurrence probabilities. In [40] non-reference image quality measures were proposed to distinguish between genuine and fake data. In this method, accurate classification of real versus fake iris, fingerprint, face, and finger vein data was achieved. In [41] a sparse representation-based blind image deblurring algorithm was proposed which helps to learn smaller sub-dictionaries from the patches clustered based on dominant orientation instead of learning a single large dictionary.

Conclusion

In this manuscript, we have presented a structured overview of the most relevant work carried out so far in the field of liveness detection. We also proposed a new classification for liveness detection methods based on multimodal biometrics and a review of the existing spoofing attacks. Further, a new classification has been developed for the multimodal biometric system based on liveness detection techniques into the following types namely (1) multiple traits based techniques, (2) multiple sensor-based techniques, (3) multiple representation-based techniques, (4) behavioural techniques, (5) image quality-based techniques. As the review has shown, several approaches have been achieved in the development of countermeasures against spoofing attacks. But, due to the development of new attacking methodologies every day, there is still the need to devote further efforts to the design of new and more efficient liveness detection approaches that may increase the reliability of biometric technology.

Compliance with Ethical Standards

Conflict of interest All authors declare that they have no conflict of interest.

References

1. Hadid A, Evans N, Marcel S, Fierrez J. Biometrics systems under spoofing attack: an evaluation methodology and lessons learned. IEEE Signal Process Mag. 2015;32:20–30.

2. Akhtar Z, Micheloni C, Foresti GL. Biometric liveness detection: challenges and research opportunities. IEEE Secu Priv. 2015;13:63–72.
3. Marcel S, Nixon MS, Li SZ. Handbook of biometric anti-spoofing. London: Springer London; 2014. p. 1–279.
4. Shahin MK, Badawi AM, Rasmy ME. A multimodal hand vein, hand geometry, and fingerprint prototype design for high security biometrics. In: Proceedings of the 2008 Cairo international biomedical engineering conference. IEEE; 2008. p. 1–6.
5. Rodrigues RN, Ling LL, Govindaraju V. Robustness of multimodal biometric fusion methods against spoof attacks. J Vis Lang Comput. 2009;20:169–79.
6. Jiang RM, Sadka AH, Crookes D. Multimodal biometric human recognition for perceptual human–computer interaction. IEEE Trans Syst Man Cybern Part C Appl Rev. 2010;40:676–81.
7. Gomez-Barrero M, Galbally J, Fierrez J. Efficient software attack to multimodal biometric systems and its application to face and iris fusion. Pattern Recognit Lett. 2014;36:243–53.
8. Das A, Pal U, Ferrer MA, Blumenstein M. A framework for liveness detection for direct attacks in the visible spectrum for multimodal ocular biometrics. Pattern Recognit Lett. 2016;82:232–41.
9. Kavitha P, Vijaya K. Optimal feature-level fusion and layered k-support vector machine for spoofing face detection. Multimed Tools Appl. 2018;77:26509–43.
10. Jain AK, Hong L, Kulkarni Y. A multimodal biometric system using fingerprint, face, and speech 1999; 10.
11. Komeili M, Armanfard N, Hatzinakos D. Liveness detection and automatic template updating using fusion of ecg and fingerprint. IEEE Trans Inf Forensics Secur. 2018;13:1810–22.
12. Walia GS, Singh T, Singh K, Verma N. Robust multimodal biometric system based on optimal score level fusion model. Expert Syst Appl. 2019;116:364–76.
13. Chetty G, Wagner M. Multi-level liveness verification for face-voice biometric authentication. In: Proceedings of the 2006 Biometrics Symposium: special session on research at the biometric consortium conference. IEEE; 2006. p. 1–6.
14. Akhtar, Z., Micheloni, C., Piciarelli, C., Foresti, G.L.: MoBio; LivDet: Mobile biometric liveness detection. In: 2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). pp. 187–192. IEEE (2014)
15. Gupta K, Walia GS, Sharma K. Multimodal biometric system using grasshopper optimization. In: 2019 international conference on computing, communication, and intelligent systems (ICCCIS); 2019. IEEE, p. 387–391.
16. Bao W, Li H, Li N, Jiang W. A liveness detection method for face recognition based on optical flow field. In: Proceedings of 2009 international conference image analysing signal process; 2009. IASP 2009, p. 233–236.
17. Sahidullah M, Thomsen DAL, Hautamaki RG, Kinnunen T, Tan ZH, Parts R, Pitkanen M. Robust voice liveness detection and speaker verification using throat microphones. IEEE/ACM Trans Audio Speech Lang Process. 2018;26:44–56.
18. Kim W, Hong W, Kim T, Kim D, Lee M. RF sensor-based liveness detection scheme with loop stability compensation circuit for a capacitive fingerprint system. IEEE Access. 2019;7:152545–51.
19. Albakri G, Alghowinem S. The effectiveness of depth data in liveness face authentication using 3D sensor cameras. Sensors. 2019;19:1928.
20. Wang, Y., Cai, W., Gu, T., Shao, W., Li, Y., Yu, Y.: Secure Your Voice. In: Proceedings of ACM interactive, mobile, wearable ubiquitous technology, vol 3. 2019. p. 1–28.
21. Ma Li, Tan T, Wang Y, Zhang D. Personal identification based on iris texture analysis. IEEE Trans Pattern Anal Mach Intell. 2003;25:1519–33.
22. Chen R, Lin X, Ding T. Liveness detection for iris recognition using multispectral images. Pattern Recognit Lett. 2012;33:1513–9.
23. Parveen S, Ahmad S, Abbas N, Adnan W, Hanafi M, Naeem N. Face liveness detection using dynamic local ternary pattern (DLTP). Computers. 2016;5:10.
24. Boulkenafet Z, Komulainen J, Hadid A. Face spoofing detection using colour texture analysis. IEEE Trans Inf Forensics Secur. 2016;11:1818–30.
25. Agarwal R, Jalal AS, Arya K. A multimodal liveness detection using statistical texture features and spatial analysis. Multimed Tools Appl. 2020;79:13621–13645.
26. Sun L, Pan G, Wu Z, Lao S. Blinking-based live face detection using conditional random fields. In: Advances in Biometrics. vol. 4642 LNCS. Springer: Berlin, Heidelberg; 2007. p. 252–260
27. Zhao G, Pietik M. Patterns with an application to facial expressions. Most. 2007;29:1–14.
28. Wang L, Ding X, Fang C. Face live detection method based on physiological motion analysis. Tsinghua Sci Technol. 2009;14:685–90.
29. Chetty G. Biometric liveness checking using multimodal fuzzy fusion. In: 2010 IEEE world congress computational intelligence; 2010. WCCI, p. 1–8.
30. Komogortsev OV, Karpov A. Liveness detection via oculomotor plant characteristics: attack of mechanical replicas. In: Proceedings of 2013 international conference biology; 2013. ICB.
31. Singh AK, Joshi P, Nandi GC. Face recognition with liveness detection using eye and mouth movement. In: 2014 international conference on signal propagation and computer technology (ICSPCT 2014); 2014. IEEE, p. 592–597.
32. Somasundaram G, Cherian A, Morellas V, Papanikolopoulos N. Action recognition using global spatio-temporal features derived from sparse representations. Comput Vis Image Underst. 2014;123:1–13.
33. Nagrani A, Albanie S, Zisserman A. Seeing voices and hearing faces: cross-modal biometric matching. In: 2018 IEEE/CVF conference on computer vision and pattern recognition; 2018. IEEE, p. 8427–8436.
34. Chen FM, Wen C, Xie K, Wen FQ, Sheng GQ, Tang XG. Face liveness detection: fusing colour texture feature and deep feature. IET Biom. 2019;8:369–77.
35. Schardosim LR, Dos Santos RR, Scharcanski J. Detection of presentation attacks using imaging and liveness attributes. Electron Lett. 2019;55:1226–9.
36. Saad MA, Bovik AC, Charrier C. Blind image quality assessment: a natural scene statistics approach in the DCT domain. IEEE Trans Image Process. 2012;21:3339–52.
37. Lee YH, Khalil-Hani M, Bakhteri R, Nambiar VP. A real-time near infrared image acquisition system based on image quality assessment. J Real Time Image Process. 2017;13:103–20.
38. Akhtar Z, Foresti GL. Face spoof attack recognition using discriminative image patches. J Electr Comput Eng. 2016;2016:1–15.
39. Xia Z, Lv R, Sun X. Rotation-invariant Weber pattern and Gabor feature for fingerprint liveness detection. Multimed Tools Appl. 2018;77:18187–200.
40. Söllinger D, Trung P, Uhl A. Non-reference image quality assessment and natural scene statistics to counter biometric sensor spoofing. IET Biom. 2018;7:314–24.
41. Singh K, Vishwakarma DK, Walia GS. Blind image deblurring via gradient orientation-based clustered coupled sparse dictionaries. Pattern Anal Appl. 2019;22:549–58.
42. Wild P, Radu P, Chen L, Ferryman J. Robust multimodal face and fingerprint fusion in the presence of spoofing attacks. Pattern Recognit. 2016;50:17–25.

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

Proceedings of the Third International Conference on Intelligent Sustainable Systems [ICISS 2020]
IEEE Xplore Part Number: CFP20M19-ART; ISBN: 978-1-7281-7089-3

A Contemporary Survey of Unimodal Liveness Detection Techniques: Challenges & Opportunities

Kavita
Electronics & Communication
Delhi Technological University
Delhi, India
kavbit.207@gmail.com

Gurjit Singh Walia
SAG,
DRDO
Delhi, India
gurjit.walia@gmail.com

Rajesh Rohilla
Electronics & Communication
Delhi Technological University
Delhi, India
rajesh@dce.ac.in

2020 3rd International Conference on Intelligent Sustainable Systems (ICISS) | 978-1-7281-7089-3/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICISS49785.2020.9316059

Abstract- Biometric recognition is an extensive and energizing field of research but spoofing attacks, "a deceiver pretense as an authorized user by reproducing the biometrics of the original user", is still the main problem of this field. Liveness detection is one prime countermeasure to spoofing attacks and is based on the principle that some additional information can be obtained or refined to verify that the produced data is authentic or not by a standard verification system. It utilized the physiological or anatomical signs of life, for example, change in, movement of the mouth, facial expression, blinking of eyes, etc. This paper presents an extensive analysis of the various unimodal biometric system based liveness/vitality detection techniques, in which physiological properties are utilized to discriminate between live and fake biometric traits. Further, these techniques are categorized based on static, dynamic, behavioral, and image quality based methods, recently proposed by various researchers. A new classification is also developed for these liveness detection techniques. The work covers theories, methodology, and evaluation datasets and also aims at future work in this field of research.

Keywords- Spoofing-attacks, Liveness detection, Unimodal biometric

I. INTRODUCTION

At present, biometric recognition system is extensively used in many security systems but presentation attack or spoofing attack is still the main issue for these systems. In a biometric presentation attack, "a deceiver pretense as an authorized user by reproducing the biometrics of the original user". However, various methods exist to find out that the person available in front of the biometric sensor is an artifact or live but the issue of spoofing attack is remained [1]. In the biometric system, two types of attacks are found, direct attack and indirect attack. Direct attacks are made on the biometric system by representing the human characteristics or an artifact in front of the biometric capture subsystem. Direct attacks take place at the biometric sensor level. In [40] Espinoza et al. (2010) presented a study to analyze the vulnerabilities of a fingerprint sensor in which the decision threshold was not applied. The study is not focused on decision outcome, scores are obtained by the comparison between real or fake. In [4] a parameterization that was based on quality-related measures has been proposed for iris liveness detection. This approach required only a single iris image to know whether the eye presented in front of the sensor is real or fake, which is the main advantage of this approach. Attacks performed by the invaders (hackers or cyber-criminals), inside the system are known as Indirect attack. These attacks are made by escaping the feature extractor or matcher, in the database by manipulating templates, and in the communication channels by exploiting possible weak points [2].To provide additional security & privacy [59], hardware or software-based system is utilized by liveness detection. In a software-based system, signal processing algorithms are used to detect the fake biometric traits when the sample is received from a standard sensor. These systems are cheaper & non-invasive as compare to the hardware-based systems. In a hardware-based system, additional hardware is added at the sensor level to find out the particular properties. An example of this may include a fingerprint recognition system in which live fingers are detected via blood pressure. Software-based systems use static and dynamic features while hardware-based systems utilize sensors and features. To resolve the problem of spoofing attacks and make the practical use of biometric applications possible, liveness detection is utilized. It operates in four modes, In the first mode, it utilizes available sensors to recognize the characteristic of attacks in the signal, secondly, to find out the evidence of genuineness it uses dedicated sensors, in the third mode for the interaction of user and system it makes use of challenge-response techniques and lastly, to robust against presentation attacks it uses recognition techniques [2]. Liveness detection, one prime countermeasure to spoofing attacks, based on the principle that some additional information can be obtained or refined to verify that the produced data is authentic or not by a standard verification system. Examples of liveness detection are blinking of eyes, facial expressions, movements of the mouth, blood pressure and electrical heart signals (ECG), brain wave (EEG), perspiration at the fingertips, etc. [3]. In [42] a component-based face coding approach has been utilized to differentiate between genuine and spoof faces. In [43] low complexity technique based on local binary pattern descriptors is presented to detect the printed iris attack. In [44] an algorithm was proposed by Pinto et al. (2015) to detect the video-based spoofing attack. The authors utilize global information for evaluation, which is unvarying to content. In [45] Chingovska & Anjos (2015) presented two solutions generative and discriminative to detect the spoofing attack in which identity information of the user was utilized. In [46] to overcome the problem of face spoofing (photographs or videos) the authors presented a diffusion speed model that uses a single image to know about the liveness of the face. The advantage of this method is that it helps to differentiate between the real and fake biometric traits, it provides better accuracy and robustness. The limitation of this method is that it required user cooperation to detect the liveness and the verification process takes some time. In this survey paper, our goal is to comprehensively analyze and discuss various liveness detection techniques based on the unimodal biometric system for user authentication. This survey is organized as follows: section 1 gives the formal introduction to liveness

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 27,2021 at 10:24:20 UTC from IEEE Xplore. Restrictions apply.

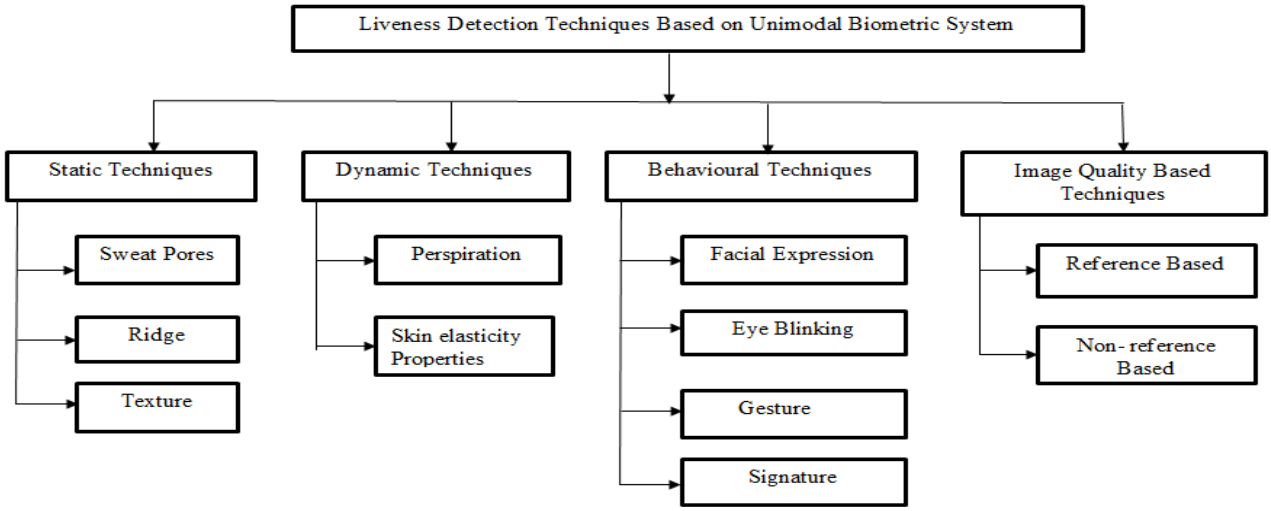


Fig. 1: Proposed Classification of Liveness Detection Techniques Based on Unimodal Biometric System

detection techniques followed by the proposed classification in section 2, different methods proposed are presented as static techniques in section 3, dynamic techniques in section 4, behavioral techniques in section 5, and image quality based techniques in section 6. Lastly, concluding remarks, research gaps, and challenges are mentioned in section 7.

II. Proposed Classification of Liveness Detection Techniques Based on Unimodal Biometric System

Unimodal biometric system uses only single characteristics (face, iris, fingerprint, gait, voice, signature, etc.) for verification. As the unimodal biometric system utilized only single trait, it has some limitations such as spoof attacks, distinctiveness, intra-class variations non-universality, and noisy data. Multiple biometric traits (such as face & iris, face & fingerprint, voice & gait, etc.) can be used to overcome these limitations, and the biometric system which utilized multiple traits known as a multimodal biometric system. Also, to resolve the problem of spoofing attacks various approaches have been proposed by the researchers some of them are discussed in this section. The approaches are divided based on static, dynamic, behavioral, and image quality based techniques. In figure 1 proposed classification for the unimodal biometric system and figure, 2 consecutive transformation of biometric liveness detection method for the unimodal system is summarized. A taxonomical chart- Table 1 which summarizes the various liveness methods of the unimodal biometric system of these four categories is also given.

III. Static Techniques

Static features include ridge and valley texture, sweat pores, perspiration, surface coarseness, texture-based statics, etc. which are discussed below:

A. Sweat pores

In the Pore-based method, pores are detected as an indication of fingerprint liveness. High-pass and correlation filtering is utilized by this method. Correlation filters are used to consider the pore spacing of an individual. For instance, In [4] high-pass and correlation filtering have been utilized to detect the active sweat pores (the small circular forms presented in the fingerprint ridges of an existing finger [36]) and the authors also suggested that this method can be used to find out the liveness. In (2007) Choi et al. [5] utilized multiple statics features to analyze the sweat pores and find out the liveness of the fingerprint.

B. Texture

In the texture analysis, it is assumed that the properties of skin, for example, color of skin and reflectance of an original face will be different from the surface properties of a fake face. In [7] Tan et al. (2010) used Lambertian models to find out the difference between spoofed and live faces. In [8] Peixoto et al. (2011) proposed the extension of [7], to find out the face liveness. The authors utilized the difference of gaussian filter to detect the liveness under bad illumination conditions. In [9] Ghiani et al. (2012) presented a method based on local phase quantization to know about the liveness of fingerprint. In [10] to detect mask attacks a countermeasure technique was proposed in which reflectance characteristics of masks and original faces were analyzed and 94.47% accuracy was achieved. In one study, to find out spoofs by making the use of printed photos, a micro-texture analysis was used and to find out spoofed contact lenses one researcher group evaluated the statistical texture of four different attributes [11]. Pereira1 et al. (2014) [13] proposed extensions of the local binary pattern operator and utilize Spatio-temporal (dynamic texture) to detect the face spoofing. The main aim of the approach was to identify real faces with the help of facial micro-textures. Iris image classification was the first time formulated by Sun et al. (2014) [14] which help to detect the iris liveness. The authors proposed a hierarchical visual codebook method to encode the distinguish and primary texture of iris images. In [15] a software-based method was presented to detect the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 27,2021 at 10:24:20 UTC from IEEE Xplore. Restrictions apply.

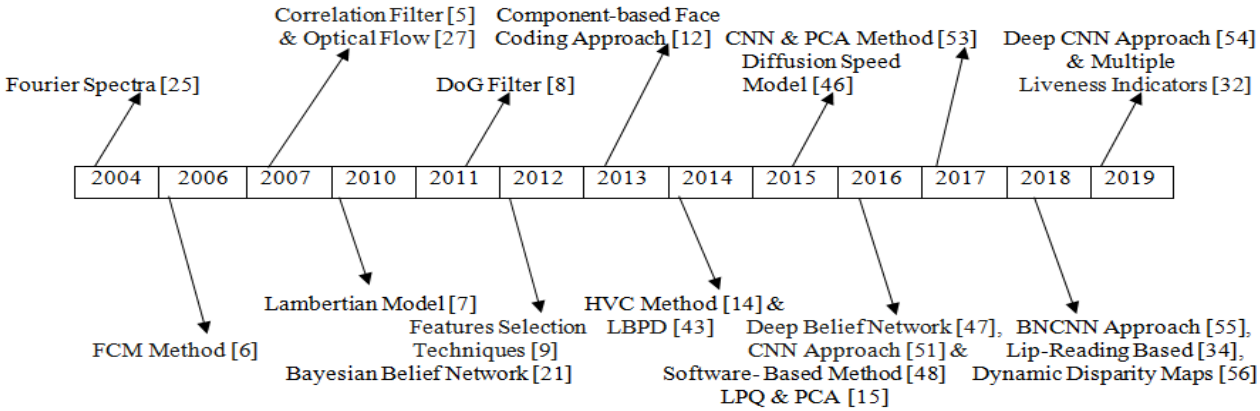


Fig. 2: Consecutive transformation of Biometric liveness detection method for unimodal system that has been covered in this field so far.

liveness of fingerprint. The authors used a multi-scale local phase quantity to construct and dimensionality of the feature vectors was reduced by principal component analysis (PCA). This method utilized only one image to detect the liveness of the fingerprint and provided better performance as compared to the previous techniques.

C. Ridge

In this method ridge of the finger is used to find the liveness, for instance, Abhyankar & Schuckers (2006) [6] proposed a fuzzy c-means clustering (FCM) method to know about the liveness of fingerprint. It was a software-based approach that utilized the density and texture of the fingerprint images. This method used local ridge frequency analysis and multiresolution texture analysis to derive the features. In [47] Kim et al. (2016) proposed a technique based on the deep belief network that provides exact detection of the fingerprint liveness with the various sensors datasets taken from fingerprint liveness detection (Liv Det) competition. In [48] Xia1 et al. (2016) proposed a software-based method to detect the liveness of fingerprint which was based on the two-class classification problem. In [50] Wangwe et al. (2016) proposed a liveness detection technique for dorsal hand vein that was based on principle component analysis and the power spectrum estimation of the autoregressive model. Yuan et al. (2017) [53] proposed a method that utilized convolutional neural network (CNN) and principal component analysis to know about the liveness of fingerprint. The authors achieved good results as compared to the other latest methods in which a support vector machine classifier was used. In [51] Nogueira et al. (2016) proposed convolutional neural networks to detect the liveness of fingerprints. The authors compared four different models CNN-VGG, CCN-random weights, CNN-Alexnet, and classical local binary pattern (LBP) technique.

To conclude this section, static techniques provide better security against spoof attacks and also help to find out the livness of the unimodal systems. In the next section, various dynamic techniques are discussed which also helps to find out the liveness and spoof attacks.

IV. DYNAMIC TECHNIQUES

Dynamic features are derived from the analysis of multiple forms of a similar modality. The dynamic property of a live finger is used to find out the liveness of a person know as perspiration. These techniques are classified based on perspiration and skin elasticity properties.

A. Perspiration

In dynamics analysis, when the illumination level is manipulated to examine pupil dilation change, various images are obtained. The fingerprint's sweating is examined in the perspiration-based method. A time-taken method is proposed by Sujan and his colleagues to find out the spoofed fingerprints by computing the finger's perspiration procedure. A perspiration detection algorithm was proposed to find out the liveness of the fingerprint in which 90% classification rate was achieved and the database contains 33 live & 33 spoof fingerprint images [19]. Marasco & Sansone (2012) [20] proposed a combined set of perspiration-based and morphological measures to know the liveness of fingerprints by utilizing a feature selection method based on a specific sensor. In [21] Marasco et al. (2012) presented a strategy based on the Bayesian belief network that represents the relationship between match scores and liveness values. The authors presented two sequential methods, bayesian belief network and classifier-based framework for the fusion. In [22] Czajka proposed a method based on pupil dynamics for eye liveness detection which utilized linear and nonlinear support vector machines to categorize spontaneous oscillations and natural reaction. For face and texture recognition Hajati et al. (2017) [23] presented a derivative sparse representation (DSR) approach in which they utilize the short-length videos and achieved better results. In [24] Galbally and his colleagues make the use of motion features produced by the sensor's motion.

B. Skin elasticity Properties

Skin elasticity is utilized in the skin deformation–based method. For example, to capture finger distortion thin-plate slats were used [34]. Special training is required in such methods.

In the above section, various dynamic techniques such as perspiration and skin elasticity properties are discussed that provide better results are compared to the static techniques to find out the presentation attacks and liveness. In the next section, various techniques based on the behavior such as facial expression, eye blinking, speech, etc. are discussed.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 27,2021 at 10:24:20 UTC from IEEE Xplore. Restrictions apply.

V. BEHAVIORAL TECHNIQUE

Nowadays behavior analysis plays an important role to find out the liveness. It includes facial expression, gesture eye blinking, and speech, etc., to know about the liveness.

A. Facial Expression

In this method, different facial expressions are used to find the liveness of a person. In [12] a component-based face coding approach was proposed for liveness detection. Gavrillescu et al. (2016) [49] proposed a soft biometric neural network system to recognize the video face. This system examines patterns in independent facial expressions. Jiangwei Li et al. (2004) [25] proposed a technique in which structure and movement of the face have been utilized to detect the liveness of the face and based on the analysis of two-dimensional (2-D) Fourier spectra. In [28] to detect the face liveness monocular camera has been utilized by Pan et al. (2010). The authors combine eyeblink and scene context, inside and outside clues of face against the spoofing (photographs, videos, and 3D models) are also used by them. In [32] face liveness detection algorithm was proposed by Singh & Arora (2018) in which multiple liveness indicators such as blinking of the eye, lips movement, and chin movement are utilized to detect the liveness, which provides higher accuracy. A modified convolution neural network approach that works on the deep features rather than the handcrafted has been proposed by De Souza et al. (2017). In this work, the authors presented two deep neural networks, LBPnet, and n-LBPnet to detect face spoofing which provides better results compared to the state of art method [52]. In [54] Rehman et al. (2018) presented an approach based on a deep CNN classifier that was reliable for the intra-database & cross-database, for face liveness detection problem when the training data is limited. In [56] Rehman et al. (2019) proposed an approach for face liveness detection, based on a stereo camera in which dynamic disparity-maps were used. These dynamic disparity maps are learned from convolutional-features with input stereo face images to direct a convolution neural network. Also, the various forms of disparity-maps are calculated.

B. Eye Blinking

In this method movement of eyes is used to know about the liveness of a person. In the motion analysis method to find out that the presented video in front of the system is fake or real, the extant of spontaneous movement clues are analyzed. For instance, In [11] to detect photo-based spoofing, the eye blink-based liveness detection method was proposed by one research group. Other researchers have used the optical flow field for face liveness detection [31]. In [26] Pan et al. (2007) proposed an approach conditional random field (CRF) framework against photo attack and they utilized the blinking of eyes, as a liveness indicator in the face recognition process. Kollreider et al. (2007) [27] utilized the optical flow of lines technique based on the optical flow method to differentiate between the motion of lines and points. To find out the liveness score optical flow estimation and face part detection were used. In [29] Komogortsev (2015) performed liveness detection at the feature and match score levels for various existing forms of eye movement which is based on different characteristics of the human visual system. In [55] Long & Zeng (2019) proposed a batch normalized convolutional neural network (BNCNN) for iris liveness detection. In BNCNN there were eighteen layers to detect the real and forge iris. In [60] a capsule network has been proposed for retina recognition that provides better accuracy & classification as compared to convolution neural network.

C. Gesture

It utilized different positions of a person to know the liveness. Czajka et al. (2017) [31] were the first who proposed a solution to automatically recognize the orientation of iris images which can be used to identify the spoofing attacks. In this, the authors compared two approaches on the same data, using feature engineering and learning, for classification convolutional neural network was used. In [33] Edmunds & Caplier (2018) have been proposed a motion-based countermeasure against the video and photo attacks in which natural and unnatural motion cues are used. Face motions was extracted with the help of conditional local neural fields face tracking algorithm.

D. Speech

In this method voice of a person is used to know about the liveness of a person. For instance, Wu et al. (2017) [30] presented the automatic speaker verification spoofing and countermeasures initiative. To find out the user's liveness, a system known as Lipass has been presented in [34], based on lip-reading and utilized acoustic sensing on smartphones to extracts behavioral trait of user's speaking mouths. The system achieved 90.2% & 93.1% accuracy in user identification and spoofing detection respectively

In this section, various techniques based on behavior such as facial expression, eye blinking, gesture, and speech are discussed. These techniques are better as compared to the static and dynamic techniques, as in these techniques liveness factors such as facial expression, eye blinking, speech, etc are used which makes it more robust and secure against spoof attacks, and also helps to know about the liveness.

VI. IMAGE QUALITY BASED TECHNIQUE

Several techniques have been developed to differentiate between fake and genuine face, image quality assessment is one of them which is widely used to detect the different types of attacks. Properties of texture or image quality are used in the image quality-based method. Quality measures based method is proposed by Galbally and his colleagues; it includes continuity, ridgepole strength, and clearness [35]. In [57] Gupta et al. (2019) presented a quality-based adaptive score fusion technique for a multimodal biometric system in which three modalities (face, iris, and fingerprint) were used. The authors evaluated reliability factor to solve the problem of a dynamic environment and spoofing attacks.

A. Reference Based

To evaluate the quality of the test sample, the reference-based method believes in an undistorted reference image. In [38] Singh et al. (2016) proposed two algorithms, edge-based and dominant orientation-based texture histogram equalization to improve the visual quality of the images.
A multi-cues integration framework has been proposed

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 27,2021 at 10:24:20 UTC from IEEE Xplore. Restrictions apply.

TABLE I: Liveness Detection Techniques Used in Unimodal Biometric System

Reference	Feature	Attack	Database	Description
Tan et al. (2010) [7]	Face	Photo-attack	Photograph imposter	Proposed Lambertian model for face liveness detection.
Pan et al. (2010) [28]	Face	Video-replay attack	Photograph imposter	Proposed a scene context matching approach against the spoofing.
Peixoto et al. (2011) [8]	Face	Photo-attack	Yale, NUAA	Proposed, a difference of Gaussian filters to detect the face liveness under bad illumination conditions.
Marasco & Sansone (2012) [20]	Fingerprint	Presentation attack	LivDet2009	Proposed a feature selection method for liveness detection.
Marasco et al. (2012) [21]	Fingerprint	Spoof-attack	LivDet2009	Proposed a strategy based on a Bayesian belief network.
Yang et al. (2013) [12]	Face	Spoof-attack	NUAA, CASIA	A component-based face coding approach is proposed.
Pereira et al. (2014) [13]	Face	Presentation attack	CASIA-FASD, Replay Attack	Proposed extensions of local binary pattern operator to detect the face spoofing.
Sun et al. (2014) [14]	Iris	Iris spoof attack	ND, Clarkson and CASIA iris images	Proposed Hierarchical Visual Codebook method which is the integration of the bag-of-words model.
Gragnaniello et al. (2014) [43]	Iris	Printed-iris attack	MobBIO multimodal and MICHE	Proposed a low complexity technique to detect the printed iris attack.
Komogortsev (2015)[29]	Eye	Presentation attack	EMBD	Proposed a method based on different characteristics of the human visual system for iris liveness detection
Pinto et al. (2015) [44]	Face	Face spoofing	Unicamp video-attack, Replay-attack	Proposed an algorithm to detect the video-based spoofing attack.
Chingovska & Anjos (2015) [45]	Face	Presentation attack	Replay-Attack	Proposed two solutions, generative and discriminative to detect spoofing attacks.
Wen et al. (2015) [37]	Face	Face spoofing attack	MSU-MFSD, Replay-Attack, CASIA-FASD,	Proposed an algorithm based on image distortion analysis.
Kim et al. (2015) [46]	Face	Presentation attack	NUAA, SFL, and Replay-attack	Proposed a diffusion speed model for face liveness detection.
Xia et al .(2016) [48]	Fingerprint	Presentation attack	LivDet2009, LivDet2011	Proposed a software-based method, based on the two-class classification problem.
Feng et al. (2016) [39]	Face	Face spoofing attack	3D-MAD, REPLAY-ATTACK, CASIA-FASD	Proposed a multi-cues integration scheme for face anti-spoofing.
Gavrillescu et al. (2016) [49]	Face	Presentation attack	Honda/UCSD video, Youtube faces.	Proposed a system for video-based face recognition known as a soft biometric neural network.
Nogueira et al. (2016)[51]	Fingerprint	Presentation attack	LivDet 2009, LivDet 2011 & LivDet 2013	Proposed a Convolutional Neural Networks for liveness detection.
Yuan et al. (2016) [15]	Fingerprint	Presentation attack	LivDet 2011	Proposed a software-based fingerprint liveness detection method.
Souza et al. (2017) [52]	Face	Face spoofing attack	NUAA	Proposed a modified convolution neural network approach that works on the deep features rather than the handcrafted.
Wu et al . (2017) [30]	Speech	Presentation attack	ASV spoof 2015 database	Presented the ASV Spoofing and Countermeasures (ASV spoof) initiative.
Czajka et al. (2017) [31]	Iris	Presentation attack	Self-constructed	Proposed a solution to automatically recognize the correct orientation of iris images.
Yuan et al. (2017) [53]	Fingerprint	Fingerprint spoof attack	LivDet (2013), LivDet (2011)	Proposed a method based on convolutional neural network and principal component analysis.
Edmunds & Caplier (2018) [33]	Face	Photo attack	3DMAD, MSU-MFSD, CASIA-FASD	Proposed a motion-based countermeasure against the video and photo spoofing attacks
Singh & Arora (2018) [32]	Face	Presentation attack	In-House, ZJU eyeblink, Print attack	Proposed an algorithm with multiple liveness indicators for face liveness detection.
Sharma & Dey (2018) [16]	Fingerprint	Presentation attack	LivDet 2009 , 2011, 2013, and 2015	Presented a software-based approach for liveness detection.
Schardosim et al. (2019) [18]	Face	Presentation attack	CASIA, NUAA	Proposed a method based on observation of imaging and liveness attributes.
Li Lu et al. (2019) [34]	Lip	Presentation attack	Self-constructed	Proposed a lip-reading-based authentication system known as Lipass to find out the liveness of the user.
Long & Zeng (2019) [55]	Iris	Presentation attack	Ndcontact, CASIA Iris Syn.	Proposed a batch normalized convolutional neural network.
Yuan et al. (2019) [17]	Fingerprint	Presentation attack	LivDet2011, LivDet 2013	Proposed deep convolution neural network with image scale equalization for fingerprint liveness detection.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 27,2021 at 10:24:20 UTC from IEEE Xplore. Restrictions apply.

by Feng et al. (2016) [39] for face anti-spoofing, that utilized a hierarchical neural network. To detect the liveness motion cues and image quality cues were fused. In this, the liveness feature based on the image quality is developed by utilizing Shearlet and dense optical flow has been utilized to extract motion-based liveness features. Yuan et al. (2019) [17] proposed a deep convolution neural network with image scale equalization to detect the liveness of fingerprint and the confusion matrix was used as a performance indicator. This method was used to maintain image resolution and preserve texture information. For liveness detection, Schardosim et al. (2019) [18] proposed a method that belief in the observation of imaging attributes (facial & background textures, and steganalysis features) and liveness attributes (face deformation energy).

B. Non- reference Based

Non-reference based method utilized some pre-trained statistical models to evaluate the quality of the test image. In [37] wen et al. (2015) was presented a technique based on image distortion analysis to detect face spoofing. The authors used chromatic moment, specular reflection, blurriness, and color diversity to form the IDA feature vector. To differentiate between genuine and fake faces support vector machine classifier was used. Sharma & Dey (2018) [16] presented a software-based approach in which quality features are used to know about the liveness of fingerprint images. To discriminate between fake and real fingerprints support vector machine boosted tree classifiers and the random forest was used.

To conclude this section, image quality assessment-based techniques are widely nowadays and in these techniques, it is assumed that the quality of real & fake samples will be different, which helps to know about liveness & spoof attacks.

VII. CONCLUSION AND FUTURE DIRECTIONS

As discussed, this paper presents the work that has been carried out concerning the vulnerabilities of the biometric system to direct and indirect attacks. Several techniques have been developed to secure the biometric system against spoofing attacks. Nowadays to overcome the problems of spoofing attacks, liveness detection that aims at detecting the physiological sign of life (facial expression, blinking of eyes, etc.) is mainly used in the biometric system to make them more secure. As a liveness detection technique confirms the availability of a person in front of the biometric system is real or fake, these are widely used. Many liveness detection techniques have been developed in which the unimodal biometric system is presented. The proposed research work has presented an extensive review of discrete liveness detection techniques for the unimodal biometric system proposed by several researchers in the past decades and various dataset used such as CASIA-FASD, 3DMAD, MSU-MFSD, CASIA Iris, LivDet2009, etc. Further, a new classification has been developed for the liveness detection techniques based on the unimodal biometric system into the following types namely (i) Static techniques (ii) Dynamic techniques(iii) Behavioural techniques (iv) Image Quality based techniques.These techniques are very helpful to make the biometric system more secure and widely used in the field of research.

The main limitations of a unimodal biometric system are noisy data, non-universality intra-class variations, etc. which can be overcome by a multimodal biometric system. Nowadays to increase the performance of biometric systems and to make them more secure against the spoofing attack multimodal biometric systems are considered better than the unimodal biometric system. These systems are more robust as compared to unimodal systems and provide better security against spoof attacks. Many techniques have been developed to prevent spoofing attacks and find out the liveness of multimodal biometric systems.

REFERENCES

[1] A. Hadid, N. Evans, S. Marcel, and J. Fierrez, "Biometrics Systems Under Spoofing Attack: An evaluation methodology and lessons learned," IEEE Signal Processing Mag., vol. 32, no. 5, pp. 20–30, 2015.

[2] Z. Akhtar, C. Micheloni, and G. L. Foresti, "Biometric Liveness Detection: Challenges and Research Opportunities," IEEE Security, Privacy, vol. 13, no. 5, pp. 63–72, 2015.

[3] S. Marcel, M. S. Nixon, and S. Z. Li, "Handbook of Biometric Anti-Spoofing-Trusted Biometrics under Spoofing Attacks", 2014.

[4] N. Manivanan, S. Memon, and W. Balachandran, "Automatic detection of active sweat pores of fingerprint using highpass and correlation filtering," Electronics. , Lett. , vol. 46, no. 18, pp. 1268–1269, 2010.

[5] H. Choi, R. Kang, K. Choi, and J. Kim, "Aliveness detection of fingerprints using multiple static features," Proc. World Acad. Sci., pp. 200–205, 2007.

[6] A. Abhyankar and S. Schuckers, "Fingerprint liveness detection using local ridge frequencies and multiresolution texture analysis techniques," Proc. - International Conference Image Processing, ICIP, pp. 321–324, 2006.

[7] X. Tan, Y. Li, J. Liu, and L. Jiang, "Face Liveness Detection from a Single Image with Sparse Low Rank Bilinear Discriminative Model," in *Lecture Notes in Computer Science*, vol. 6316 LNCS, no. PART 6, pp. 504–517, 2010.

[8] B. Peixoto, C. Michelassi, and A. Rocha, "Face liveness detection under bad illumination conditions," in 2011 18th IEEE International Conference on Image Processing, no. 29, pp. 3557–3560, 2011.

[9] L. Ghiani, G. L. Marcialis, and F. Roli, "Fingerprint liveness detection by local phase quantization," Proc. - International Conference Pattern Recognition, no. Icpr, pp. 537–540, 2012.

[10] N. Kose and J. L. Dugelay, "Reflectance analysis based countermeasure technique to detect face mask attacks," 2013 18th International Conference Digital Signal Process. DSP 2013, vol. 1, pp. 0–5, 2013.

[11] M. Hassaballah and S. Aly, "Face recognition: Challenges, achievements and future directions," IET Comput. Vis., vol. 9, no. 4, pp. 614–626, 2015.

[12] J. Yang, Z. Lei, S. Liao, and S. Z. Li, "Face liveness detection with component dependent descriptor," Proc. - 2013 Int. Conf. Biometrics, ICB 2013, 2013.

[13] T. De et al., "Face liveness detection using dynamic texture," EURASIP J. Image Video Process., no. 2, pp. 1–15, 2014.

[14] Z. Sun, H. Zhang, T. Tan, and J. Wang, "Iris image classification based on hierarchical visual codebook," IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 6, pp. 1120–1133, 2014.

[15] C. Yuan, X. Sun, and R. Lv, "Fingerprint liveness detection based on multi-scale LPQ and PCA," China Commun., vol. 13, no. 7, pp. 60–65, 2016.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 27,2021 at 10:24:20 UTC from IEEE Xplore. Restrictions apply.

[16] R. P. Sharma and S. Dey, "Fingerprint liveness detection using local quality features," Vis. Comput., vol. 35, no. 10, pp. 1393–1410, 2019.
[17] C. Yuan, Z. Xia, L. Jiang, Y. Cao, Q. M. J. Wu, and X. Sun, "Fingerprint Liveness Detection Using an Improved CNN with Image Scale Equalization," IEEE Access, vol.7, pp. 26953–26966, 2019.
[18] L. R. Schardosim, R. R. Dos Santos, and J. Scharcanski, "Detection of presentation attacks using imaging and liveness attributes," Electron. Lett., vol. 55, no. 23, pp. 1226–1229, 2019.
[19] S. T. V. Parthasaradhi, R. Derakhshani, L. A. Hornak, and S. A. C. Schuckers, "Time-series detection of perspiration as a liveness test in fingerprint devices," IEEE Trans. Syst. Man Cybern. Part C Appl. Rev., vol. 35, no. 3, pp. 335–343, 2005.
[20] E. Marasco and C. Sansone, "Combining perspiration- and morphology-based static features for fingerprint liveness detection," Pattern Recognit. Lett. , vol. 33, no. 9, pp.1148–1156, 2012.
[21] E. Marasco, Y. Ding, and A. Ross, "Combining match scores with liveness values in a fingerprint verification system," 2012 IEEE 5th Int. Conf. Biometrics Theory, Appl. Syst. BTAS 2012, pp. 418–425, 2012.
[22] A. Czajka, "Pupil dynamics for iris liveness detection," IEEE Trans. Inf. Forensics Secur., vol. 10, no. 4, pp. 726–735, 2015.
[23] F. Hajati, M. Tavakolian, S. Gheisari, Y. Gao, and A. S. Mian, "Dynamic Texture Comparison Using Derivative Sparse Representation: Application to Video-Based Face Recognition," *IEEE Trans. Human-Machine Syst.*, vol. 47, no. 6, pp. 970–982, 2017.
[24] J. Galbally, J. Ortiz-Lopez, J. Fierrez, and J. Ortega-Garcia, "Iris liveness detection based on quality related features," Proc. - 2012 5th IAPR Int. Conf. Biometrics, ICB 2012, no. LivDet 2009, pp. 271–276, 2012.
[25] J. Li, Y. Wang, T. Tan, and A. K. Jain, "Live face detection based on the analysis of Fourier spectra," in Biometric Technology for Human Identification, Aug. 2004, vol. 5404, pp. 296–303.
[26] G. Pan, L. Sun, and Z. Wu, "Eyeblink-based Anti-Spoofing in Face Recognition from a Generic Web camera," 2007.
[27] K. Kollreider, H. Fronthaler, and J. Bigun, "Non-intrusive liveness detection by face images," Image Vis. Comput., vol. 27, no. 3, pp. 233–244, 2009.
[28] G. Pan, L. Sun, Z. Wu, and Y. Wang, "Monocular camera-based face liveness detection by combining eye blink and scene context," Telecommun. Syst., vol. 47, no. 3–4, pp. 215–225, 2011.
[29] O. V. Komogortsev, A. Karpov, and C. D. Holland, "Attack of mechanical replicas: Liveness detection with eye movements," IEEE Trans. Inf. Forensics Secur., vol. 10, no.4, pp. 716–725, 2015.
[30] Z. Wu et al., "ASVspoof: The automatic speaker verification spoofing and countermeasures challenge," IEEE J. Sel. Top. Signal Process, vol. 11, no. 4, pp. 588– 604, 2017.
[31] A. Czajka, K. W. Bowyer, M. Krumdick, and R. G. Vidalmata, "Recognition of Image-Orientation-Based Iris Spoofing," IEEE Trans. Inf. Forensics Secur., vol. 12, no. 9, pp. 2184–2196, 2017.
[32] M. Singh and A. S. Arora, "A Novel Face Liveness Detection Algorithm with Multiple Liveness Indicators," Wirel. Pers. Commun., vol. 100, no. 4, pp. 1677–1687, 2018.
[33] T. Edmunds and A. Caplier, "Motion-based countermeasure against photo and video spoofing attacks in face recognition," J. Vis. Commun. Image Represent. vol. 50, no. October 2017, pp. 314–332, 2018.
[34] L. Lu et al., "Lip Reading-Based User Authentication Through Acoustic Sensing on Smartphones," IEEE/ACM Trans. Netw., vol. 27, no. 1, pp. 447–460, 2019.
[35] J. Galbally, F. Alonso-Fernandez, J. Fierrez, and J. Ortega-Garcia, "A high-performance fingerprint liveness detection method based on quality related features," Futur. Gener. Comput. Syst., vol. 28, no. 1, pp. 311–321, 2012.
[36] C. Sousedik and C. Busch, "Presentation attack detection methods for fingerprint recognition systems: A survey," IET Biometrics, vol. 3, no. 4, pp. 219–233, 2014.
[37] D. Wen, H. Han, and A. K. Jain, "Face spoof detection with image distortion analysis," IEEE Trans. Inf. Forensics Secur., vol. 10, no. 4, pp. 746–761, 2015.
[38] K. Singh, D. K. Vishwakarma, G. S. Walia, and R. Kapoor, "Contrast enhancement via texture region based histogram equalization," *J. Mod. Opt.*, vol. 63, no. 15, pp. 1444–1450, 2016.
[39] L. Feng et al., "Integration of image quality and motion cues for face anti-spoofing: A neural network approach," J. Vis. Commun. Image Represent. , vol. 38, pp. 451–460, 2016.
[40] M. Espinoza, C. Champod, and P. Margot, "Vulnerabilities of fingerprint reader to fake fingerprints attacks," Forensic Sci. Int., vol. 204, no. 1–3, pp. 41–49, 2011.
[41] J. Galbally, J. Ortiz-Lopez, J. Fierrez, and J. Ortega-Garcia, "Iris liveness detection based on quality related features," Proc. - 2012 5th IAPR Int. Conf. Biometrics, ICB 2012, no. LivDet 2009, pp. 271–276, 2012.
[42] J. Yang, Z. Lei, S. Liao, and S. Z. Li, "Face liveness detection with component dependent descriptor," Proc. - 2013 Int. Conf. Biometrics, ICB 2013, pp. 1–6, 2013.
[43] D. Gragnaniello, C. Sansone, and L. Verdoliva, "Iris liveness detection for mobile devices based on local descriptors," Pattern Recognit. Lett., vol. 57, pp. 81–87, 2015.
[44] A. Pinto, W. R. Schwartz, H. Pedrini, and A. D. R. Rocha, "Using visual rhythms for detecting video-based facial spoof attacks," IEEE Trans. Inf. Forensics Secur., vol. 10, no. 5, pp. 1025–1038, 2015.
[45] I. Chingovska and A. R. Dos Anjos, "On the Use of Client Identity Information for Face Antispoofing," IEEE Trans. Inf. Forensics Secur., vol. 10, no. 4, pp. 787–796, 2015.
[46] W. Kim, S. Suh, and J. J. Han, "Face liveness detection from a single image via diffusion speed model," IEEE Trans. Image Process., vol. 24, no. 8, pp. 2456–2465, 2015.
[47] S. Kim, B. Park, B. S. Song, and S. Yang, "Deep belief network based statistical feature learning for fingerprint liveness detection," Pattern Recognit. Lett., vol. 77, pp. 58–65, 2016.
[48] Z. Xia, R. Lv, Y. Zhu, P. Ji, H. Sun, and Y. Q. Shi, "Fingerprint liveness detection using gradient-based texture features," Signal, Image Video Process, vol. 11, no. 2, pp. 381–388, 2017.
[49] M. Gavrilescu, "Study on using individual differences in facial expressions for a face recognition system immune to spoofing attacks," IET Biometrics, vol. 5, no. 3, pp. 236–242, 2016.
[50] Y. Wang, D. Zhang, and Q. Qi, "Liveness detection for dorsal hand vein recognition," Pers. Ubiquitous Comput., vol. 20, no. 3, pp. 447–455, 2016.
[51] R. F. Nogueira, R. De Alencar Lotufo, and R. Campos MacHado, "Fingerprint Liveness Detection Using Convolutional Neural Networks," IEEE Trans. Inf. Forensics Secur., vol. 11, no. 6, pp. 1206–1213, 2016.
[52] G. B. de Souza, D. F. da Silva Santos, R. G. Pires, A. N. Marana, and J. P. Papa, "Deep Texture Features for Robust Face Spoofing Detection," *IEEE Trans. Circuits Syst. II Express Briefs*, vol. 64, no. 12, pp. 1397–1401, 2017.
[53] C. Yuan, X. Li, Q. M. J. Wu, J. Li, and X. Sun, "Fingerprint liveness detection from different fingerprint materials using convolutional neural network and principal component analysis," Comput. Mater. Contin, vol. 53, no. 4, pp. 357–371, 2017.
[54] Y. A. U. Rehman, L. M. Po, and M. Liu, "LiveNet: Improving features generalization for face liveness detection using convolution neural networks," Expert Syst. Appl., vol.108, pp. 159–169, 2018.
[55] M. Long and Y. Zeng, "Detecting iris liveness with batch normalized convolutional neural network," Comput. Mater. Contin., vol. 58, no. 2, pp. 493–504, 2019.
[56] Y. A. U. Rehman, L. M. Po, and M. Liu, "SLNet: Stereo face liveness detection via dynamic disparity-maps and convolutional neural network," Expert Syst. Appl., vol.142, p. 113002, 2020.
[57] K. Gupta, G. S. Walia, and K. Sharma, "Quality based adaptive score fusion approach for multimodal biometric system," *Appl. Intell.*, vol. 50, no. 4, pp. 1086–1099, Apr. 2020.
[58] I.J. Jacob, "Capsule network based biometric recognition system," J. Artif. Intell. Capsul. Networks, vol. 1, no. 02, pp.83-94, 2019.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 27,2021 at 10:24:20 UTC from IEEE Xplore. Restrictions apply.

[59] V.Suma, "Security and privacy mechanism using blockchain,"J. Ubiquito. Comput. and Commun. Tech., vol.1, no.01, pp. 45-54, 2019.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 27,2021 at 10:24:20 UTC from IEEE Xplore. Restrictions apply.

Journal Pre-proof



A hitchhiker's guide to single-cell transcriptomics and data analysis pipelines

Richa Nayak, Yasha Hasija

PII: S0888-7543(21)00033-1

DOI: https://doi.org/10.1016/j.ygeno.2021.01.007

Reference: YGENO 9902

To appear in: *Genomics*

Received date: 9 August 2020

Revised date: 30 December 2020

Accepted date: 18 January 2021

Please cite this article as: R. Nayak and Y. Hasija, A hitchhiker's guide to single-cell transcriptomics and data analysis pipelines, *Genomics* (2021), https://doi.org/10.1016/j.ygeno.2021.01.007

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2021 Published by Elsevier.

Journal Pre-proof

A hitchhiker's guide to single-cell transcriptomics and data analysis pipelines

Richa Nayak, Yasha Hasija* yashahasija06@gmail.com

Department of Biotechnology, Delhi Technological University, Delhi, 110042, India

*Corresponding author.

ABSTRACT

Single-cell transcriptomics (SCT) is a tour de force in the era of big omics data that has led to the accumulation of massive cellular transcription data at an astounding resolution of single cells. It provides valuable insights into cells previously unachieved by bulk cell analysis and is proving crucial in uncovering cellular heterogeneity, identifying rare cell populations, distinct cell-lineage trajectories, and mechanisms involved in complex cellular processes. SCT data is highly complex and necessitates advanced statistical and computational methods for analysis. This review provides a comprehensive overview of the steps in a typical SCT workflow, starting from experimental protocol to data analysis, deliberating various pipelines used. We discuss recent trends, challenges, machine learning methods for data analysis, and future prospects. We conclude by listing the multitude of scRNA-seq data applications and how it shall revolutionize our understanding of cellular biology and diseases.

Keywords

Single-cell transcriptomics; Single-cell RNA sequencing; Single-cell data analysis; Computational Approach; Machine Learning

Journal Pre-proof

1. Introduction

As an elementary school textbook would exclaim, cells are the fundamental, structural, and functional unit of all living organisms. Understanding the biology of cells has been at the center of our pursuit to unravel the complexities that make an organism. Cell biology research has undergone a remarkable transformation in recent years with the advent of single-cell multi-omics technology. The genome structure of every cell is essentially the same for any given individual organism; however, the genome's expression pattern determines the physiological fate of the cell. The observed diversity of phenotypes is due to the genotype and the varying expression pattern, abnormalities in which

form the basis of various diseases. Mapping of this unique genotype-phenotype relationship requires transcriptome profiling, and recent progress made in high throughput sequencing technologies has enabled the measurement of transcriptomic information at an unprecedented resolution of single cells[1].

Transcriptome profiling has revealed that, for any given cell, the transcriptome information reveals the activity of merely a subset of genes[2], and each cell type has a unique transcriptomic fingerprint. Earlier, transcriptome profiling was based on the assumption that all cells from any given tissue are homogenous, and bulk population sequencing followed by average expression analysis would provide us with sufficient information to understand gene expression in both standard and abnormal cell states. Increasing evidence suggests that even in similar cells, the gene expression pattern can be heterogeneous[3,4]. Although bulk expression analysis could simultaneously assess gene expression levels and differentiate between abundant known cell types, it could obscure the identification of rare cell types, subtypes and fail to distinguish cell to cell variability[5]. Thus, the understanding of stochastic cellular processes necessitated a more precise transcriptome analysis technique to overcome the averaging phenomenon inherent to bulk analysis. Unabated technological advancements in NGS, molecular biology, cell biology, and bioinformatics has fostered a new wave of profiling single cells at genomics, transcriptomics, proteomics, and epigenomics level[6,7].

Single-cell transcriptomics (SCT) involves profiling the complete set of RNA transcripts of each individual cell for a given population of cells[8]. Transcriptome analysis of single-cell was pioneered two decades ago, in two separate historical experiments, one by Norman Iscove[9], and James Eberwine and group [10,11], that laid the groundwork for single-cell transcriptome analysis based on high throughput sequencing technologies[12]. scRNA-sequencing is a fast-growing and promising technology for SCT[13] and has rendered microarrays and qPCR obsolete.

The volume and complexity of scRNA-seq data make it a paradigm of big data[1], it has opened doors to a multitude of possibilities in biomedical research, but we have only tapped a fraction of the potential of such a large and versatile dataset. scRNA-seq transcriptome profiles have paved the way for identification of rare cell types in complex tissues[14], cell lineage relationships in early development[15], antigen specificity of immune cells[16], inferring cellular trajectory[17], determination of cell fate[18], distinguishing between normal and abnormal cell[19], understanding tumor heterogeneity[20], identifying regulatory signatures in cancer[21], deciphering immune repertoire for infectious diseases[22], elucidating the mechanism for drug resistance, and relapse in cancer treatment[23]. With better analysis methods, we are uncovering more applications.

scRNA-seq data, although highly potent, poses many challenges on various fronts owing to its big-data characteristics such as sophisticated data acquisition techniques, data storage, management, and analysis[24]. A single scRNA-seq experiment generates a larger volume of high-dimensional raw data than bulk sequencing methods as it

retains the information of the stochastic expression of genes for individual cells. In addition, scRNA-seq experimental protocols have more steps compared to bulk sequencing, which gives rise to more technical biases and artifacts.

Experimental techniques for scRNA-seq have mushroomed and improved over time, which has led to the generation of a massive amount of data and an increasing demand for computational techniques for data analysis. That has led to a spike in developing new experimental protocols, algorithms, and tools to analyze the raw data. Several research groups and commercial companies have designed software tools and packages for data preprocessing and downstream analysis. Machine Learning (ML) approaches, preeminent in big data analysis, have been a noteworthy addition to the list of approaches used for the underlying computational challenges of dimensionality reduction, clustering, and differential expression (DE) analysis[25,26]. Furthermore, there are choices between various programming languages like R, Java, MATLAB, C++, and Python. The development of analysis tools is still in its infancy, and current tools have many shortcomings. The challenge to improve the reciprocation between speed and accuracy in analysis remains. Despite the abundance of techniques, it is hard to establish a standard that can be used across disciplines, and it is crucial to make an informed decision while proceeding for analysis as it can have a tremendous impact on the findings. The scRNA-seq analysis tool choice can influence detecting a biological signal comparable to quadrupling the sample size[27].

This review outlines the general workflow involved in single-cell RNA-seq protocols and discusses the popular and promising new computational tools for analysis. It provides a comprehensive account of each step of the analysis, starting from data preprocessing, imputation of dropouts to tools used for pseudotime ordering, and rare cell type identification. It also discusses ML approaches in the analysis steps, wherever applicable. The review concludes with a discussion of applications across fields in biological sciences, remaining challenges, and prospects.

2. Single-Cell RNA Sequencing Technology

A wide range of scRNA-seq protocols has been developed to accommodate the high demand for improved techniques with high throughput. New methods are being developed to counter batch effects and technical noise since it is important to regulate the initial steps to alleviate the computational burden during data analysis.

scRNA-seq technologies currently in use can be divided into four broad classes based on transcript coverage approach[28]: (i) full-length transcript sequencing [example- MATQ-seq[29], SMART-seq2[30], ICELL8[31] SUPeR-seq[32]], (ii) 5’-end transcript sequencing [example- STRT-seq[33,34]], (iii)3’-end transcript sequencing [example- Chromium[35]10X Genomics, Fluidigm C1[36], Drop-seq[37], inDrop[38]]. With full-length transcript sequencing approach, there is an issue of resolution, speed, and sequencing cost. On the other hand, a major drawback of cDNA sequencing prioritizing either 5’ or 3’-end transcripts of the DNA is incapable of examining allele-specific

expression or alternative splice forms. Some methods rely on FACS based sorting, such as MARS-seq, that make them reliant on a larger initial volume[39] and is disadvantageous when the initial volume is low, as in fine-needle aspirates. Another drawback of using FACS is the requirement of antibodies that target specific proteins for sorting; this poses problems while sorting rare cell subtypes[40]. Thus, each protocol has its sets of advantages and disadvantages that determine the "depth" (reads/cell) of a given dataset, and it could ultimately affect the statistical and biological insight[41]. scRNA-seq is not a "one-size-fits-all" technique like the bulk sequencing approach since the depth can vary with the protocol being used, cell types being examined, capture method, sequencing technique, and alignment stringency during library construction[42]. We discuss a typical scRNA-seq experimental workflow and complementary technologies in use.

2.1. *scRNA-seq workflow*

Single-cell isolation

The generation of scRNA-seq data from a tissue sample involves multiple steps. First, the tissue is digested to ensure dissociation, which gives rise to the single-cell suspension from which single-cells are isolated so that each cell's mRNA can be profiled separately. Single-cell isolation technique predominantly used for scRNA-seq are plate-based techniques and microfluidic-based techniques. Plate-based techniques involve capturing or sorting cells on multi-well plates or microfuge tubes, followed by FACS based sorting. Some full-length scRNA-seq techniques like SUPeR-seq, SMART-seq2, MATQ-seq, Cel-seq rely on plate-based techniques. However, there are many limitations, one being fewer cells per assay than droplet-based technologies. Microfluidic technology involves capturing cells in its microfluidic droplet. Microfluidics-based techniques have swiftly gained popularity amongst single-cell isolation techniques as it requires less initial volume, is cost-effective, and aids in massively parallel quantification of single-cell gene expression profiles[43,44]. Microfluidics techniques can be of two types, i.e. (i) continuous-flow microfluidics, like the Fluidigm's C1 Single-Cell Auto Preparation System, (ii) droplet-based microfluidics like InDrop, Drop-seq, and 10X Genomics. Comparison between single-cell isolation techniques is well depicted by Chen X. et al. 2018[45] and Ziegenhain C. et al. 2017[46]. Droplet-based platforms are readily automated and easily optimizable to suit individual experimental needs depending upon the number of cells to be captured and sequenced. Recent studies have an added advantage of not requiring zero inflation over plate-based techniques that need zero inflation for accurate simulation[27].

Another technique for single-cell isolation from solid tissue is Laser Capture Microdissection, LCM-seq, a laser system-aided isolation of cells directly from solid tissue, coupled with in-situ RNA-sequencing techniques[47,48], which conserves spatial information of mRNA expression within the morphology of a tissue. This method enables

isolation of rare cell types even in highly heterogeneous clinical samples, with increased accuracy and understanding of dynamic cellular systems.

In the lookout to overcome these techniques' inefficiencies, nanowell based single-cell isolation like Seq-Well promises cost-effectiveness, throughput, and portability and requires only nanoliter sized initial volume[49]. Some newer techniques eliminate the single-cell isolation step, like in SPLiT-seq (split-pool ligation-based transcriptome sequencing) and sci-RNA-seq (Single-cell Combinatorial Indexing RNA sequencing[50]). SPLiT-seq allows for simplified and low-cost transcriptome profiling compatible with fixed cells or nuclei and offers high-resolution[51].

2.2. *scRNA-seq Library Preparation*

Much like any RNA library preparation, it roughly entails reverse transcription of captured mRNA into first-strand cDNA synthesis, second-strand synthesis, and cDNA amplification followed by sequencing[52]. But it is more challenging as the amount of RNA per cell is low compared to bulk RNA-seq experiments. A thorough analysis of scRNA-seq required the profiling of a large number of represented individual cells, which is a task worthy of Sisyphus and significantly adds to the cost of carrying out sequencing. The use of Unique Molecular Identifiers (UMI) or cellular barcodes somewhat simplified the process. Cell barcodes are primarily designed to be able to distinguish between read transcripts originating from different cells. To fully determine the uniqueness of the reads, UMIs (short molecular tags composed of a unique random sequence) are added to the reverse transcription step (5'end in template switching or 3' end in oligo-dT primer)[53]. They constitute the second portion of a barcode and primarily detect and quantify unique mRNA transcripts[46] such that amplicons of the same transcript are only counted once. This allows for multiplexing of scRNA-seq, even for low abundant transcripts that show poor reproducibility with previous quantification methods based on the number of sequencing reads[54]. Still, current UMI based approaches are poorly suited for the identification of allele-specific expression or alternative splice variants[2]. Post library construction, cDNA libraries labeled with cellular barcodes are pooled for sequencing, mapping, and alignment. Current RNA-seq technologies rely on pooled-sequencing in order to generate high throughput data. This allows for amplification and sequencing of multiple cells parallelly in the same pool that generates batch-specific output data containing sequences from multiple cells.

3. **Preprocessing scRNA-seq Data**

Journal Pre-proof

Sequencing generates reads (raw data) that need to undergo quality control (QC) before downstream analysis. scRNA-seq data contains many technical artifacts that may arise due to cell bursting leading to RNA leakage, multiple cells sticking together, and lowly expressed RNA leading to dropouts, amplification bias, transcriptional bursting, RNA degradation, and batch effect. Before performing downstream analysis, it is crucial to ensure that all the cellular barcode data obtained from scRNA-seq correspond to viable cells[55]. Another challenge is to prevent false interpretation of technical artifacts (cells that show technical noise that appears as distinguishable gene expression pattern) as biological heterogeneity[56].

A typical scRNA-seq dataset constitutes of three files, genes quantified (gene IDs), cells quantified (cellular barcode data), and a count matrix, irrespective of technology or pipeline used. These files are crucial for building quality matrices for QC assessment. The barcodes are extracted and annotated, called demultiplexing or barcode extraction, followed by mapping and alignment of the read data using read processing pipelines. Post alignment, feature annotation, and quantification are carried out on the data to generate gene expression matrices (*N(cells) x M(genes)*) indicative of the level of gene expression in each cell, based on the molecular counts or read counts. These correspond to high mapping quality exonic loci[55]. A list of preprocessing and downstream analysis tools is shown in Table 1.

Fig. 1. Preprocessing (a) The first step after obtaining the scRNA-seq data is demultiplexing FASTQ batch data. (b) The demultiplexed data (modified FASTQ file) is mapped against respective genome using an alignment program. (c) Feature annotation is carried out using gene annotation file that contains all information on genes, exons, introns and regions of interest (RefSeq, GENCODE). All the reads are filtered, keeping reads that align to the forward and reverse strand with less than three mismatches and mapped only once to the reference genome. (d) Once the gene names for specific reads are obtained, how many reads correspond to which genes can be determined. Tabular count matrix depicting genes/features as rows and cell labels as columns is generated. (e) Multiple single count matrices are joined together to form a combined matrix. (f) Not all batches use same barcodes thus barcodes are filtered for the combined matrix to ensure there is no cross-contamination and a final QC-filtered count matrix is generated. (g) Feature selection and dimensionality reduction is carried out on the matrix carried out on the count matrix followed by downstream analysis.

Journal Pre-proof

3.1. *Quality Control*

QC filtering can be performed using a combination of strategies. Some common strategies are based on assigned barcodes- number of counts per barcode and number of genes per barcode[57]. Cells that show unique gene counts and genes expressed in very few cells are not always indicative of biological heterogeneity, low or high count-depth are indicative of quiescent/damaged cells or doublets/multiplets, respectively. Another covariate used for QC is the fraction of mitochondrial genes per barcode. Elevated levels of mitochondrial genes (above 5-10%) in a cell is an indication that the cell may have broken and the cytoplasmic mRNA content has leaked. Furthermore, RNA spike-ins

(synthetically generated short RNA polymers of known quantity) are used for calibration purposes, where a low mapping ratio between endogenous RNA and spike-ins is indicative of a low-quality library[58].

The quality metrices are visualized to determine the outlier cells. Low-quality cells are filtered out by setting appropriate thresholds. While filtering out outlier cells, multiple independent variables must be considered together rather than individual ones as it can lead to misinterpretation of biological heterogeneity. Thresholds can be fixed or adaptive. Setting fixed thresholds requires experience as suitable thresholds may vary for each experimental protocol or biological systems[59]. An alternative is adaptive thresholds that are decided based on the outlier peaks for the QC covariates. It is essential to reevaluate the QC metrices after filtering before proceeding for further analysis.

3.2. *Normalization*

UMI-based protocols inherently reduce amplification biases, and the addition of spike-ins enables assessment of sensitivity. However, cell-level (counts comparable between cells) and gene-level (counts comparable between genes) normalization is carried out to cater to sampling effects or technical biases that remain in the data due to variability in the protocols. Once the count matrix is obtained, normalization is sought to address the gene expression variability between cells in count data to prevent the highly expressed genes from influencing the analysis. Popular normalization methods have been derived from bulk RNA-seq analysis methods and have been successfully applied to scRNA-seq data such as DESeq and Trimmed Mean of M-values[60]. Popularly, raw scRNA-seq read library normalization is carried out using read count normalization/ CPM (Counts Per Million) methods like RPKM (Read Per Kilobase Million), TPM (Transcripts Per Million), and FPKM (Fragments Per Kilobase Million). The scaling factors in these methods are based on the assumption that the majority of the genes are not differentially expressed, so they might fail when fold-change of DE genes is high across cell populations under study[60]. These library-based or global-scaling normalization methods derived from bulk RNA analysis, other than being computationally intensive, have their shortcomings when used for scRNA-seq data. Given the added complexity of scRNA-seq data due to data sparsity and high heterogeneity, it requires advanced normalization strategies to address specific biases.

SINCERA is a commonly used normalization pipeline, in which gene normalization is performed by z-score, while cell normalization is performed by trimmed mean. For example, during clustering, BISCUIT uses iterative normalization by learning features representing technical modifications. RaceID (Rare cell type identification) normalizes the total transcript count within each cell to the median transcript number across cells [61]. Clustering based on Transcript Compatibility Counts (TCC) uses equivalence classes in place of genes as parameters and normalizes each parameter by distributing the total count across all the cells. sctransform pipeline interfaces with Seurat and Pearson residuals from negative binomial regression, which has been regularized. In regression, sequencing depth is used as a covariate to eliminate technical artifacts[62]. SCnorm uses quantile regression to

approximate the dependency of expression of transcript or depth of sequencing per gene[63]. Similar dependency genes are clustered together, and then a second quantile regression is utilized to approximate scale parameters in every group. In-group correction is achieved by using the approximate scale parameters to deliver normalized estimations of expression.

The diversity of scRNA-seq protocols makes it difficult to standardize any one normalization method. It has been observed that different normalization methods perform optimally for different datasets, and the same goes for cell-level and gene-level normalization. Post normalization, log(x+1) transformed count matrices are obtained that give a simplified account of expression levels in terms of log-fold changes and bring down the skewness of the data[55]. Downstream analyses that are based on the assumption that scRNA-seq data is normally distributed and perform analysis on log-transformed data may sometimes result in counterfeiting DE effects. Thus, there is a pressing need to develop more precise and robust normalization methods designed explicitly for scRNA-seq data.

3.3. *Data Correction*

Normalized data successfully removes amplification and count depth biases; however, a few challenging technical and biological biases remain. Data correction deals with batch effects, dropouts, and biological effects. scRNA-seq data is prone to zero-inflated values, otherwise known as dropouts[64], resulting from low sensitivity of scRNA-seq protocols, inefficient capture of mRNA, low amounts of mRNA in cells, or transient gene expression[2]. The dependency of downstream analysis on the accuracy of gene expression profiles makes the imputation of dropouts a crucial step. Many analysis pipelines account for dropouts during analysis; however, recent findings can change how we look at dropouts. Imputation is carried out either by direct expression analysis or model-based. Newer and more robust ML-based algorithms have taken over popular imputation techniques like Markov Transition Matrix-based MAGIC[65], Clustering-based DrImpute[66], LASSO regression-based ScImpute[64]. SAVER(Single-cell Analysis Via Expression Recovery) uses gene-gene relationships to recover true expression levels of each cell[67], RESCUE (REcovery of Single-Cell Under-detected Expression) enhances cell-type identification based on an ensemble-based method to minimize feature selection bias and count error and perform imputation by comparing gene expression levels of other cells with similar patterns[68]. SCRABBLE, a matrix regularization framework, uses bulk RNA seq data as a constraint that improves the accuracy and estimation of gene expression distribution across cells compared to scRNA-seq analysis in isolation[69]. For datasets that suffer from imbalance and limited sample sizes, scHinter, with a hierarchical framework for random interpolation by leveraging minority oversampling technique[70], proves to be a robust technique than its predecessors. A recent study explores the opposite view of dropouts, where instead of

imputation, it can be used as a signal. It was observed that binary dropout patterns prove almost equally informative as quantitative expression patterns of highly variable genes in cell type identification[71].

Furthermore, studies have been conducted to establish that the excess of dropouts is consistent with stochastic sampling of molecular counts, and any additional zero values may result from biological variation[72]. Such studies suggested that a negative binomial distribution model for UMI based scRNA-seq count data would suffice, and zero-inflation may not always be necessary [46,72,73]. It can also be inferred that the number of dropouts can be decreased by increasing the depth of sequencing or increasing global count with more efficient capturing methods.

Apart from dropouts, several other technical covariates like batch effects and biological covariates need to be considered. Removal of such biases must be carried out simultaneously as there might be a dependency between multiple covariates under consideration. Batch effects arise from data handling in different experiment batches or time points and are highly nonlinear variations. Batch effects can have a significant impact on DE analysis. Some methods include aggregation-based methods[74] that pool cells from batches to form a pseudo-bulk sample and use bulk analysis approaches or nested fixed[75] and mixed effect models[76] that treat batch effects as fixed effects nested within each group or random-effects shared between cells from each batch, respectively. A comparison between different batch correction methods is given in Chen et al., 2020[77], and some popular batch correction methods are listed in table 1. Upon removal of batch effects, data is merged for further unbiased analysis. Biological covariates may arise due to important biological processes like cell cycle effects that affect cell-size and mRNA counts. Correcting for such variation helps reveal important biological signals and processes. Linear regression against a cell cycle score, and correction for cell size during normalization are some ways to remove the effects of the cell cycle[55]. However, data correction for biological effects may not always be in the best interest, and correction for one effect may mask another. Thus, it is advised first to evaluate the study's objective and context before deciding on data correction measures.

Despite the numerous QC measures, it is hard to determine each step's stringency before assessing its effect on downstream analysis. Thus, a feedback system should be followed to regulate QC stringency alongside downstream analysis.

Table 1. List of some popular preprocessing pipelines and downstream analysis tools used for scRNA-seq data Analysis

Pipeline	Environment	Description

category			
Overall analysis	Seurat[78]	R	A comprehensive tool to perform QC analysis on diverse types of single-cell data. Spatial interference using in situ RNA patterns as a reference. Compatible with multimodal data. http://satijalab.org/seurat/
	Scanpy[79]	Python	A scalable analytical framework for scRNA-seq analysis starting from preprocessing, visualization, DE, clustering, and TI works in tandem with AnnData- annotated data matrices. Fast and efficient for large data sets. https://github.com/theislab/scanpy
	PyMINEr[80]	Python	An automated tool for cell-type identification, filtering enriched genes, network pathway analysis, and visualization of analysis. Non-cell type determining gene expression may influence cellular graphs. https://www.sciencesolott.com/pyminer
Pre-processing	dropEst[81]	R	Accurate estimation of quality control, correcting composition bias, and sequencing errors for droplet-based scRNA-seq data. Provides configuration options for accommodating different scRNA-seq protocols. More efficient with smaller datasets than larger datasets. http://github.com/hms-dbmi/dropEst
Visualization	Cerebro[82]	R	An interactive environment, compatible with Seurat objects. https://github.com/romanhaa/Cerebro
	iSEE[83]		Interactive visualization, reproducible, and compatible with existing R/Bioconductor packages. https://github.com/csoneson/iSEE
Imputation	bayNorm[73]	R	An integrated platform for normalization, imputation, and batch effect correction. Improves accuracy and sensitivity of DE analysis. https://bioconductor.org/packages/release/bioc/html/bayNorm.html
	DeepImpute[84]		Employs a deep neural network-based algorithm that allows for improved speed, accuracy, and scalability. It is also well suited for large ever-increasing datasets. https://github.com/lanagarmire/DeepImpute
Batch effect, merging	BERMUDA[85]	R	Deep Autoencoders. Data is obtained from scRNA-seq data, and one common cell type is required at the least from batches for the process https://github.com/txWang/BERMUDA

Journal Pre-proof

	MNN[86]	R	Mutual Nearest Neighbor for batch correction of single cells. Based on the established assumption that a batch is orthogonal to biology and that MNN exists between batches. https://github.com/MarioniLab/MNN2017/
Normalization	SCnorm[63]	R	Makes use of quantile regression to approximate the dependency of expression of a transcript on the depth of sequencing per gene. Similar dependency genes are clustered together, and then a second quantile regression is utilized to approximate scale parameters in every group. For depth sequencing, then in-group correction is achieved by using the approximate scale parameters to deliver normalized estimations of expression. https://www.biostat.wisc.edu/~kendzior/SCNORM/
Dimensionality Reduction	DR-A[87]	Python	Dimensionality Reduction with Adversarial variational autoencoder
	SAUCIE[88]	Python	Deep Multitasking Neural Networks for DR
Cell clustering	Refer to table 3		
Cluster Annotation	scMAP[89]	R	Automated cluster annotation technique follows a projection-based approach where scRNA-seq data is projected onto a previously annotated cell type or dataset.
Pseudotime trajectory inference/reconstruction	TinGa[90]	R	TI based on Growing Neural Graphs. Scalable, time-efficient, and accurate on complex trajectories. Does not require prior specification of the topology by the user. https://github.com/Helena-todd/TinGa
	ReCAT[91]	R	Hidden Markov model-based method for reconstructing cell cycle pseudotime in time-series data https://github.com/tinglab/reCAT
Differential Expression	MAST	R	Uses a linear hurdle model to account for confounders, and DE is determined using the likelihood ratio test. https://github.com/RGLab/MAST\
	SCDE	R	Uses a Bayesian approach that incorporates an evidence-based approach to evaluate the likelihood of the average level of gene expression for individual cells and measure the fold changes. Highly sensitive. https://hms-dbmi.github.io/scde/index.html

Note: https://www.scrna-tools.org/ is a catalog of tools for analyzing single-cell RNA sequencing data

3.4. ***Dimensionality Reduction (DR)***

Journal Pre-proof

scRNA-seq data is computationally intensive, noisy, and suffers from the curse of dimensionality. scRNA-seq expression metrics are of high dimension, but not all genes are required for meaningful classification of cellular expression profiles, and it can practically be explained in fewer dimensions, focusing only on relevant biological signals. DR enables better data visualization and resolves the statistical issue of data sparsity. An effective low-dimensional representation should summarize the data in a few optimal dimensions that must retain the underlying structure in the data to describe the variability of the dataset. Some of the popular techniques for DR are Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding(t-SNE), Uniform Manifold Approximation, and Projection (UMAP), Self Organizing Maps (SOM), and Model embedded dimension reduction[92]. Some clustering pipelines come integrated for both dropout imputation and DR like CIDR[93].DR has two components: feature selection, where one selects a smaller subset from the original set of variables, and the other part is feature extraction, where the high dimension data is projected to a lower dimension. Feature selection is carried out based on the expression variability of genes according to the assumption that genes showing high variability correspond to biological variation. Per-gene variation can be quantified by calculating the variance of log-normalized value. A subset of highly variable genes (HVG), depending on the type of dataset, is selected for further analysis. Selecting a larger subset of HVGs may increase the noise but reduces the risk of discarding biologically relevant signals. Seurat performs feature selection by modeling the mean-variance relationship[94].

After feature selection, DR is carried out using linear or nonlinear techniques. PCA is a linear projection method that linearly transforms the original dataset into PCs ranked in decreasing order of variance, and the data's variance is maximized in the lower dimensional space. It is computationally efficient and removes redundant features, but since scRNA-seq has a highly nonlinear structure, PCA alone is not best suited for data visualization. However, t-SNE, a graph-based nonlinear technique, is much better suited for the task. It is often performed in tandem with PCA. t-SNE is based on a probabilistic distance model. It creates probability distributions to establish a relationship between two data points in high dimensional space and reconstructs it in a lower-dimensional space by optimizing using gradient descent. Although t-SNE is favorable for data visualization, good with nonlinear datasets, and preserves local structures in lower dimensions, it often ignores the global structure, which may lead to misinterpretation of differences between cell populations. Owing to these drawbacks, a superior manifold learning-based DR technique has emerged called Uniform Manifold Approximation and Projection (UMAP)[95]. It is principally similar to t-SNE but also preserves the global structure. It is also computationally more efficient but may sometimes result in spurious signals in smaller datasets. For complex dataset visualization, partition-based graph abstraction (PAGA) is used with UMAP. ZIFA[96] and ZINB-WaVE[97] are model-embedded dimension reduction algorithm carried out on zero-inflated data. ML methods, like Deep Learning and Autoencoders, have shown efficiency in DR problems. Latent Dirichlet Allocation (LDA), a Natural Language Processing-based algorithm, and SAUCIE[88], a neural network-based algorithm, are promising new approaches. DR-A, an autoencoder based framework, provides precise low

Journal Pre-proof

dimensional representation, enhances downstream clustering performance, and could potentially be used for lineage estimation[87].

4. Downstream Analysis

The unique features of single-cell data make downstream analysis elaborate and diversified. There are different stages of preprocessed data such as log-transformed data, batch corrected data, feature selected data, dimensionality reduced data, etc. Depending on the requirement specifications of downstream analysis and availability, certain preprocessed data is chosen. Some open-source scRNA-seq repositories or reference databases are Human Cell Atlas[98], Broad Institute's Single Cell Portal[99], EMBL-EBI's Single Cell Expression Atlas[100], PanglaoDB[101], and OmicX Jingle Bells. 10X Genomics offers datasets at various preprocessed levels for downstream analysis.

Downstream analysis of scRNA-seq data can be of cellular level or gene level. This section will discuss various analysis tools and their applications that have led to prominent discoveries.

4.1. *Cell- level Analysis*

The cell-level analysis is about understanding cell subtypes, cell differentiation patterns, cell lineage, cellular trajectories, identify novel cellular markers, and many unique cellular features. It helps characterize known properties of cells and uncover previously unidentified characteristics. Cell-level analysis is not exclusive to gene-level analysis. We briefly discuss important cell-level analysis techniques that have found major implications in advancing cell biology research.

4.1.1. *Cluster Analysis*

Clustering entails categorizing cells into clusters to enable the identification of cell types and subtypes. It is ideally performed on a dimensionally reduced dataset, and clusters are made based on cell-specific molecular profiles. Results of a clustering analysis can itself be of much significance or can serve as a covariate in other downstream analyses. There has been an attempt to develop robust cluster analysis algorithms (as shown in Table 2) that can address the vast heterogeneity of cells, but there is still a lack of an optimal algorithm that fares well across datasets.

Table 2. Classes of clustering algorithms used in scRNA-seq following ways[92]

Class of clustering algorithm	Principle	Limitations	Pipeline

Journal Pre-proof

Distance-matrix	Unsupervised learning algorithms like k- means falls in this category. This algorithm first identifies k centroids or means iteratively, and data points are assigned to the cluster around the nearest centroids. During cluster allocation of datapoints, the in-cluster sum of squares is reduced, and the position of centroids is iteratively optimized. Scalable and time-efficient.	Sensitive to outliers, biased towards data shape/cluster shape, and the number of clusters must be specified beforehand.	SCUBA[102], PCAKmeans, pcaReduce, SAIC[103], scVCMD[104]
Hierarchical clustering	Generates clusters into a hierarchical structure and is popular in gene expression analysis. It overcomes the limitation of k-means of specifying the number of clusters a priori and handling different shapes of clusters. No assumptions are made about the distribution of data points, and each cluster links to another by branches and is nested like a hierarchical tree in the form of a dendrogram. This representation facilitates meaningful data interpretation.	Time intensive	BackSPIN, cellTree[105], DendroSplit[106] CIDR[93]
Graph-based	Supervised-learning algorithms. Projects a graph representation of data in which the nodes correspond to datapoints/cells, and edges correspond to pairwise similarity between the datapoints, for example, K-Nearest Neighbor. Clusters are based on neighboring cell pairs. Graph-based clustering techniques have various subtypes like Louvain clustering that is used by popular scRNA-seq analysis pipeline. Louvain algorithm is largely getting taken over by the Leiden algorithm for cluster detection, which uses a smart local move approach faster and shows more proficiency in detecting well-connected communities[107].	Reliance on heuristic solutions sometimes leads to spurious results, and iteration sometimes masks small communities.	Seurat[78] scanpy[79] SNN-Cliq[108]
Mixture Models	Clustering is based on the probability distribution of the datapoints. It is well suited for the identification of subpopulations and integrates prior knowledge as assumptions of probability distributions.	Computationally intensive and relies on the accuracy of assumptions of probability distribution.	BISCUIT[109], Seurat, TSCAN[110]
Density-based	Density-based algorithms assign clusters based on high-density regions of datapoints. It is a highly efficient clustering technique but is sensitive to parameters.	Sensitive to parameters, time-intensive	Monocle2[111] (for the identification of

			outlier cells)
Neural Network	Supervised-learning methods, inspired by the neural network of the human nervous system. Highly efficient in performing clustering and classification tasks. Kohonen networks are bilayer networks that use competitive learning for clustering. Deep learning and autoencoders are also used. These are efficient, scalable, and information on relationships among clusters can be incorporated.	Sensitive to parameters	SAUCIE[88] scDeepCluster[112]
Ensemble clustering	Ensemble clustering algorithms are a solution to the lack of any one optimal algorithm as it makes use of multiple clustering algorithms on the same dataset, and a consensus result is obtained that is more precise than independent algorithms.	Disadvantages of individual algorithms add up	SAFE-clustering uses SC3, CIDR, Seurat, and t-SNE + k-means for clustering and then combines the result to obtain one consensus result by using a hypergraph partitioning algorithm[113]

After clustering, the clusters need to be annotated in order to give them biological relevance. *Cluster annotation* can be achieved either by thorough examination of literature, reference cell databases, or by identifying gene signatures or differentially expressed marker genes. CellAssign[114] and scMAP[89] rely on the former technique for cluster annotation, but since most marker genes have been identified by bulk analysis in the reference datasets, it is limited to giving a classical view of the cell types. Also, it is not necessary that cell types in reference databases will correspond to all the cell types present in the dataset under investigation. Analysis pipelines like Seurat, SC3, scVDMC, make use of differential gene expression approach. Full gene expression profiles are used in DE analysis for marker gene identification and cluster annotation, that is performed using simple statistical tests. The quantitative levels in gene expression are measured amongst clusters and all the cells in the dataset. Based on statistical tests like Wilcoxon rank sum test (used in Seurat), Welch's t test, Kruskal-Wallis test, etc. marker gene sets are identified i.e., the top-ranked genes from these tests. DE analysis could either be carried out in succession to clustering like in Seurat, simultaneously like in scVDMC and DendroSplit, or by using DE software like MAST[115], SCDE[116], and ZingeR. Gene set enrichment analysis is carried out against reference gene sets (set of genes grouped as they share common chromosome location, biological function, or regulation) using statistical parameters like Jaccard index, and clusters are annotated accordingly. The important thing to note here is the *p*-value that is based on the assumption that

Journal Pre-proof

the marker gene identified represents the biological phenomenon, but the *p*-value is often inflated and leads to an overestimation of marker genes. Most existing GSEA methods have been developed for bulk RNA seq analysis and perform poorly in case of scRNA-seq, thus Ma, Y., Sun, S., Shang, X. et al. came up with an integrative DE-GSE analysis technique called iDEA[117] that makes use of DE summary statistic and thus easy to use with current DE methods and efficiently produces well-assessed *p*-values for enriched gene set detection. Nowadays, automated cluster annotation techniques are becoming increasingly available like scMAP[89], which follows a projection-based approach where scRNA-seq data is projected onto a previously annotated cell type or dataset. Garnett uses a supervised classification approach for rapid annotation[118], scCATCH[119] makes use of CellMatch reference database for annotation followed by evidence-based scoring for increased performance. Another important goal is to identify rare cell types that may appear as outliers in clustering results that only consider global differences in gene expression. RaceID[61] and GiniClust[120] are clustering algorithms sensitive to identifying rare cell types. RaceID is based on the assumption that a given cell-type must express some genes that are specific to the cell type and appear as outliers but if the focus is shifted from global to such cells and the technical and the biological noise are accounted for by setting appropriate thresholds, it will enable the identification of rare cell types.

To an extent, the determination of cell-types is dependent on user-defined criteria, since for different researchers, the level of cluster resolution may vary, and in some cases, sub clustering of clusters may also be required. The choice of the resolution also has a significant effect on the results. DendroSplit, a clustering framework, allows the user to cluster using feature selection that enables identifying multiple levels of biologically meaningful cell populations in a dataset, also suitable for detecting rare cells. [106]. Despite continual attempts on developing new algorithms, clustering and cluster annotations suffer from various challenges in both biological and computational front. It is advisable to use a cocktail of automated and manual cluster annotation measures to get precise results. But since clustering is an unsupervised learning approach, several parameters are needed for reliable evaluation. Statistical and experimental validation is often needed. Transient biological states make it difficult to identify cell states. However, the generation of more comprehensive and extensive cell atlases will facilitate better clustering, cluster annotation results, and better computational approaches to help overcome technical challenges.

4.1.2. *Trajectory Inference*

Trajectory Inference (TI), also known as pseudo-temporal ordering, is a process of characterization of underlying dynamic cellular processes. Although clustering successfully builds discrete clusters of cell types and subtypes, it does not account for the variability due to dynamic cellular processes like transient cell states in cell differentiation, cell cycle, or environmental effect. TI deals with this blind spot by ordering cells along a continuous path that minimizes transcriptional changes between successive cell pairs, called pseudotime (one dimensional manifold), that represents the progression of the cell through its dynamic processes measured in terms of transcriptional changes that a cell undergoes during a biological process. Some datasets have an expected temporal component such as cells from

developing embryos, immune cells during an immune response, tumor cells, progenitor stem cells, etc. Understanding of cell differentiation via bulk RNA analysis gave us the impression that cell differentiation occurs in discrete stages, but in reality, it is a continuous process that may appear chaotic but can be ordered along continuous trajectories. Following the cells along a pseudo-temporal trajectory and analyzing gene expression changes yields valuable insights into the cellular regulatory processes, dynamic states, and abnormal cell states. The progression of cells in a given cellular process is rarely synchronized. Capturing this asynchrony poses a unique challenge of deciphering the sequence of regulatory events. TI takes into account a snapshot view of these events and uses computational techniques to infer the order of the cells along their developmental trajectories. The derived trajectory topologies can be linear, bifurcating, multifurcating, complex tree structures, or graph structures[121]. TI is mostly used for cell-lineage construction. During cellular developmental stages, cells express unique cellular markers and various lineage marks that can serve helpful for tracing lineage along pseudotime trajectories, such as somatic mutations, single nucleotide polymorphisms, copy number variants, microsatellites, transposons, and retroviral sequences.

Two broad strategies used for TI are DR-based methods (Monocle[122], Wishbone[123],), which use the reduced latent space as the first phase in inference and assigns pseudotime to individual cells, or clustering-based methods (TSCAN[110], SCUBA[102], ÉCLAIR[124]) which builds a network connecting clusters and applies pseudo-temporal ordering of clusters[125]. Monocle is a pioneering pseudo-temporal ordering algorithm to demonstrate how pseudotime analysis can reveal important cellular regulatory interactions[122]. It uses an unsupervised learning approach that constructs minimum spanning trees (MST) for ordering cells along pseudotime. Several other algorithms like Wishbone, TSCAN, versions of Monocle (currently Monocle 3) have been developed that are more robust and accurate, and a detailed comparative account can be found in Saelens et al. 2018[121]. While most TI algorithms are unsupervised learning-based models, Ouija is a supervised learning-based algorithm that was developed keeping in mind that several confounding factors that affect biological processes like cell-cycle and apoptosis sometimes need to be accounted for to get biologically plausible pseudotime trajectories[126]. It uses switch-like marker genes and can be used as a complementary method with existing methods owing to its consideration of gene-specific behaviors as opposed to unsupervised methods. Despite the availability of more than 70 TI methods, researchers find it challenging to determine which is best suited for their analysis. Selection depends on the task like types of biological processes being studied, whether it is a cell differentiation process (Wishbone), lineage trees (MerLOT[127]), cell cycle (Cyclone[128], reCAT[91]), or downstream analysis, like inferring GRN (SCODE[129]), DE. Each trajectory method has its pros and cons, and there is a lack of standardization. dynverse is a collection of R packages specifically designed to address this issue so that researchers can perform TI, quantify it, or compare it to other available methods to decide on the best approach for their dataset[121]. Recent advancements have led to the use of time-series data in place of snapshot data. Tempora is an upcoming algorithm that may prove to

Journal Pre-proof

be more biologically relevant than previous methods as it uses biological pathway information and identifies time-dependent pathways for ordering and inferring time-series data[130].

Undoubtedly, TI is becoming a popular tool for studying biological processes like cellular differentiation, immune response, tumor progression, and resistance, but inferring trajectories alone cannot be reliable as it needs to be validated using supporting biological evidence. With better algorithms and more time-series data availability, TI can be used to predict pre-disease state of cells, which may help in the early detection of diseases.

4.2. ***Gene-level Analysis***

Gene-level analysis is an integral part of cell-level analysis for studying cellular structures and identities, but independently gene-level analysis of single cells reveals a more comprehensive inference of cellular pathways and regulatory networks. It involves DE analysis, pathway analysis, gene regulatory networks (GRN), and gene set analysis. We shall be discussing some of the important gene-level analysis and the information they have revealed.

4.2.1. *Differential Expression (DE)*

We have discussed DE testing in previous sections while discussing clustering, but at the gene-level, we focus more on the stochastic nature of gene expression, and distinctive signatures only observed at the single-cell level. Although principally the same as bulk DE analysis, DE for single-cell data was developed to deal with artifacts inherent to scRNA-seq such as multimodality, dropouts, and heterogeneity. Among popular DE tools for scRNA-seq, MAST uses a linear hurdle model to account for confounders, and DE is determined using the likelihood ratio test[115]. SCDE uses a Bayesian approach that incorporates an evidence-based approach to evaluate the likelihood of the average level of gene expression for individual cells and measure the fold changes[116]. These techniques have shown higher sensitivity to other techniques. In some cases, bulk DE methods, when used with gene-weights, exceeds performance on scRNA-seq specific DE methods but at the price of being computationally intensive. Apart from having low detection accuracy for true DE genes, there is also a lack of agreement between various available tools. This demands better tools that account for the multimodality of scRNA-seq data, its artifacts, and identifies true DE genes having biological relevance. Wang, Tianyu et al. performed a recent comparative analysis of DE tools that could guide researchers to evaluate DE tools, choose appropriate ones for their analysis, and improve upon existing techniques[131]. Trajectory-based DE methods have also been developed called tradeSeq, enabling DE analysis between-lineage and within-lineage, providing a continuous resolution of gene expression changes through a dynamic process[132].

DE studies allow us to identify distinct expression profiles of cellular pathways that help us understand the effect of perturbations and the underlying mechanisms of disease pathologies.

4.2.2. *Gene Regulatory Network (GRN)*

The above discussion on gene expression poses another question on how gene expression is regulated in cells. Stochastic gene expression observed amongst single-cells indicates that gene regulation that relies on transcription factors, signaling molecule, and co-factors is regulated in a specific way. Uncovering these GRNs will reveal the basis of gene expression stochasticity and provide mechanistic insights into normal and abnormal cellular phenotypes. Several tools have been developed for inferring GRN, some derived from bulk analysis methods and some designed explicitly for scRNA-seq data, but GRNs are incredibly complex to decipher. The SCENIC algorithm simultaneously constructs gene regulatory networks and performs clustering[133]. It bases the identification of stable cellular states on the activity of GRNs in each cell and is well suited for the discovery of cell states that are driven by transcription factors and cis-regulatory sequences. SCGRN is a supervised feature learning-based approach for inferring GRNs that employs three different ML techniques[134], is promising against unsupervised learning approaches used earlier. Another study by Qin et al. presents a toolkit, Scribe, for inferring causal GRNs from single-cell datasets that indicate that pseudotime data perform poorly compared to true time-series data[136]. Such insights will encourage newer studies that focus on deriving true biological insights from scRNA-seq data. Recently, researchers explored GRN inference algorithms and developed a framework called BEELINE, where they used synthetic networks with predictable trajectories, literature curated Boolean models, and diverse transcriptional regulatory networks to assess the accuracy of GRN methods[135]. This study can be used as a benchmark for selecting GRN algorithms or for developing new strategies.

The study of GRNs at single-cell resolution will contribute significantly to system biology approaches and help build more precise models. These models can further help discover dynamical network biomarkers that can predict disease and pre-disease states.

5. **Applications and Future Prospects**

The past decade has seen a momentous upsurge in SCT and its applications across neurology, microbiology, cell biology, molecular biology, immunology, cancer biology, bioinformatics, stem cell, biomedical sciences, and clinical and diagnostic applications. Recent research efforts evince its potential to be of use in translational research. To list all the applications is beyond the scope of this article. We list some of the major areas that have benefited from SCT are listed in Table 3.

Journal Pre-proof

Table 3. Current and future applications of scRNA-seq analysis in major fields of biomedical sciences*

Field of study	Scope of analysis	Applications
Immunology	• Clustering of regional immune cells • Trajectory analysis of individual immune lineages	Identification of novel immune cell subtypes[137], revealing immune microenvironment across tissues[138], understanding regional immunity in tumors[139]. Build developmental trajectory of immune cells, and gain mechanistic insights[140]. Immunology studies with the aid of SCT will help us develop better and targeted immunotherapies.
Cancer biology	• Clustering of tumor microenvironment • DE analysis of tumors • Construction of gene regulatory maps	Researchers and clinicians have struggled with understanding tumor heterogeneity for a long time. scRNA-seq has revealed intra-tumor, inter-tumor heterogeneity, and rare tumor subpopulations[141]. It will help understand cell-cell interactions in the tumor ecosystem, tumor resistance, refractory, and recurrence mechanisms. It can help elucidate genetic and non-genetic mechanisms for cancer and help devise better treatment regimes.
Cell-cell communication studies	• Integrating the count matrix generated from scRNA-seq with known ligand-receptor interaction matrix • Construction of GRNs	Deciphering cell-cell interactions is crucial to understanding both cellular development and diseases. SCT has enabled the inference of ligand-receptor interactions at an unprecedented resolution. Employing SCT we can identify communication patterns and use them to predict functions of poorly studied pathways. Tools like NATMI[142] is being used to identify which cell-type pairs or cellular communities communicate more frequently or specifically, what ligand-receptor pairs are the most active within a network, and has offered insights into autocrine signaling in cell-cell communication.
Stem cell	• Trajectory analysis of stem cells • DE analysis of progenitor cells	When used to construct hematopoietic lineage trees, SCT analysis revealed that the differentiation process is continuous instead of the traditional belief that it is a stepwise process[143]. Also helped reveal a novel pathway used by stem-cells for self-renewal[144]. Another group of researchers developed a scRNA-seq based CRISPER interference technique to study transcription factors involved in human endoderm development that revealed underlying factors and effects of perturbation[145]. It can help devise novel methods for treating genetic diseases and developmental disorders and better stem-cell therapies.
Neurobiology	• Clustering of	It has been used to identify neuronal subtypes[146],

Journal Pre-proof

	regional brain cells • Pseudo temporal ordering • DE analysis of partitioned brain functions	understanding cellular programs involved in early development, cell populations in neuronal diseases, tracking transcriptional landscape of aging, resolving regional cell type landscape, and other nuances of brain function[147]. scRNA-seq has led to rapid accumulation of normal and tumor brain cell data, which can be further used to understand brain function and diseases.
Infection biology	• Classification of cellular and viral transcriptomes • Genome profiling of host cell during infection	The recent COVID-19 pandemic has drawn all the attention to infectious diseases and host-pathogen interactions. It has also made us highly aware of the lack of robust techniques to understand infection mechanisms and devise treatment strategies. scRNA-seq analysis can help uncover host-pathogen relationships, map infection progression, and help identify druggable targets. The study of human microbiome interaction with immune cells will indicate the pathological state that develops when it is perturbed[148,149].
Diagnostics	• DE analysis and Gene regulatory network construction for identification of novel diagnostic biomarkers.	scRNA-seq data is becoming increasingly available for pre-clinical and clinical samples of diseases. It enables the construction of high-resolution cellular maps for diseases, helps identify novel biomarkers, and unravels underlying disease mechanisms that aids in developing better treatment regimes. Used in Cerebrospinal Fluid Research and diagnostics[150], a molecular diagnostic test using scRNA-seq analysis identified two gene-sets involved in the autoimmune response that is suggestive of disease progression and could drive lupus nephritis treatment[151]. Researchers successfully identified and validated diagnostic and therapeutic biomarkers for rheumatoid arthritis for mice and humans, using scRNA-seq analysis and network tools[152]. In the future, it may also help in identifying dynamical network biomarkers that will help predict a pre-disease state.
Precision Medicine	• Longitudinal profiling with scRNA-seq. • Pre- and post-treatment analysis	scRNA-seq helps identify the more notorious clones or malignant cells in cancer, thus tailored therapy could be designed to target this particular group of cells. Profiling patient samples can help monitor disease states, response to therapies, and mechanisms of resistance[153,154]. However, for this ambitious idea to become a reality, it would require more cost-effective and reliable analysis techniques than

Journal Pre-proof

currently available.

** applications in different fields are not mutually exclusive*

scRNA transcriptomics can be used with multiple other technologies to yield more comprehensive results for understanding cellular biology. Spatial transcriptomics is one such technology that preserves the spatial location of gene expression in cells during analysis, and when used with scRNA-seq, is an excellent way to study tissue microenvironment. The possibilities of what can be achieved with this technology are countless.

6. Conclusion

Single-cell RNA sequencing technology has zoomed in on cellular biology like never before. This review objectively points out that single-cell analysis is a multi-step process, and no one step is exclusive of the other. All the steps, only when carefully monitored in tandem, will give precise results. Leveraging SCT and multiple single-cell modalities at once bears a remarkable ingenuity in understanding complex cellular processes, capturing cellular heterogeneity, and disease states. It opens new frontiers of research. Efforts to characterize all cells in a human body, such as the Chan-Zuckerberg Initiative- Human Cell Atlas, serve as major reservoirs for researchers across several fields from biological sciences to computational sciences to come together and develop from. There is an increasing demand for better tools, techniques, analysis algorithms, and experimental validation measures that can rapidly materialize the vision of understanding biological processes at a single-cell resolution.

Acknowledgments

This work was supported by the project "Genetic Analysis of Dermatological Disorders" (BT/PR5402/BID/7/408/2012 dated: 6/7/2017), Department of Biotechnology, Government of India.

Conflict of interest

The authors do not have any conflicts of interests to declare.

References

[1] P. Angerer, L. Simon, S. Tritschler, F.A. Wolf, D. Fischer, F.J. Theis, Single cells make big data: New challenges and opportunities in transcriptomics, Curr. Opin. Syst. Biol. (2017). doi:10.1016/j.coisb.2017.07.004.

[2] B. Hwang, J.H. Lee, D. Bang, Single-cell RNA sequencing technologies and bioinformatics pipelines, Exp. Mol. Med. (2018). doi:10.1038/s12276-018-0071-8.

[3] S. Huang, Non-genetic heterogeneity of cells in development: More than just noise, Development. (2009). doi:10.1242/dev.035139.

Journal Pre-proof

[4] N. Li, H. Clevers, Coexistence of quiescent and active adult stem cells in mammals, Science (80-.). (2010). doi:10.1126/science.1180794.

[5] B.D. Aevermann, M. Novotny, T. Bakken, J.A. Miller, A.D. Diehl, D. Osumi-Sutherland, R.S. Lasken, E.S. Lein, R.H. Scheuermann, Cell type discovery using single-cell transcriptomics: Implications for ontological representation, Hum. Mol. Genet. (2018). doi:10.1093/hmg/ddy100.

[6] S. Linnarsson, S.A. Teichmann, Single-cell genomics: Coming of age, Genome Biol. (2016). doi:10.1186/s13059-016-0960-x.

[7] E. Shapiro, T. Biezuner, S. Linnarsson, Single-cell sequencing-based technologies will revolutionize whole-organism science, Nat. Rev. Genet. (2013). doi:10.1038/nrg3542.

[8] P. Angerer, L. Simon, S. Tritschler, F.A. Wolf, D. Fischer, F.J. Theis, Single cells make big data: New challenges and opportunities in transcriptomics, Curr. Opin. Syst. Biol. 4 (2017) 85-91. doi:10.1016/j.coisb.2017.07.004.

[9] G. Brady, M. Barbara, N.N. Iscove, Representative in vitro cDNA amplification from individual hemopoietic cells and colonies, Methods Mol. Cell. Biol. (1990).

[10] J. Eberwine, H. Yeh, K. Miyashiro, Y. Cao, S. Nair, R. Finnell, M. Zettel, P. Coleman, Analysis of gene expression in single live neurons, Proc. Natl. Acad. Sci. U. S. A. (1992). doi:10.1073/pnas.89.7.3010.

[11] F. Tang, K. Lao, M.A. Surani, Development and applications of single-cell transcriptome analysis, Nat. Methods. (2011). doi:10.1038/nmeth.1557.

[12] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B.B. Tuch, A. Siddiqui, K. Lao, M.A. Surani, mRNA-Seq whole-transcriptome analysis of a single cell, Nat. Methods. (2009). doi:10.1038/nmeth.1315.

[13] D. Hebenstreit, Methods, challenges and potentials of single cell RNA-seq, Biology (Basel). (2012). doi:10.3390/biology1030658.

[14] D. Grün, A. Lyubimova, L. Kester, K. Wiebrands, O. Basak, N. Sasaki, H. Clevers, A. Van Oudenaarden, Single-cell messenger RNA sequencing reveals rare intestinal cell types, Nature. (2015). doi:10.1038/nature14966.

[15] S. Petropoulos, D. Edsgärd, B. Reinius, Q. Deng, S.P. Panula, S. Codeluppi, A. Plaza Reyes, S. Linnarsson, R. Sandberg, F. Lanner, Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos, Cell. (2016). doi:10.1016/j.cell.2016.03.023.

[16] A.A. Tu, T.M. Gierahn, B. Monian, D.M. Morgan, N.K. Mehta, B. Ruiter, W.G. Shreffler, A.K. Shalek, J.C. Love, TCR sequencing paired with massively parallel 3′ RNA-seq reveals clonotypic T cell signatures, Nat. Immunol. (2019). doi:10.1038/s41590-019-0544-5.

[17] R.J. Miragaia, T. Gomes, A. Chomka, L. Jardine, A. Riedel, A.N. Hegazy, N. Whibley, A. Tucci, X. Chen, I. Lindeman, G. Emerton, T. Krausgruber, J. Shields, M. Haniffa, F. Powrie, S.A. Teichmann, Single-Cell Transcriptomics of Regulatory T Cells Reveals Trajectories of Tissue Adaptation, Immunity. (2019). doi:10.1016/j.immuni.2019.01.001.

[18] M.J.T. Stubbington, T. Lönnberg, V. Proserpio, S. Clare, A.O. Speak, G. Dougan, S.A. Teichmann, T cell fate and clonality inference from single-cell transcriptomes, Nat. Methods. (2016). doi:10.1038/nmeth.3800.

[19] A.K. Shalek, R. Satija, X. Adiconis, R.S. Gertner, J.T. Gaublomme, R. Raychowdhury, S. Schwartz, N. Yosef, C. Malboeuf, D. Lu, J.J. Trombetta, D. Gennert, A. Gnirke, A. Goren, N. Hacohen, J.Z. Levin, H. Park, A. Regev, Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells, Nature. (2013). doi:10.1038/nature12172.

[20] J. Wagner, M.A. Rapsomaniki, S. Chevrier, T. Anzeneder, C. Langwieder, A. Dykgers, M. Rees, A. Ramaswamy, S. Muenst, S.D. Soysal, A. Jacobs, J. Windhager, K. Silina, M. van den Broek, K.J. Dedes, M. Rodríguez Martínez, W.P. Weber, B. Bodenmiller, A Single-Cell Atlas of the Tumor and Immune Ecosystem of Human Breast Cancer, Cell. (2019). doi:10.1016/j.cell.2019.03.005.

[21] J.M. Granja, S. Klemm, L.M. McGinnis, A.S. Kathiria, A. Mezger, M.R. Corces, B. Parks, E. Gars, M. Liedtke, G.X.Y. Zheng, H.Y. Chang, R. Majeti, W.J. Greenleaf, Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia, Nat. Biotechnol. (2019). doi:10.1038/s41587-019-0332-7.

[22] C. Yao, H.W. Sun, N.E. Lacey, Y. Ji, E.A. Moseman, H.Y. Shih, E.F. Heuston, M. Kirby, S. Anderson, J. Cheng, O. Khan, R. Handon, J. Reilley, J. Fioravanti, J. Hu, S. Gossa, E.J. Wherry, L. Gattinoni, D.B. McGavern, J.J. O'Shea, P.L. Schwartzberg, T. Wu, Single-cell RNA-seq reveals TOX as a key regulator of CD8+ T cell persistence in chronic infection, Nat. Immunol. (2019). doi:10.1038/s41590-019-0403-4.

[23] S.M. Shaffer, M.C. Dunagin, S.R. Torborg, E.A. Torre, B. Emert, C. Krepler, M. Beqiri, K. Sproesser, P.A. Brafford, M. Xiao, E. Eggan, I.N. Anastopoulos, C.A. Vargas-Garcia, A. Singh, K.L. Nathanson, M. Herlyn, A. Raj, Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance, Nature. (2017). doi:10.1038/nature22794.

Journal Pre-proof

[24] P. Yu, W. Lin, Single-cell Transcriptome Study as Big Data, Genomics, Proteomics Bioinforma. (2016). doi:10.1016/j.gpb.2016.01.005.

[25] J. Zheng, K. Wang, Emerging deep learning methods for single-cell RNA-seq data analysis, Quant. Biol. (2019). doi:10.1007/s40484-019-0189-2.

[26] R. Petegrosso, Z. Li, R. Kuang, Machine learning and statistical methods for clustering single-cell RNA-sequencing data, Brief. Bioinform. (2019). doi:10.1093/bib/bbz063.

[27] B. Vieth, S. Parekh, C. Ziegenhain, W. Enard, I. Hellmann, A systematic evaluation of single cell RNA-seq analysis pipelines, Nat. Commun. (2019). doi:10.1038/s41467-019-12266-7.

[28] G. Chen, B. Ning, T. Shi, Single-cell RNA-seq technologies and related computational data analysis, Front. Genet. (2019). doi:10.3389/fgene.2019.00317.

[29] K. Sheng, W. Cao, Y. Niu, Q. Deng, C. Zong, Effective detection of variation in single-cell transcriptomes using MATQ-seq, Nat. Methods. (2017). doi:10.1038/nmeth.4145.

[30] S. Picelli, Å.K. Björklund, O.R. Faridani, S. Sagasser, G. Winberg, R. Sandberg, Smart-seq2 for sensitive full-length transcriptome profiling in single cells, Nat. Methods. (2013). doi:10.1038/nmeth.2639.

[31] L.D. Goldstein, Y.J.J. Chen, J. Dunne, A. Mir, H. Hubschle, J. Guillory, W. Yuan, J. Zhang, J. Stinson, B. Jaiswal, K.B. Pahuja, I. Mann, T. Schaal, L. Chan, S. Anandakrishnan, C. wah Lin, P. Espinoza, S. Husain, L. Shapiro, K. Swaminathan, S. Wei, M. Srinivasan, S. Seshagiri, Z. Modrusan, Massively parallel nanowell-based single-cell gene expression profiling, BMC Genomics. (2017). doi:10.1186/s12864-017-3893-1.

[32] X. Fan, X. Zhang, X. Wu, H. Guo, Y. Hu, F. Tang, Y. Huang, Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos, Genome Biol. (2015). doi:10.1186/s13059-015-0706-1.

[33] S. Islam, U. Kjällquist, A. Moliner, P. Zajac, J.B. Fan, P. Lönnerberg, S. Linnarsson, Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq, Genome Res. (2011). doi:10.1101/gr.110882.110.

[34] S. Islam, U. Kjällquist, A. Moliner, P. Zajac, J.B. Fan, P. Lönnerberg, S. Linnarsson, Highly multiplexed and strand-specific single-cell RNA 5′ end sequencing, Nat. Protoc. (2012). doi:10.1038/nprot.2012.022.

[35] G.X.Y. Zheng, J.M. Terry, P. Belgrader, P. Ryvkin, Z.W. Bent, R. Wilson, S.B. Ziraldo, T.D. Wheeler, G.P. McDermott, J. Zhu, M.T. Gregory, J. Shuga, L. Montesclaros, J.G. Underwood, D.A. Masquelier, S.Y. Nishimura, M. Schnall-Levin, P.W. Wyatt, C.M. Hindson, R. Bharadwaj, A. Wong, K.D. Ness, L.W. Beppu, H.J. Deeg, C. McFarland, K.R. Loeb, W.J. Valente, N.G. Ericson, E.A. Stevens, J.P. Radich, T.S. Mikkelsen, B.J. Hindson, J.H. Bielas, Massively parallel digital transcriptional profiling of single cells, Nat. Commun. (2017). doi:10.1038/ncomms14049.

[36] D.M. DeLaughter, The Use of the Fluidigm C1 for RNA Expression Analyses of Single Cells, Curr. Protoc. Mol. Biol. (2018). doi:10.1002/cpmb.55.

[37] E.Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A.R. Bialas, N. Kamitaki, E.M. Martersteck, J.J. Trombetta, D.A. Weitz, J.R. Sanes, A.K. Shalek, A. Regev, S.A. McCarroll, Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets, Cell. (2015). doi:10.1016/j.cell.2015.05.002.

[38] A.M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D.A. Weitz, M.W. Kirschner, Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells, Cell. (2015). doi:10.1016/j.cell.2015.04.044.

[39] P. Hu, W. Zhang, H. Xin, G. Deng, Single cell isolation and analysis, Front. Cell Dev. Biol. (2016). doi:10.3389/fcell.2016.00116.

[40] A.E. Saliba, A.J. Westermann, S.A. Gorski, J. Vogel, Single-cell RNA-seq: Advances and future challenges, Nucleic Acids Res. (2014). doi:10.1093/nar/gku555.

[41] V. Menon, Clustering single cells: A review of approaches on high-and low-depth single-cell RNA-seq data, Brief. Funct. Genomics. (2018). doi:10.1093/bfgp/elx044.

[42] M.S. Cembrowski, Single-cell transcriptomics as a framework and roadmap for understanding the brain, J. Neurosci. Methods. (2019). doi:10.1016/j.jneumeth.2019.108353.

[43] D.B. Weibel, G.M. Whitesides, Applications of microfluidics in chemical biology, Curr. Opin. Chem. Biol. (2006). doi:10.1016/j.cbpa.2006.10.016.

[44] J.S. Marcus, W.F. Anderson, S.R. Quake, Microfluidic single-cell mRNA isolation and analysis, Anal. Chem. (2006).

Journal Pre-proof

doi:10.1021/ac0519460.

[45] X. Chen, S.A. Teichmann, K.B. Meyer, From Tissues to Cell Types and Back: Single-Cell Gene Expression Analysis of Tissue Architecture, Annu. Rev. Biomed. Data Sci. (2018). doi:10.1146/annurev-biodatasci-080917-013452.

[46] C. Ziegenhain, B. Vieth, S. Parekh, B. Reinius, A. Guillaumet-Adkins, M. Smets, H. Leonhardt, H. Heyn, I. Hellmann, W. Enard, Comparative Analysis of Single-Cell RNA Sequencing Methods, Mol. Cell. (2017). doi:10.1016/j.molcel.2017.01.023.

[47] V. Espina, J.D. Wulfkuhle, V.S. Calvert, A. VanMeter, W. Zhou, G. Coukos, D.H. Geho, E.F. Petricoin, L.A. Liotta, Laser-capture microdissection, Nat. Protoc. (2006). doi:10.1038/nprot.2006.85.

[48] S. Nichterwitz, G. Chen, J. Aguila Benitez, M. Yilmaz, H. Storvall, M. Cao, R. Sandberg, Q. Deng, E. Hedlund, Laser capture microscopy coupled with Smart-seq2 for precise spatial transcriptomic profiling, Nat. Commun. (2016). doi:10.1038/ncomms12139.

[49] T.M. Gierahn, M.H. Wadsworth, T.K. Hughes, B.D. Bryson, A. Butler, R. Satija, S. Fortune, J. Christopher Love, A.K. Shalek, Seq-Well: Portable, low-cost rna sequencing of single cells at high throughput, Nat. Methods. 14 (2017) 395-398. doi:10.1038/nmeth.4179.

[50] J. Cao, J.S. Packer, V. Ramani, D.A. Cusanovich, C. Huynh, R. Daza, X. Qiu, C. Lee, S.N. Furlan, F.J. Steemers, A. Adey, R.H. Waterston, C. Trapnell, J. Shendure, Comprehensive single-cell transcriptional profiling of a multicellular organism, Science (80-.). (2017). doi:10.1126/science.aam8940.

[51] A.B. Rosenberg, C.M. Roco, R.A. Muscat, A. Kuchina, P. Sample, Z. Yao, L.T. Graybuck, D.J. Peeler, S. Mukherjee, W. Chen, S.H. Pun, D.L. Sellers, B. Tasic, G. Seelig, Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding, Science (80-.). 360 (2018) 176-182. doi:10.1126/science.aam8999.

[52] B. Hwang, J.H. Lee, D. Bang, Single-cell RNA sequencing technologies and bioinformatics pipelines, Exp. Mol. Med. 50 (2018). doi:10.1038/s12276-018-0071-8.

[53] T. Kivioja, A. Vähärautio, K. Karlsson, M. Bonke, M. Enge, S. Linnarsson, J. Taipale, Counting absolute numbers of molecules using unique molecular identifiers, Nat. Methods. (2012). doi:10.1038/nmeth.1778.

[54] S. Islam, A. Zeisel, S. Joost, G. La Manno, P. Zajac, M. Kasper, P. Lönnerberg, S. Linnarsson, Quantitative single-cell RNA-seq with unique molecular identifiers, Nat. Methods. (2014). doi:10.1038/nmeth.2772.

[55] M.D. Luecken, F.J. Theis, Current best practices in single-cell RNA-seq analysis: a tutorial, Mol. Syst. Biol. (2019). doi:10.15252/msb.20188746.

[56] O. Stegle, S.A. Teichmann, J.C. Marioni, Computational and analytical challenges in single-cell transcriptomics, Nat. Rev. Genet. (2015). doi:10.1038/nrg3833.

[57] T. Ilicic, J.K. Kim, A.A. Kolodziejczyk, F.O. Bagger, D.J. McCarthy, J.C. Marioni, S.A. Teichmann, Classification of low quality cells from single-cell RNA-seq data, Genome Biol. (2016). doi:10.1186/s13059-016-0888-1.

[58] P. Brennecke, S. Anders, J.K. Kim, A.A. Kołodziejczyk, X. Zhang, V. Proserpio, B. Baying, V. Benes, S.A. Teichmann, J.C. Marioni, M.G. Heisler, Accounting for technical noise in single-cell RNA-seq experiments, Nat. Methods. (2013). doi:10.1038/nmeth.2645.

[59] Chapter 6 Quality Control | Orchestrating Single-Cell Analysis with Bioconductor, (s.d.). https://osca.bioconductor.org/quality-control.html (accedit 6 agost 2020).

[60] C.A. Vallejos, D. Risso, A. Scialdone, S. Dudoit, J.C. Marioni, Normalizing single-cell RNA sequencing data: Challenges and opportunities, Nat. Methods. (2017). doi:10.1038/nmeth.4292.

[61] L. Wen, F. Tang, How to catch rare cell types, Nature. (2015). doi:10.1038/nature15204.

[62] C. Hafemeister, R. Satija, Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression, Genome Biol. (2019). doi:10.1186/s13059-019-1874-1.

[63] R. Bacher, L.F. Chu, N. Leng, A.P. Gasch, J.A. Thomson, R.M. Stewart, M. Newton, C. Kendziorski, SCnorm: Robust normalization of single-cell RNA-seq data, Nat. Methods. (2017). doi:10.1038/nmeth.4263.

[64] W.V. Li, J.J. Li, An accurate and robust imputation method scImpute for single-cell RNA-seq data, Nat. Commun. (2018). doi:10.1038/s41467-018-03405-7.

[65] D. van Dijk, R. Sharma, J. Nainys, K. Yim, P. Kathail, A.J. Carr, C. Burdziak, K.R. Moon, C.L. Chaffer, D. Pattabiraman, B. Bierie, L. Mazutis, G. Wolf, S. Krishnaswamy, D. Pe'er, Recovering Gene Interactions from Single-Cell Data Using Data Diffusion, Cell. (2018). doi:10.1016/j.cell.2018.05.061.

Journal Pre-proof

[66] W. Gong, I.Y. Kwak, P. Pota, N. Koyano-Nakagawa, D.J. Garry, DrImpute: Imputing dropout events in single cell RNA sequencing data, BMC Bioinformatics. (2018). doi:10.1186/s12859-018-2226-y.

[67] M. Huang, J. Wang, E. Torre, H. Dueck, S. Shaffer, R. Bonasio, J.I. Murray, A. Raj, M. Li, N.R. Zhang, SAVER: Gene expression recovery for single-cell RNA sequencing, Nat. Methods. (2018). doi:10.1038/s41592-018-0033-z.

[68] S. Tracy, G.C. Yuan, R. Dries, RESCUE: Imputing dropout events in single-cell RNA-sequencing data, BMC Bioinformatics. (2019). doi:10.1186/s12859-019-2977-0.

[69] T. Peng, Q. Zhu, P. Yin, K. Tan, SCRABBLE: Single-cell RNA-seq imputation constrained by bulk RNA-seq data, Genome Biol. (2019). doi:10.1186/s13059-019-1681-8.

[70] P. Ye, W. Ye, C. Ye, S. Li, L. Ye, G. Ji, X. Wu, scHinter: imputing dropout events for single-cell RNA-seq data with limited sample size, Bioinformatics. (2020). doi:10.1093/bioinformatics/btz627.

[71] P. Qiu, Embracing the dropouts in single-cell RNA-seq data, bioRxiv. (2018). doi:10.1101/468025.

[72] V. Svensson, Droplet scRNA-seq is not zero-inflated, Nat. Biotechnol. (2020). doi:10.1038/s41587-019-0379-5.

[73] W. Tang, F. Bertaux, P. Thomas, C. Stefanelli, M. Saint, S. Marguerat, V. Shahrezaei, Bay Norm: Bayesian gene expression recovery, imputation and normalization for single-cell RNA-sequencing data, Bioinformatics. (2020). doi:10.1093/bioinformatics/btz726.

[74] A.T.L. Lun, J.C. Marioni, Overcoming confounding plate effects in differential expression analyses of single-cell RNA-seq data, Biostatistics. (2017). doi:10.1093/biostatistics/kxw055.

[75] M.B. Cole, D. Risso, A. Wagner, D. DeTomaso, J. Ngai, E. Purdom, S. Dudoit, N. Yosef, Performance Assessment and Selection of Normalization Procedures for Single-Cell RNA-Seq, Cell Syst. (2019). doi:10.1016/j.cels.2019.03.010.

[76] P.Y. Tung, J.D. Blischak, C.J. Hsiao, D.A. Knowles, J.E. Burnett, J.K. Pritchard, Y. Gilad, Batch effects and the effective design of single-cell gene expression studies, Sci. Rep. (2017). doi:10.1038/srep39921.

[77] W. Chen, S. Zhang, J. Williams, B. Ju, B. Shaner, J. Easton, J. Wu, X. Chen, A comparison of methods accounting for batch effects in differential expression analysis of UMI count based single cell RNA sequencing, Comput. Struct. Biotechnol. J. (2020). doi:10.1016/j.csbj.2020.03.026.

[78] R. Satija, J.A. Farrell, D. Gennert, A.F. Schier, A. Regev, Spatial reconstruction of single-cell gene expression data, Nat. Biotechnol. (2015). doi:10.1038/nbt.3192.

[79] F.A. Wolf, P. Angerer, F.J. Theis, SCANPY: Large-scale single-cell gene expression data analysis, Genome Biol. (2018). doi:10.1186/s13059-017-1382-0.

[80] S.R. Tyler, P.G. Rotti, X. Sun, Y. Yi, W. Xie, M.C. Winter, M.J. Flamme-Wiese, B.A. Tucker, R.F. Mullins, A.W. Norris, J.F. Engelhardt, PyMINEr Finds Gene and Autocrine-Paracrine Networks from Human Islet scRNA-Seq, Cell Rep. (2019). doi:10.1016/j.celrep.2019.01.063.

[81] V. Petukhov, J. Guo, N. Baryawno, N. Severe, D.T. Scadden, M.G. Samsonova, P. V. Kharchenko, dropEst: Pipeline for accurate estimation of molecular counts in droplet-based single-cell RNA-seq experiments, Genome Biol. (2018). doi:10.1186/s13059-018-1449-6.

[82] R. Hillje, P.G. Pelicci, L. Luzi, Cerebro: interactive visualization of scRNA-seq data, Bioinformatics. (2019). doi:10.1093/bioinformatics/btz877.

[83] K. Rue-Albrecht, F. Marini, C. Soneson, A.T.L. Lun, iSEE: Interactive SummarizedExperiment Explorer, F1000Research. (2018). doi:10.12688/f1000research.14966.1.

[84] C. Arisdakessian, O. Poirion, B. Yunits, X. Zhu, L.X. Garmire, DeepImpute: An accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data, Genome Biol. (2019). doi:10.1186/s13059-019-1837-6.

[85] T. Wang, T.S. Johnson, W. Shao, Z. Lu, B.R. Helm, J. Zhang, K. Huang, BERMUDA: A novel deep transfer learning method for single-cell RNA sequencing batch correction reveals hidden high-resolution cellular subtypes, Genome Biol. (2019). doi:10.1186/s13059-019-1764-6.

[86] L. Haghverdi, A.T.L. Lun, M.D. Morgan, J.C. Marioni, Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors, Nat. Biotechnol. (2018). doi:10.1038/nbt.4091.

[87] E. Lin, S. Mukherjee, S. Kannan, A deep adversarial variational autoencoder model for dimensionality reduction in single-cell RNA

sequencing analysis, BMC Bioinformatics. 21 (2020) 64. doi:10.1186/s12859-020-3401-5.

[88] M. Amodio, D. van Dijk, K. Srinivasan, W.S. Chen, H. Mohsen, K.R. Moon, A. Campbell, Y. Zhao, X. Wang, M. Venkataswamy, A. Desai, V. Ravi, P. Kumar, R. Montgomery, G. Wolf, S. Krishnaswamy, Exploring single-cell data with deep multitasking neural networks, Nat. Methods. (2019). doi:10.1038/s41592-019-0576-7.

[89] V.Y. Kiselev, A. Yiu, M. Hemberg, Scmap: Projection of single-cell RNA-seq data across data sets, Nat. Methods. (2018). doi:10.1038/nmeth.4644.

[90] H. Todorov, R. Cannoodt, W. Saelens, Y. Saeys, TinGa: fast and flexible trajectory inference with Growing Neural Gas, Bioinformatics. (2020). doi:10.1093/bioinformatics/btaa463.

[91] Z. Liu, H. Lou, K. Xie, H. Wang, N. Chen, O.M. Aparicio, M.Q. Zhang, R. Jiang, T. Chen, Reconstructing cell cycle pseudo time-series via single-cell transcriptome data, Nat. Commun. (2017). doi:10.1038/s41467-017-00039-z.

[92] R. Petegrosso, Z. Li, R. Kuang, Machine learning and statistical methods for clustering single-cell RNA-sequencing data, Brief. Bioinform. (2019). doi:10.1093/bib/bbz063.

[93] P. Lin, M. Troup, J.W.K. Ho, CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data, Genome Biol. (2017). doi:10.1186/s13059-017-1188-0.

[94] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, R. Satija, Integrating single-cell transcriptomic data across different conditions, technologies, and species, Nat. Biotechnol. (2018). doi:10.1038/nbt.4096.

[95] L. McInnes, J. Healy, N. Saul, L. Großberger, UMAP: Uniform Manifold Approximation and Projection, J. Open Source Softw. (2018). doi:10.21105/joss.00861.

[96] E. Pierson, C. Yau, ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis, Genome Biol. (2015). doi:10.1186/s13059-015-0805-z.

[97] D. Risso, F. Perraudeau, S. Gribkova, S. Dudoit, J.P. Vert, A general and flexible method for signal extraction from single-cell RNA-seq data, Nat. Commun. (2018). doi:10.1038/s41467-017-02554-5.

[98] Data Portal – Human Cell Atlas, (s.d.). https://www.humancellatlas.org/data-portal/ (accedit 18 març 2020).

[99] Single Cell Portal, (s.d.). https://singlecell.broadinstitute.org/single_cell (accedit 18 març 2020).

[100] Home < Single Cell Expression Atlas < EMBL-EBI, (s.d.). https://www.ebi.ac.uk/gxa/sc/home (accedit 18 març 2020).

[101] Samples | PanglaoDB, (s.d.). https://panglaodb.se/samples.html?species=human&protocol=all protocols&sort=mostrecent (accedit 18 març 2020).

[102] M. Eugenio, R.L. Karp, G. Guo, P. Robson, A.H. Hart, L. Trippa, G.C. Yuan, Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape, Proc. Natl. Acad. Sci. U. S. A. (2014). doi:10.1073/pnas.1408993111.

[103] L. Yang, J. Liu, Q. Lu, A.D. Riggs, X. Wu, SAIC: An iterative clustering approach for analysis of single cell RNA-seq data, BMC Genomics. (2017). doi:10.1186/s12864-017-4019-5.

[104] H. Zhang, C.A.A. Lee, Z. Li, J.R. Garbe, C.R. Eide, R. Petegrosso, R. Kuang, J. Tolar, A multitask clustering approach for single-cell RNA-seq analysis in Recessive Dystrophic Epidermolysis Bullosa, PLoS Comput. Biol. (2018). doi:10.1371/journal.pcbi.1006053.

[105] D.A. duVerle, S. Yotsukura, S. Nomura, H. Aburatani, K. Tsuda, CellTree: An R/bioconductor package to infer the hierarchical structure of cell populations from single-cell RNA-seq data, BMC Bioinformatics. (2016). doi:10.1186/s12859-016-1175-6.

[106] J.M. Zhang, J. Fan, H.C. Fan, D. Rosenfeld, D.N. Tse, An interpretable framework for clustering single-cell RNA-Seq datasets, BMC Bioinformatics. (2018). doi:10.1186/s12859-018-2092-7.

[107] V.A. Traag, L. Waltman, N.J. van Eck, From Louvain to Leiden: guaranteeing well-connected communities, Sci. Rep. (2019). doi:10.1038/s41598-019-41695-z.

[108] C. Xu, Z. Su, Identification of cell types from single-cell transcriptomes using a novel clustering method, Bioinformatics. (2015). doi:10.1093/bioinformatics/btv088.

[109] E. Azizi, A.J. Carr, G. Plitas, A.E. Cornish, C. Konopacki, S. Prabhakaran, J. Nainys, K. Wu, V. Kiseliovas, M. Setty, K. Choi, R.M. Fromme, P. Dao, P.T. McKenney, R.C. Wasti, K. Kadaveru, L. Mazutis, A.Y. Rudensky, D. Pe'er, Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment, Cell. (2018). doi:10.1016/j.cell.2018.05.060.

Journal Pre-proof

[110] Z. Ji, H. Ji, TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis, Nucleic Acids Res. (2016). doi:10.1093/nar/gkw430.

[111] X. Qiu, Q. Mao, Y. Tang, L. Wang, R. Chawla, H.A. Pliner, C. Trapnell, Reversed graph embedding resolves complex single-cell trajectories, Nat. Methods. (2017). doi:10.1038/nmeth.4402.

[112] T. Tian, J. Wan, Q. Song, Z. Wei, Clustering single-cell RNA-seq data with a model-based deep learning approach, Nat. Mach. Intell. (2019). doi:10.1038/s42256-019-0037-0.

[113] Y. Yang, R. Huh, H.W. Culpepper, Y. Lin, M.I. Love, Y. Li, SAFE-clustering: Single-cell Aggregated (from Ensemble) clustering for single-cell RNA-seq data, Bioinformatics. (2019). doi:10.1093/bioinformatics/bty793.

[114] A.W. Zhang, C.O. Flanagan, E.A. Chavez, J.L.P. Lim, N. Ceglia, A. Mcpherson, M. Wiens, P. Walters, T. Chan, B. Hewitson, D. Lai, A. Mottok, C. Sarkozy, L. Chong, T. Aoki, X. Wang, A.P. Weng, J.N. Mcalpine, S. Aparicio, C. Steidl, K.R. Campbell, S.P. Shah, RNA-seq for tumor microenvironment profiling, Nat. Methods. (2019). doi:10.1038/s41592-019-0529-1.

[115] G. Finak, A. McDavid, M. Yajima, J. Deng, V. Gersuk, A.K. Shalek, C.K. Slichter, H. W. Miller, M.J. McElrath, M. Prlic, P.S. Linsley, R. Gottardo, MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data, Genome Biol. (2015). doi:10.1186/s13059-015-0844-5.

[116] P. V. Kharchenko, L. Silberstein, D.T. Scadden, Bayesian approach to single-cell differential expression analysis, Nat. Methods. (2014). doi:10.1038/nmeth.2967.

[117] Y. Ma, S. Sun, X. Shang, E.T. Keller, M. Chen, X. Zhou, Integrative differential expression and gene set enrichment analysis using summary statistics for scRNA-seq studies, Nat. Commun. (2020). doi:10.1038/s41467-020-15298-6.

[118] H.A. Pliner, J. Shendure, C. Trapnell, Supervised classification enables rapid annotation of cell atlases, Nat. Methods. (2019). doi:10.1038/s41592-019-0535-3.

[119] X. Shao, J. Liao, X. Lu, R. Xue, N. Ai, X. Fan, scCATCH: Automatic Annotation on Cell Types of Clusters from Single-Cell RNA Sequencing Data, iScience. (2020). doi:10.1016/j.isci.2020.100882.

[120] L. Jiang, Rare cell type detection, en: Methods Mol. Biol. 2019. doi:10.1007/978-1-4939-9057-3_5.

[121] W. Saelens, R. Cannoodt, H. Todorov, Y. Saeys, A comparison of single-cell trajectory inference methods, Nat. Biotechnol. (2019). doi:10.1038/s41587-019-0071-9.

[122] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N.J. Lennon, K.J. Livak, T.S. Mikkelsen, J.L. Rinn, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells, Nat. Biotechnol. (2014). doi:10.1038/nbt.2859.

[123] M. Setty, M.D. Tadmor, S. Reich-Zeliger, O. Angel, T.M. Salame, P. Kathail, K. Choi, S. Bendall, N. Friedman, D. Pe'Er, Wishbone identifies bifurcating developmental trajectories from single-cell data, Nat. Biotechnol. (2016). doi:10.1038/nbt.3569.

[124] G. Giecold, E. Marco, S.P. Garcia, L. Trippa, G.C. Yuan, Robust lineage reconstruction from high-dimensional single-cell data, Nucleic Acids Res. (2016). doi:10.1093/nar/gkw452.

[125] J. Chen, L. Rénia, F. Ginhoux, Constructing cell lineages from single-cell transcriptomes, Mol. Aspects Med. (2018). doi:10.1016/j.mam.2017.10.004.

[126] K. Campbell, C. Yau, Ouija: Incorporating prior knowledge in single-cell trajectory learning using Bayesian nonlinear factor analysis, bioRxiv. (2016). doi:10.1101/060442.

[127] R. Gonzalo Parra, N. Papadopoulos, L. Ahumada-Arranz, J. El Kholtei, N. Mottelson, Y. Horokhovsky, B. Treutlein, J. Soeding, Reconstructing complex lineage trees from scRNA-seq data using MERLoT, Nucleic Acids Res. (2019). doi:10.1093/nar/gkz706.

[128] A. Scialdone, K.N. Natarajan, L.R. Saraiva, V. Proserpio, S.A. Teichmann, O. Stegle, J.C. Marioni, F. Buettner, Computational assignment of cell-cycle stage from single-cell transcriptome data, Methods. (2015). doi:10.1016/j.ymeth.2015.06.021.

[129] H. Matsumoto, H. Kiryu, C. Furusawa, M.S.H. Ko, S.B.H. Ko, N. Gouda, T. Hayashi, I. Nikaido, SCODE: An efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation, Bioinformatics. 33 (2017) 2314-2321. doi:10.1093/bioinformatics/btx194.

[130] T.N. Tran, G. Bader, Tempora: cell trajectory inference using time-series single-cell RNA sequencing data, bioRxiv. (2019). doi:10.1101/846907.

Journal Pre-proof

[131] T. Wang, B. Li, C.E. Nelson, S. Nabavi, Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data, BMC Bioinformatics. (2019). doi:10.1186/s12859-019-2599-6.

[132] K. Van den Berge, H. Roux de Bézieux, K. Street, W. Saelens, R. Cannoodt, Y. Saeys, S. Dudoit, L. Clement, Trajectory-based differential expression analysis for single-cell sequencing data, Nat. Commun. (2020). doi:10.1038/s41467-020-14766-3.

[133] S. Aibar, C.B. González-Blas, T. Moerman, V.A. Huynh-Thu, H. Imrichova, G. Hulselmans, F. Rambow, J.C. Marine, P. Geurts, J. Aerts, J. Van Den Oord, Z.K. Atak, J. Wouters, S. Aerts, SCENIC: Single-cell regulatory network inference and clustering, Nat. Methods. (2017). doi:10.1038/nmeth.4463.

[134] T. Turki, Y. h. Taguchi, SCGRNs: Novel supervised inference of single-cell gene regulatory networks of complex diseases, Comput. Biol. Med. (2020). doi:10.1016/j.compbiomed.2020.103656.

[135] A. Pratapa, A.P. Jalihal, J.N. Law, A. Bharadwaj, T.M. Murali, Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data, Nat. Methods. (2020). doi:10.1038/s41592-019-0690-6.

[136] X. Qiu, A. Rahimzamani, L. Wang, B. Ren, Q. Mao, T. Durham, J.L. McFaline-Figueroa, L. Saunders, C. Trapnell, S. Kannan, Inferring Causal Gene Regulatory Networks from Coupled Single-Cell Expression Dynamics Using Scribe, Cell Syst. (2020). doi:10.1016/j.cels.2020.02.003.

[137] P. Savas, B. Virassamy, C. Ye, A. Salim, C.P. Mintoff, F. Caramia, R. Salgado, D.J. Byrne, Z.L. Teo, S. Dushyanthen, A. Byrne, L. Wein, S.J. Luen, C. Poliness, S.S. Nightingale, A.S. Skandarajah, D.E. Gyorki, C.M. Thornton, P.A. Beavis, S.B. Fox, P.K. Darcy, T.P. Speed, L.K. MacKay, P.J. Neeson, S. Loi, Single-cell profiling of breast cancer T cells reveals a tissue-resident memory subset associated with improved prognosis, Nat. Med. (2018). doi:10.1038/s41591-018-0078-7.

[138] E. Papalexi, R. Satija, Single-cell RNA sequencing to explore immune cell heterogeneity, Nat. Rev. Immunol. (2018). doi:10.1038/nri.2017.76.

[139] X. Yu, Y.A. Chen, J.R. Conejo-Garcia, C.H. Chung, X. Wang, Estimation of immune cell content in tumor using single-cell RNA-seq reference data, BMC Cancer. (2019). doi:10.1186/s12885-019-5927-5.

[140] A.L. Roy, Transcriptional Regulation in the Immune System: One Cell at a Time, Front. Immunol. 10 (2019) 1355. doi:10.3389/fimmu.2019.01355.

[141] M.L. Suvà, I. Tirosh, Single-Cell RNA Sequencing in Cancer: Lessons Learned and Emerging Challenges, Mol. Cell. (2019). doi:10.1016/j.molcel.2019.05.003.

[142] R. Hou, E. Denisenko, H.T. Ong, J.A. Ramilowski, A.R.R. Forrest, Predicting cell-to-cell communication networks using NATMI, Nat. Commun. (2020). doi:10.1038/s41467-020-14887-x.

[143] I.C. Macaulay, V. Svensson, C. Labalette, L. Ferreira, F. Hamey, T. Voet, S.A. Teichmann, A. Cvejic, Single-Cell RNA-Sequencing Reveals a Continuous Spectrum of Differentiation in Hematopoietic Cells, Cell Rep. (2016). doi:10.1016/j.celrep.2015.12.082.

[144] K.S. Yan, C.Y. Janda, J. Chang, G.X.Y. Zheng, K.A. Larkin, V.C. Luca, L.A. Chia, A.T. Mah, A. Han, J.M. Terry, A. Ootani, K. Roelf, M. Lee, J. Yuan, X. Li, C.R. Bolen, J. Wilhelmy, P.S. Davies, H. Ueno, R.J. Von Furstenberg, P. Belgrader, S.B. Ziraldo, H. Ordonez, S.J. Henning, M.H. Wong, M.P. Snyder, I.L. Weissman, A.J. Hsueh, T.S. Mikkelsen, K.C. Garcia, C.J. Kuo, Non-equivalence of Wnt and R-spondin ligands during Lgr5 + intestinal stem-cell self-renewal, Nature. (2017). doi:10.1038/nature22313.

[145] R.M.J. Genga, E.M. Kernfeld, K.M. Parsi, T.J. Parsons, M.J. Ziller, R. Maehr, Single-Cell RNA-Sequencing-Based CRISPRi Screening Resolves Molecular Drivers of Early Human Endoderm Development, Cell Rep. (2019). doi:10.1016/j.celrep.2019.03.076.

[146] S. Darmanis, S.A. Sloan, Y. Zhang, M. Enge, C. Caneda, L.M. Shuer, M.G.H. Gephart, B.A. Barres, S.R. Quake, A survey of human brain transcriptome diversity at the single cell level, Proc. Natl. Acad. Sci. U. S. A. (2015). doi:10.1073/pnas.1507125112.

[147] Q. Mu, Y. Chen, J. Wang, Deciphering Brain Complexity Using Single-cell Sequencing, Genomics, Proteomics Bioinforma. (2019). doi:10.1016/j.gpb.2018.07.007.

[148] A.C. Tolonen, R.J. Xavier, Dissecting the human microbiome with single-cell genomics, Genome Med. (2017). doi:10.1186/s13073-017-0448-7.

[149] P.M. Strzelecka, A.M. Ranzoni, A. Cvejic, Dissecting human disease with single-cell omics: Application in model systems and in the clinic, DMM Dis. Model. Mech. (2018). doi:10.1242/dmm.036525.

[150] T. V. Lanz, A.K. Pröbstel, I. Mildenberger, M. Platten, L. Schirmer, Single-cell high-throughput technologies in cerebrospinal fluid research and diagnostics, Front. Immunol. (2019). doi:10.3389/fimmu.2019.01302.

Journal Pre-proof

[151] E. Der, H. Suryawanshi, P. Morozov, M. Kustagi, B. Goilav, S. Ranabathou, P. Izmirly, R. Clancy, H.M. Belmont, M. Koenigsberg, M. Mokrzycki, H. Rominieki, J.A. Graham, J.P. Rocca, N. Bornkamp, N. Jordan, E. Schulte, M. Wu, J. Pullman, K. Slowikowski, S. Raychaudhuri, J. Guthridge, J. James, J. Buyon, T. Tuschl, C. Putterman, J. Anolik, W. Apruzzese, A. Arazi, C. Berthier, M. Brenner, J. Buyon, R. Clancy, S. Connery, M. Cunningham, M. Dall'Era, A. Davidson, E. Der, A. Fava, C. Fonseka, R. Furie, D. Goldman, R. Gupta, J. Guthridge, N. Hacohen, D. Hildeman, P. Hoover, R. Hsu, J. James, R. Kado, K. Kalunian, D. Kamen, M. Kretzler, H. Maecker, E. Massarotti, W. McCune, M. McMahon, M. Park, F. Payan-Schober, W. Pendergraft, M. Petri, M. Pichavant, C. Putterman, D. Rao, S. Raychaudhuri, K. Slowikowski, H. Suryawanshi, T. Tuschl, P. Utz, D. Waguespack, D. Wofsy, F. Zhang, Tubular cell and keratinocyte single-cell transcriptomics applied to lupus nephritis reveal type I IFN and fibrosis relevant pathways, Nat. Immunol. (2019). doi:10.1038/s41590-019-0386-1.

[152] D.R. Gawel, J. Serra-Musach, S. Lilja, J. Aagesen, A. Arenas, B. Asking, M. Bengnér, J. Björkander, S. Biggs, J. Ernerudh, H. Hjortswang, J.E. Karlsson, M. Köpsen, E.J. Lee, A. Lentini, X. Li, M. Magnusson, D. Martínez-Enguita, A. Matussek, C.E. Nestor, S. Schäfer, O. Seifert, C. Sonmez, H. Stjernman, A. Tjärnberg, S. Wu, K. Åkesson, A.K. Shalek, M. Stenmarker, H. Zhang, M. Gustafsson, M. Benson, A validated single-cell-based strategy to identify diagnostic and therapeutic targets in complex diseases, Genome Med. (2019). doi:10.1186/s13073-019-0657-3.

[153] A.K. Shalek, M. Benson, Single-cell analyses to tailor treatments, Sci. Transl. Med. (2017). doi:10.1126/scitranslmed.aan4730.

[154] A.J. Wilk, A. Rustagi, N.Q. Zhao, J. Roque, G.J. Martínez-Colón, J.L. McKechnie, G.T. Ivison, T. Ranganath, R. Vergara, T. Hollis, L.J. Simpson, P. Grant, A. Subramanian, A.J. Rogers, C.A. Blish, A single-cell atlas of the peripheral immune response in patients with severe COVID-19, Nat. Med. 26 (2020) 1070-1076. doi:10.1038/s41591-020-0944-y.

Graphical abstract

Highlights

- Advancements and applications of Single-cell transcriptomics are described
- Stepwise experimental and data analysis workflow are discussed
- Development of new strategies for scRNA-seq is an active area of research
- Computational and machine learning approaches for data analysis are reviewed

Journal Pre-proof

A Hybrid Grey Wolf-Whale Optimization Algorithm for Optimizing SVM in Breast Cancer Diagnosis

Indu Singh
Department of Computer Science Engineering, Delhi Technological University
New Delhi, India
Email: indusingh@dtu.ac.in

Raghav Bansal
Department of Electronics and Communication Engineering, Sardar Vallabhbhai National Institute of Technology
Surat, India
Email: raghav12bansal@gmail.com

Anchal Gupta
Department of Computer Science Engineering, Delhi Technological University
New Delhi, India
Email: anchalgupta.cse@gmail.com

Ayush Singh
Department of Computer Science Engineering, Delhi Technological University
New Delhi, India
Email: ayushsinghayush5@gmail.com

2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC) | 978-1-7281-7132-6/20/$31.00 ©2020 IEEE | DOI: 10.1109/PDGC50313.2020.9315816

***Abstract*—Breast cancer is one of the major causes of death among women in many parts of the world but, it is confirmed that early detection and accurate diagnosis of this disease ensures the long survival of the patient. In our research, a hybrid meta-heuristic swarm intelligence based Support Vector Machine ClassifiercalledGrey Wolf-Whale Optimization Algorithm with Support Vector Machine (GWWOA-SVM) is proposed for early-stage detection of the disease. Anamalgamation of Whale Optimization Algorithm (WOA) and Grey Wolf Optimization (GWO) is used for tuning the hyper-parameters of SVM. The effectiveness of GWWOA-SVM is evaluated against the Wisconsin Breast Cancer Dataset (WDBC). The performance of the proposed model is evaluated on various metrics such as accuracy, precision, recall, specificity, and F1 Score. Our model achieves a classification accuracy of 97.721% for WDBC dataset. The GWWOA algorithm is compared with GWO and WOA based on their extent of optimization of the fitness function. The comparison of breast cancer diagnosiscorroborates that ourmodel evaluatesmore nuanced performance metrics.**

Keywords—Breast cancer detection,Whale Optimization Algorithm, Grey Wolf Optimization, Support Vector Machine, Hyper-parameter tuning.

I. INTRODUCTION

The rise and escalation of atypical cells originating in the breast tissue are referred to as Breast cancer. Considering the World Health Organization (WHO) cancer statistics [2], breast cancer affects 2.1 million women per year and is a primary reason for deaths related to cancer amongst women. According to the American Cancer Society in 2019, around 334,200 new cases of breast cancer appear among men and women, and almost 42,260 patients pass away due to breast cancer bifurcated into 41,760 women and 500 men.

A lot of treatments available for breast cancer diagnosis, such as surgery, chemotherapy, radiation, and hormone therapies[1]. In the initial stages of the treatment, there are chances of recurrence of breast cancer, which might vary from cases to cases. Nevertheless, most reoccurrences happen in the first five years of cure. The most usualareas for recurrence of the cancer are bones, lungs, liver, brain, and so forth outside the breast region. The main topic of this research is a hybrid meta-heuristic swarm intelligence based Support Vector Machine Classifier Grey Wolf-Whale Optimization Algorithm with Support Vector Machine(GWWOA-SVM). GWWOA algorithm is a hybrid of Whale Optimization and Grey Wolf optimization algorithm. In our proposed GWWOA algorithm, we have updated the value of $\vec{a}$ for better performance. Also, different performance evaluation parameters are evaluated, which are classification accuracy, precision, recall, specificity, F1 Score.

Thepaper'sorganization is described below. Section II is for related works. The background for algorithms is given in Section III. In Section IV, the proposed GWWOA algorithm and fitness function is described. In Section V, the performance of the proposed algorithms is evaluated and investigated. The conclusion is presented in Section VI. Finally, references are given in Section VII.

II. RELATED WORK

Different models and algorithms have been created on the characterization of breast cancer.M. R. Krishnan et al. [3] have utilizeda Gaussian radial basis function and polynomial kernel function. Execution of the classifier was assessed on various metrics like accuracy, confusion matrix, sensitivity, and specificity. In radial basis function, the most note-worthyaccuracy was 92.105%. For the polynomial kernel function, the 4th-degree polynomial function gives the best accuracy as 92.627%.

A. Christobel et al. [4] juxtaposed the performance of NaïveBayes , C4.5, Support Vector Machine and K- Nearest Neighbor (KNN)for finding the best classifier and SVM ended up being precisewith an accuracy of 96.99%. T. Joachims [5] achieved atan accuracy of 95.06% with the help of fuzzy neuron techniques.

A. Rojas-Dominguez et al.[6] used BAT algorithm, Firefly algorithm, Fruit-firefly algorithm, Particle swarm optimization for choosing the optimal parameters of SVM. They evaluated accuracy for different datasets such as Breast Cancer Wisconsin Diagnostic, Breast Cancer Wisconsin Original, and several other datasets.N.Khuriwal et al. [7] presented a new method called adaptive ensemble voting

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on February 11,2021 at 06:13:35 UTC from IEEE Xplore. Restrictions apply.

method for breast cancer diagnosis on Wisconsin breast cancer dataset. The algorithm was used for feature selection, and only 16 features were used for Classification.

X. Yin et al. [8] introduced a new parameter for SVM optimization model by using advanced Whale Optimization Algorithm (AWOA) with external archiving strategy. AWOA solved the traditional fast convergence problem of WOA and also solved the problem of local minima.Q. Al-Tashi et al. [9] proposed a binary version of the hybridparticle swarm optimization and grey wolf optimization for feature selection. The wrapper-based method K-nearest neighbor classifier with Euclidean separation matrix is utilized for finding the best solution.

III. BACKGROUND

A. Whale Optimization Algorithm

Whale optimization algorithm (WOA) is dependent on the bubble-net hunting technique of humpback whales. S. Mirjalili et al. [10] first developed this algorithm. WOA is a two-stage algorithm. The first stage is an exploration which includes searching prey randomly. The second stage is called the exploitation phase and is done via the spiral bubble-net attack. The working of each step of WOA isexplainedin the following subsections:

1) ***Encircling prey and attacking-bubble net mechanism***

The whale's position must be optimized before the begin of the hunt to locate its prey. The following equations can do this:

$$\vec{X}(i+1) = \vec{X}^*(i) - \vec{A} \cdot \vec{D} \quad (1)$$

$$\vec{D} = | \vec{C} \cdot \vec{X}^*(i) - \vec{X}(i) | \quad (2)$$

$$\vec{A} = 2 \cdot \vec{a} \cdot \vec{r} - \vec{a} \quad (3)$$

$$\vec{C} = 2 \cdot \vec{r} \quad (4)$$

$\vec{X}^*(i)$depicts the best position of the whale. $\vec{X}(i+1)$tells us the present position of the search agent, and $\vec{D}$depictsthe distance between prey and whale. $\vec{C}$ and $\vec{A}$ are the coefficients. The value of $\vec{a}$reduces from 2 to 0, $\vec{a}$achieves 0 when it is at the last iteration.$\vec{r}$is a random variable and its value is between 0 and 1.The value of $\vec{A}$ is assigned in the [-1,1] range.

Equation (5) calculatesthe separation between the best position and the present position of the search agent. Equation (5) also create a spiral-shaped approach. The encircling shrinking mechanism and spiral-shaped approach, both of them have a 50% probability of being selected throughout the total iterations using (7).

$$\vec{X}(i+1) = \begin{cases} \vec{X}* - \vec{A} \cdot \vec{D}, & p < 0.5 \\ exp(bk) \cdot cos(2\pi k) \cdot \vec{D}* + \vec{X}*(i), & p \geq 0.5 \end{cases} \quad (1)$$

2) ***Searching for prey***

$\vec{A}$controls the search agents i.e. whales to move away from the local whale. In this stage, the ever-changing locationof whale is dependent on the random search($X_{rand}(i)$) instead of the best position($\vec{X}^*(i)$).

$$\vec{X}(i+1) = \vec{X}_{rand} - \vec{A} \cdot \vec{D} \quad (8)$$

$$\vec{D} = | \vec{C} \cdot \vec{X}_{rand} - \vec{X} | \quad (9)$$

The next step is to calculate the fitness of every search agent.The process mentioned above advances until it comes to the best solution($X^*(i)$). After obtaining the best solution($\vec{X}^*(i)$), the coefficients variables ($\vec{C}$ and $\vec{A}$) are updated, and a random number between 0 and 1, updates the position of search agents using (2) for $\vec{D}$ and (8) or (5) for $\vec{X}(i+1)$.

B. Grey Wolf Optimization Algorithm

GWO is an algorithm inspired by grey wolves proposed by S. Mirjalili et al. [11]. Grey wolves follow the hierarchy and are arranged in 4 categories from top-to-bottom - Alpha (α) leader of the wolves, Beta (β) assisting the leader, Delta (δ) follows both previous wolves and Omega (ω) [11].

1) ***Encircling prey***

Grey wolves hunt prey by encircling it using (10) and (11):

$$\vec{X}(i+1) = \vec{X}prey(i) - \vec{A} \cdot \vec{D} \quad (10)$$

WOA Algorithm Pseudocode

```
Initialize the population of whales Xn where (n = 1, 2, 3,, m)
Evaluate the fitness function for each search agent
X* = the best search agent of whale
  While (iter < maxIterations)
    For each solution
      Update a, A, C, L, and p
      If1 (p<0.5)
        If2 (|A| < 1)
        Calculate the new location of the present search agent by (1)
        Else If2(|A| ≥ 1)
          Randomly choose search agent (X_rand)
          Change the location of the present search agent by (7)
        End If2
      Else If1 (p ≥ 0.5)
          Location of the current search agent updated by (5)
      End If1
    End For
  Return search agents to inside the search space if it goes beyond the search space
  Fitness value for each search agent is calculated
  Updated X* if there is a better solution
  iter = iter + 1
  End While
Return X*
```

where $\vec{X}prey(i)$ tells prey's location. $\vec{X}(i+1)$represents the current position of a grey wolf.The coefficient vectors$\vec{A}$and $\vec{C}$ are calculated as follows:

$$\vec{A} = 2 \cdot \vec{a} \cdot \vec{r}_1 - \vec{a} \quad (12)$$

$$\vec{C} = 2 \cdot \vec{r}_2 \quad (13)$$

In both the phases, the value of $\vec{a}$ is decreased from 2 to 0. The value of $\vec{a}$ is 0when GWO

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on February 11,2021 at 06:13:35 UTC from IEEE Xplore. Restrictions apply.

reaches the maximum iteration. The random numbers $\vec{r}_1$and $\vec{r}_2$are selected from a range between 0 and 1.

*2) **Hunting the prey***

A grey wolf hunts the best possible solution after encircling it. The Alpha wolf stores the best solution in every iteration, and if the solution is improved the same changes. Beta wolves and Delta wolves identify the position of the prey. The position of the grey wolf is updated with the help of the following equations:

$$\vec{D}_\alpha = |\vec{C}_1 \cdot \vec{X}_\alpha - \vec{X}| \quad (14)$$
$$\vec{D}_\beta = |\vec{C}_1 \cdot \vec{X}_\beta - \vec{X}|$$
$$\vec{D}_\delta = |\vec{C}_1 \cdot \vec{X}_\delta - \vec{X}|$$
$$\vec{X}_1 = \vec{X}_\alpha - \vec{A}_1 \cdot \vec{D}_\alpha \quad (15)$$
$$\vec{X}_2 = \vec{X}_\beta - \vec{A}_2 \cdot \vec{D}_\beta$$
$$\vec{X}_3 = \vec{X}_\delta - \vec{A}_3 \cdot \vec{D}_\delta$$
$$\vec{X}(i+1) = \frac{\vec{X}_1 + \vec{X}_2 + \vec{X}_3}{3} \quad (16)$$

*3) **Exploration***

Searching mechanisms are affected by Alpha, Beta, Gama and Delta. If $\vec{A}$ is more than 1 or less than -1, then the search agents bifurcate.If $\vec{A}$ is more than 1, then the wolves search agents try to search for a better location of the prey.In GWO, variable $\vec{C}$ also influences the exploration phase.

*4) **Exploitation***

The prey in motion has to be stopped for it to be attacked. The mechanism mentioned above is achieved by reducing the value of $\vec{a}$. The value of$\vec{A}$, is in turn,is abated by the value $\vec{a}$, and is in the range between -1 and 1. If $\vec{A}$is more than -1 and less than 1, then prey can be attacked.

Grey Wolf Optimization Algorithm Pseudocode

```
Grey wolf population initialization X_n (n=1,2,……,m)
Initialize variables such as a, A and C
Evaluate the fitness of search agent
X_α = the best search agent
X_β = the second-best search agent
X_δ = the third best search agent
  While (iter < maxIterations)
    For each search agent
    Update the location of the present search agent by (16)
    End For
  Update a,A and C
  Search agents evaluated by a fitness
  Update X_α , X_β , X_δ iter = iter + 1
  End While
Return X_α
```

C. Support Vector Machine

Support Vector Machine is a popular regression and classification supervised algorithm. The SVM gives the most optimal results for binary classification problems.Vapnik et al. [12] introduced SVM in the 1990s for minimizing the structural risk, which makes the model strongly resistant to over-emphasizing problem.Classification in SVM is done by creating a hyperplane. A hyperplane is defined as a partition that totallydivides the data points into their corresponding classes, as in Fig. 1.

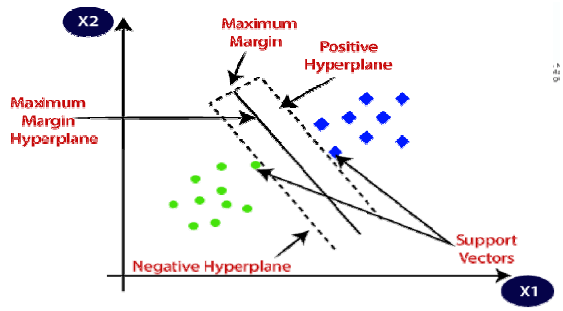


Fig. 1. SVM Hyper-Plane

D. Hyper-parameter tuning of SVM

The two main parameters affecting the performance of SVM classifier are C and γ. The parameter C controls the misclassification cost on the training data. A large value of C gives a low value of bias and a high value of variance and vice versa. γ is the parameter that explains how much the influence of a single training example reaches. A large value of γ gives low variance and high bias and vice versa. Therefore, optimal hyper-parameters (C and γ) have to be chosen for enhancing the performance of the classifier.

E. Wisconsin Breast Cancer Dataset (WDBC)

The efficiency of our proposed model is tested on Wisconsin Breast Cancer Dataset [13] from the University of Wisconsin Hospitals, Madison from Dr W. H. Wolberg. This dataset has a total of 699 samples in which 458 samples are Benign, and 241 are Malignant. The dataset has a total of 9 features, and the output labels are"M' and 'B' as Malignant and Benign respectively. The dataset has the following attributes: (1) Thickness of clump, (2) Cell size uniformity, (3) Cell shapeuniformity, (4) Marginal adhesion, (5) Single epithelial cell size, (6) Bare nuclei, (7) Bland chromatin, (8) Normal nucleoli, (9) Mitoses.

IV. PROPOSED MODEL

GWWOA Algorithm Generally, Whale Optimization Algorithm (WOA) performs well for optimization problems. However, in the exploitation phase, WOA is not effective. Because of the encircling mechanism, WOA suffers from local minima. Grey Wolf Optimization (GWO) also have good performance for optimization problems, but GWO lacks to establish a balance between the exploration and exploitation phase.

In our algorithm GWWOA, WOA is amalgamated with GWO. In GWWOA, firstly, we improved the hunting mechanism by adding a condition (16). From the (16), $\vec{A}_1$, $\vec{A}_2$, $\vec{A}_3$ have a more criticaleffect on the exploitation phase. So, we are adding one condition to the standard WOA exploitation phase to avoid local optima when $\vec{A}$ is less than 1 or greater than -1. Secondly, we are using (14), (15), and (16). In the exploitation phase, we are incorporating $\vec{A}_1$, $\vec{A}_2$, and$\vec{A}_3$.Lastly, it does not allow thesearch agent (whale) to update a position which is worse than itsearlier position.

The first step in GWWOA is the initialization of population size (consisting of both whales and wolves). After initialization, the population executes a mechanism to alter the agents, which are outside the search space. Then the value of fitness function is evaluated. After calculating fitness value, $\vec{a}$, $\vec{A}$, $\vec{C}$, L, *and* $p$are updated. The value of $\vec{a}$ has been changed in our model for better performance:

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on February 11,2021 at 06:13:35 UTC from IEEE Xplore. Restrictions apply.

$$\vec{a} = 2(1-\frac{\text{iter}^2}{\text{iter}^2_{\max}}) \quad (17)$$

Then, in the range between 0 and 1, a random number (p) is engendered.If the value of p is smaller than 0.5, then it moves to the condition in which ($|\vec{A}|<1$). If this condition comes out to be true, then the new location of the search agent is updated by the (1). Then, if the newly calculatedlocation is better than the earlierlocation, theearlierlocation is updated.

If the value of p is more than or the sameas 0.5, the new position is updated using (2). Then again, if the newly calculatedlocation is superiortothe earlier position, the earlier position will be updated. If the value of p is more than or same as 1, the new condition is taken if ($\vec{A}_1$>-1 || $\vec{A}_1$<1). If this condition is correct, then the *alphaPosition*, *betaPosition* and *deltaPosition* are calculated using (15). The new position is calculated using (16). After this step, we check either it is beyond the search space or not. After checking, a new fitness valuecomputed, and finally, the best position($X^*(i)$) is returned, which is our optimized value.

A. Fitness Function

SVM is trained using the training set, and the misclassification rate is calculated by the test set in each iteration. The rate is the number of the ratio between the number of misclassified samples (Ne) to the total number of testing samples (N) as given by (18).

$$\text{Minimize } F = \frac{N_e}{N} \quad (18)$$

B. SVM hyper- parameter tuning by GWWOA Algorithm

The WDBC dataset is divided into training and testing set.The value of $iter_{max}$ = 500, the population size is set as 30. The training set is passed on to SVM classifier, and then the SVM parameters are initialized. The initialized SVM parameters are passed on to the GWWOA algorithm for obtaining the optimal parameters. When the of the value of *iter* = 500, the GWWOA algorithm is terminated, and the parameter optimization is finished. The optimized SVM parameters are obtained, and the optimal parameters (C and γ) are passed to the SVM classifier, where the classification accuracy, precision, recall, specificity and F1 Score is calculated. Binary Classification is then performed on testing data. The above-mentioned workflow isdepicted in Fig. 2. The kernel used for SVM is "RBF".

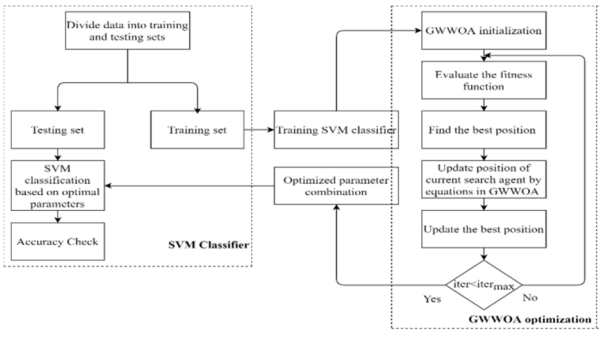


Fig. 2. The architecture of the proposed model GWWOA-SVM

V. RESULTS

All the algorithms discussed in our research are implemented in python on Windows 10 using Intel (R) Core (TM) i5-7300HQ CPU, 2.50GHz, 8GB RAM computer. This section focuses on evaluating the performance measures of our proposed GWWOA-SVM algorithm in terms of accuracy, precision, recall, specificity, and F1 Score. For our algorithm, we have divided our dataset into 70% training data and 30% testing data.

A. Evaluating the performance of the optimization algorithm in terms of fitness value

Our main objective is the minimization of the fitness function. Lower the value of the misclassification rate more will be our accuracy. For our comparison, we are comparing standard WOA, GWO, and our proposed hybrid Grey Wolf-Whale Optimized Algorithm (GWWOA).In our comparison, we ran each algorithm for 100 iterations. From Fig. 3, we observed that fitness value for GWO at the end of 100 iterations was 0.009 and fitness value obtained from WOA was also 0.009 at the end of 100 iterations. For our algorithm, the fitness value is around 0.007, which is minimum among the above mentioned three algorithms. GWWOA algorithm also achieved minimum value for less number of iterations in comparison to GWO algorithm. From the fitness value, we can conclude that hybrid GWWOA algorithm outperformed GWO and WOA algorithm.

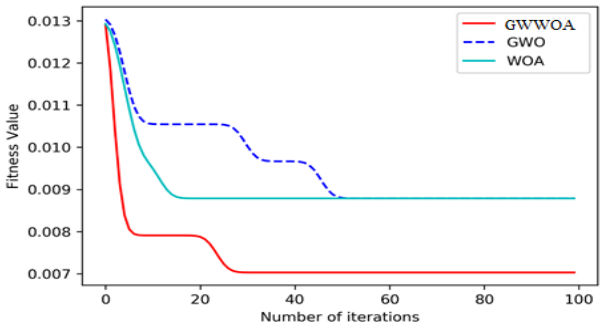


Fig. 3. Optimization of fitness function using GWWOA, GWO and WOA

B. Optimization of SVM using GWWOA Algorithm

C and γ parameter affects the performance of the SVM classifier. Optimization for ten different value of the number of iterations is performed. With these parameters, binary Classification on the Wisconsin Breast Cancer Dataset is performed. The classification accuracy is increasing with the increase in the number of iterations. As the value of *iter* increases, the extent of optimization increases, and the value of fitness function is getting minimized. The least value of fitness is obtained at *iter*= 500, as the accuracy is maximum at *iter* = 500. The optimized SVM parameters, i.e. C and γ, are obtained at *iter* =500. The maximum accuracy achieved for WDBC dataset using GWWOA-SVM algorithm is 97.721%. In Table I, we have a different value of C and γ for a different number of iterations.

TABLE I. SVM PARAMETERS

Number of iterations	*C*	γ	*Accuracy(%)*
50	4.5581	0.6807	93.859
100	4.6621	0.0581	94.736
150	8.1633	0.2545	95.614
200	3.7955	0.2286	95.892
250	7.1493	0.9305	96.215
300	5.1700	1.0	96.491
350	10.2184	0.5283	96.578
400	16.7294	0.6145	97.021
450	17.6162	0.5506	97.445
500	**15.9472**	**0.5955**	**97.721**

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on February 11,2021 at 06:13:35 UTC from IEEE Xplore. Restrictions apply.

C. Evaluation Parameters

In Table II, different evaluation parameters are calculated for different value ofthe number of iterations. From the Table I and Table II, it is evident that the highest specificity is 0.9578 when *iter= 50, C* = 4.5581, and γ = 0.6807, it implies that the classifier classifies maximum true negatives to their corresponding class.Similarly, the maximum value of precision is obtained for *iter= 500, C* = 15.9472, and γ = 0.5955, i.e., at *iter= 500*, precision computes what proportion of people predicted with breast cancer actually have cancer. For *iter = 50*, the maximum number of positive samples are found as the recall is highest—the value of F1 Score increases when the number of false-positive and false-negative decreases.Accuracy is defined as the number of correct predictions to the total number of predictions. It is also evident that accuracy is highest (97.721%), when the value of *iter* = 500, and the optimal parameters *C*= 15.9472 and γ = 0.5955.

TABLE II. PERFORMANCE EVALUATION PARAMETERS

Iterations	*Accuracy(%)*	*Precision*	*Recall*	*Specificity*	*F1 Score*
50	93.859	0.9014	0.9761	0.9302	0.9318
100	94.736	0.8814	0.9534	0.9295	0.9213
150	95.614	0.9130	0.9333	0.9420	0.9230
200	95.892	0.8742	0.9724	0.9305	0.9318
250	96.215	0.8913	0.9318	0.9285	0.9111
300	96.491	0.8659	0.9523	0.9167	0.9090
350	96.578	0.9130	0.9333	0.9420	0.9230
400	97.021	0.9325	0.9555	0.9565	0.9450
450	97.445	0.9347	0.9347	0.9558	0.9347
500	**97.721**	**0.9350**	**0.9622**	**0.9578**	**0.9542**

D. Comparison with previous studies

To exhibit the robustness of our model, the results obtained from our proposed approach are contrasted with the results of some recently created models. Some newly developed models on the Wisconsin Breast Cancer Dataset (WDBC) have been discussed in Table III. In 2016, E. Tube et al. [14] proposed an Adjusted BAT Algorithm for WDBC Dataset and achieved an accuracy of 97%. In 2017, H. A. Alhakbani et al. [15] optimized SVM parameters using Particle Swarm Optimization (PSO) and achieved an accuracy of 92.98%. In 2018, S. R. Kamel et al. [16] used SVM as a classifier on WDBC Dataset and computed accuracy of 85.89%. In 2020, M. Wang et al. [18] used WOA algorithm for optimizing SVM parameters and achieved an accuracy of 96.65%. From Table III, it is evident that our proposed model GWWOA-SVM has the highest accuracy and outperforms the previous algorithms for breast cancer diagnosis.

TABLE III. COMPARISON OF OUR MODEL WITH PREVIOUS STUDIES

Study	*Year*	*Algorithm*	*Accuracy*
E. Tuba et al.[14]	2016	ABA-SVM	97%
H. A. Alhakbani et al.[15]	2017	PSO-SVM	92.98%
S. R. Kamel et al. [16]	2018	SVM	85.89%
M. Wang et. Al [17]	2020	WOA-SVM	96.65%
Our proposed model	**2020**	**GWWOA-SVM**	**97.721%**

VI. CONCLUSION AND FUTURE WORK

In our study, we have inspected an improved hybrid of Grey Wolf Whale Optimization Algorithm hyper-parameter tuning of SVM using the swarm mentioned above algorithms.

Results have shown that the proposed model, i.e. GWWOA-SVM better at minimization of the fitness function than GWO and WOA. The experimental results have given a 97.721% classification accuracy for C = 15.9472 and γ = 0.5955.

Our future work will emphasize on a further detailed formulation of novel formulas in hybrid swarm algorithms, which can enhance the performance of the system.

VII. REFERENCES

[1] R. Lin and P.Tripuraneni,"Radiation therapy in early-stage invasive breast cancer," Indian J SurgOncol, vol.2, pp. 101–111, 2011.

[2] https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/(ACCESS.13.09.2020).

[3] M. R. Krishnan, S. Banerjee, C. Chakraborty, C. Chakraborty, and A. K. Ray, "Statistical analysis of mammographic features and its classification using support vector machine", Expert Systems with Applications 37, page: 470-478, 2010.

[4] Y. A. Christobel and P. Sivaprakasam, "An Empirical Comparison of Data Mining Classification Methods," vol. 3, no. 2, pp. 24–28, 2011.

[5] T.Joachims, "Transductive inference for text classification using support vector machines". Proceedings of international conference machine learning. Slovenia. 1999.

[6] A. Rojas-Domínguez, L. C. Padierna, J. M. Carpio Valadez, H. J. Puga-Soberanes and H. J. Fraire, "Optimal Hyper-Parameter Tuning of SVM Classifiers with Application to Medical Diagnosis," in IEEE Access, vol. 6, pp. 7164-7176, 2018.

[7] N. Khuriwal and N. Mishra, "Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm," 2018 IEEMA Engineer Infinite Conference (eTechNxT), New Delhi, 2018, pp. 1-5.

[8] X. Yin, Y. Hou, J. Yin, and C. Li, 2019 J. Phys.: Conf. Ser. 1237 022140.

[9] Q. Al-Tashi, S. J. Abdul Kadir, H. M. Rais, S. Mirjalili, and H. Alhussian, "Binary Optimization Using Hybrid Grey Wolf Optimization for Feature Selection," in IEEE Access, vol. 7, pp. 39496-39508, 2019.

[10] S. Mirjalili and A. Lewis (2016),"The whale optimization algorithm". AdvEngSoftw 95:51–67.

[11] M. M. Mafarja and S. Mirjalili (2017),"Hybrid whale optimization algorithm with simulated annealing for feature selection". Neurocomputing 206:302–312.

[12] C. Cortes and V. Vapnik(1995),"Support-vector networks. Machine Learning", 20(3): 273-297.

[13] "UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set".

[14] E. Tuba, M. Tuba and D. Simian,"Adjusted bat algorithm for tuning of support vector machine parameters". In: 2016 IEEE congress on evolutionary computation (CEC), 24–29 July 2016, pp 2225– 2232.

[15] H. A. Alhakbani and M. M. al-Rifaie,"Optimizing SVM to classify imbalanced data using Dispersive Flies Optimization," 2017 Federated Conference on Computer Science and Information Systems (FedCSIS), Prague, 2017, pp. 399-402.

[16] S. R. Kamel, R. YaghoubZadeh, and M. Kheirabadi,"Improving the performance of support-vector machine by selecting the best features by Gray Wolf algorithm to increase the accuracy of diagnosis of breast cancer". *J Big Data* 6, 90 (2019).

[17] M. Wang and H. Chen, ''Chaotic multi-swarm whale optimizer boosted support vector machine for medical diagnosis,'' Appl. Soft Computing., vol. 88, Mar. 2020, Art. no. 105946.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on February 11,2021 at 06:13:35 UTC from IEEE Xplore. Restrictions apply.

A Literature Review On Sentiment Analysis Techniques Involving Social Media Platforms

Samarth Garg
Department of Computer Science and Engineering
Delhi Technological University
Delhi, India
samarthg1299@gmail.com

Divyansh Singh Panwar
Department of Computer Science and Engineering
Delhi Technological University
Delhi, India
divyansh8june@gmail.com

Aakansha Gupta
Department of Computer Science and Engineering
Delhi Technological University
Delhi, India
aakanshagupta.74@gmail.com

Rahul Katarya
Department of Computer Science and Engineering
Delhi Technological University
Delhi, India
rahulkatarya@dtu.ac.in

2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC) | 978-1-7281-7132-6/20/$31.00 ©2020 IEEE | DOI: 10.1109/PDGC50313.2020.9315735

***Abstract*— Sentiment analysis refers to the active field of Natural language processing that extracts the attitude and emotion of a human being. With the growth of social media, more people are using online platforms such as Twitter, Facebook, YouTube, etc. to express their opinions. Twitter is considered to be the purest platform to express one's views. Mostly all personalities from diverse backgrounds use twitter. Therefore, it becomes a need of the hour to study public opinion. This provides us valuable information and helps organizations and governments to contemplate mass public opinion and take better decisions accordingly. In this review paper, an extensive and exhaustive guide to the subfield of Natural language processing (NLP), focusing precisely on sentiment analysis on twitter dataset, has been presented. It highlights three main approaches to analyze the sentiment. We have summarized and compared the approaches on different metrics opted by various researchers in the field of sentiment analysis using the twitter dataset. With so much active work in this field, this review paper would assist all future researchers.**

Keywords*— *Text classification, Sentiment Analysis, machine learning, Social media, Deep Learning, Natural Language Processing

I. INTRODUCTION

Artificial Intelligence right from its origination, has made a remarkable contribution in providing realistic and non-theoretical solutions to essential societal and human issues under a variety of different domains, including NLP, where computational and linguistics techniques are used in order to assist the computers to recognize and produce the desired results. Several noteworthy developments in the area of NLP cover sentiment analysis (SA), information retrieval (IR), emotion detection (ED), text summarization systems, and questions and answering (Q & A) systems. The Core intent of Sentiment analysis is to inspect the human language and extract significant patterns of information. It analyses the sentiments or the attitude of the user towards the subject of interest. This helps the organisation to consider public opinion and make decisions accordingly. It is used to classify sentiments as objective or subjective, positive or negative or neutral, summarizing opinions and detecting spams.

Micro blogging websites such as Twitter have emerged as a giant reservoir consisting of a vast variety of information. Users of these websites proactively participate in discussions on various current issues/topics and are freely able to express their concerns and opinions on them. According to a Forbes report, on an average 456000 tweets were made by different twitter users present all over the globe every minute in the year 2018 [1]. Hence this development of vast unstructured data available has provided data scientists and researchers with a huge opportunity of exploration in the field of sentiment analysis. According to statistics, Twitter presently has a huge 330 million active monthly user base and thereby is accessible through its website interface, SMS, and other mobile devices [2]. Twitter remains to be a better choice among researchers working in the field of sentiment analysis of text in comparison to its other popular existing counterparts. Though Facebook has a much higher user database as compared to Twitter, the user content available is mixed with photos, videos, etc, which is not of much use to researchers working in the field of sentiment text classification. Hence in this way, Twitter is viewed as the purest form of ideas and thoughts expression social media platform. Another probable reason for the widespread use of Twitter could be the regular updating of the twitter corpus with the required information only. [3]

This review paper compares and summarizes the recent work of researchers and data scientists working in the domain of SA, focusing on Twitter as a primary source of data. This paper comprises VII sections, while section II deals with the various levels of SA. Section III deals with the associated work done until now in the domain of SA in detail. Section IV briefly outlines each technique used in the research areas. Section V provides the general phases of sentiment analysis and contains the comparison table (Table-1). Finally, section VI discusses the various quality attributes used by researchers in the work, and section VII concludes the paper.

The literature search was performed for research papers published by different publishers in recent years (Fig. 1). The investigation focused on retrieving documents written in English and included a set of keywords involving the topic "Sentiment analysis on twitter dataset." Various approaches could be used to analyze the sentiment ranging from lexicon to deep-learning. With so much active research in this domain, our review paper could witness a valuable resource to future researchers.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 27,2021 at 10:20:03 UTC from IEEE Xplore. Restrictions apply.

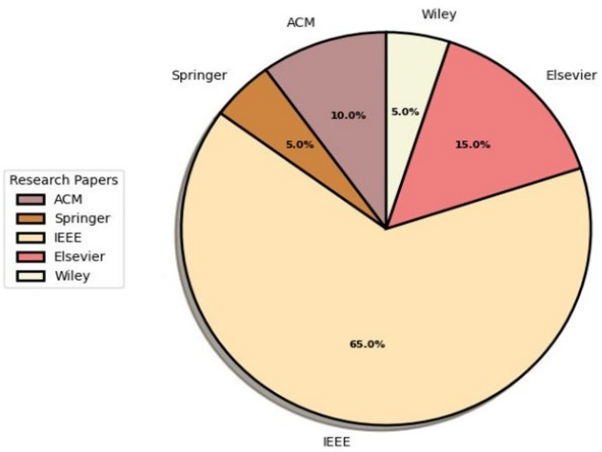


Fig. 1 Number of Published Papers in 2019-2020

II. Sentiment Analysis

Sentiment analysis is meant to inspect the emotions, attitude, and expression of a person and classify it accordingly into positive, negative, and neutral labels. Broadly speaking, sentiment analysis is subdivided into 3 main levels which are as follows:

1. Document Level Analysis: At this stage, the task is to analyze the documents sentiment as a whole. The sentiment of the whole document is then further classified as positive, negative, or neutral depending upon the text. Thus a comparative learning text can't be considered under this level.

2. Sentence Level Analysis: At this stage, the task is to analyze a particular sentence and then decide whether that sentence represents a negative, positive, or a neutral opinion. A sentence that does not show any opinion is classified as a neutral sentence.

3. Aspect Level Analysis: Both the above discussed levels are unable to evaluate one's likes and dislikes. Hence aspect level presents a throughout analysis. The fundamental key task of the entity level is identification. Earlier Aspect level was known as feature level.

III. Background

Research papers by different authors can be classified on the basis of approaches they have adopted.

Singh et al.,[4] analyzed the sentiments of people regarding the Motor Vehicle Act, 2019. The author analyzed polarity on the twitter data extracted using Twitter streaming API. The author recorded most tweets are positive even if the new rules are a burden for the public, they tend to tweet positive because of the betterment of the society [4]. This work would help the government to consider public opinion and make decisions accordingly.

On similar lines, Muntazar et al [5] correctly analyzed the tweets and predicted the opinion of the UK parliament and EU over Brexit. The author used Textblob to measure polarity[5].

V. Sindhu et al.,[6] successfully depicted the people's positive, negative and neutral remarks about a person, trend, or object using visualization techniques such as histogram, word cloud, and Pie chart. The author used Bag of Words for feature extraction and classification over manually extracted twitter data [6].

Olivier Kraaijeveld, Johannes De Smedt [7] used a cryptocurrency-specific lexicon-based sentiment analysis approach which was then integrated with bilateral Granger-causality testing to predict the price return of nine cryptocurrencies[7]. The VADER lexicon was accompanied by the addition of tokens which were obtained from the Loughran & McDonald financial corpus, 2016, which was then applied over a dataset having 24 million entries which were obtained using various twitter APIs and CoinMarketCap. Finally, a heuristic approach was used which led to the conclusion that about 1% to14% of the tweets obtained were posted on twitter by various Twitter ‘‘bot" accounts. [7]

Ansari et al., [8] in their paper analyzed the sentiment for various national political parties of India with reference to the Lok Sabha elections(2019) using classical ML classifiers with feature extraction using the TF-IDF approach. The ML classifiers used were Random Forest, LSTM, Logistic Regression, SVM, and Decision Tree. According to them, their study revealed that LSTM and Random Forest performed better on various metrics while the overall analysis revealed that the performance of SVM was the poorest amongst all. [8]

Liang, Umarani Ganesh Babu, Thomas Thorne [9] in their approach used a new ML network called Gaussian Process Dynamic Bayesian Network that integrates Gaussian Process Regression to evaluate the sentiment in the previous period at a given point of time based upon the relevant domain. This paper also builds a Sequential Monte Carlo sampler which executes Bayesian inference on a Dynamic Bayesian Network model. This helps us to infer networks showing how opinion in a particular domain can lead to a change of opinion in the other domain.[9]

Zhai Penghua, Zhang Dingyi [10], used various deep learning approaches to analyze the sentiment of the dataset obtained from SemEval-2014-Task2014 and twitter. They proposed a bidirectional GRU neural network model that embeds the attention mechanism to solve the task of aspect-level sentiment analysis. [10] The mentioned model extracted features using the GloVe technique and achieves good performance with different datasets. Authors compared the model with various state-of-the-art models such as GRU, CNN, and TC-LSTM. [10] The comparison between DL approaches could also be studied in the paper by Cheng, Song Tsai [11].

MSR Hitesh, Vedhosi, Y.J. Abhishek Kalki, Suraj Harshaand, Santoshi Kumari [2] compared various feature extraction techniques such as word2vec, TF-IDF, and Bag of words using Random Forest classifier. They found word2vec performs better than other embedding techniques [2].

The rest of all research papers are reviewed and mentioned in Table 1.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 27,2021 at 10:20:03 UTC from IEEE Xplore. Restrictions apply.

Table 1: The following table contains different techniques opted by researchers on Sentiment Analysis performed on various social media platforms

Year, Publication	Dataset used	Technique	Feature Represent-ation	Observations	Accuracy
2020,[10] (ACM)	SemEval-2014 Task2014, Twitter	GRU, CNN, TC-LSTM, ATAE-LSTM, ABAE-Bi-GRU	GloVe	• Used a bidirectional GRU neural network model that merged the attention mechanism to solve aspect-level sentiment analysis • GRU has a worst performance amongst all.	ATAE LSTM: 91% ABAE-LSTM:81.2%
2020,[1] (IEEE)	Sentiment140	Logistic Regression	LSA TF-IDF , word2vec, ELMo	• Showed that word2vec has the highest accuracy of 86.87% for feature extraction as compared to TF-IDF,ELMo, and LSA over RF classifier . • Showed the accuracy of ELMo is low when compared with Word2Vec, probably due to small dataset.	Word2vec: 77% ELMo: 73% LSA –TFIDF:71.0%
2020,[12] (Springer)	Live tweets from Twitter	SVM,KNN	N-gram	• Employed N-Gram modelling approach to design the feature extraction algorithm. • Observed that absolute competence is still not achieved.	KNN : 91.45 % SVM : 81.51%
2020,[13] (IEEE)	Twitter	Hadoop based deep RNN method		• Hadoop based deep RNN classifier model delivered maximal specificity, accuracy, sensitivity and of 0.9157, 0.9302, 0.9404 respectively.	93.02%
2020,[14] (IEEE)	Airline sentiment (Welkin10)	SVM,KNN, Random Forest (RF)		• Used K-fold cross-validation (K-10) for training and testing the available datasets. • Showed Random forest classifier as compared to other classifiers had more accuracy.	RF: 81% SVM: 74% KNN: 58%
2020,[15] (Elsevier)	Chilean Earthquake, 2010, Catalan Independence ,2017	TAN,NB, BF-TAN, SVM, Random forest	Bag of words	• Observed that SVM obtained the best performance in dataset 1 and RF in dataset 2.	SVM: 81.2% BFTAN: 76% NB: 74.2% RF: 72.5% TAN: 72.1%
2020,[16] (IEEE)	Github and kaggle (Demonetizati-on sentiment analysis)	SVM, Random Forest, Decision Tree	Bag of words	• Categorized every single feature extracted from the text data into morphological and Word N gram features. • Showed that Decision tree and random forest have more accuracy than SVM for sentiment analysis.	RF: 99.4% Decision Tree: 99.3% SVM: 91.6%
2020, [17] (IEEE)	Sentiment140 , Crowdflower Data	SentiWordNet, MNB, LR, SVM, RNN with LSTM, Ensemble models		• Investigated that Lexicon based approach were less accurate as compared to ML approach while RNN is observed to be most accurate	RNN With LSTM :82% SVM: 77.98% Ensemble 77.67% MNB: 76.52% LR: 76.44% SentiWordNet: 45%
2020,[18] (IEEE)	Live tweets: bitcoin tweets, First GOP debate, Imdb movies reviews	NB, MNB, Bernoulli classifier, LR, SGD, Linear SVC, NuSVC, LSTM, CNN		• Showed Deep learning techniques are more accurate than ML techniques or polarity based techniques	LSTM: 97% CNN LSTM : 95% MLclassifiers:82%
2020,[9] (IEEE)	9.72M manually annotated tweets	Gaussian Process Dynamic Bayesian Network		• Proposed Gaussian Process Dynamic Bayesian Network is 92% accurate to evaluate the sentiment in the previous period at a given point of time based upon the relevant domain.	92%
2020,[19] (Wiley)	Data collected from 100 twitter users with 10,275 entries.	SVM, Gaussian models: (SGM) & (MGM)		• Proposed a multivariate Gaussian to identify potential abnormal emotions of a user. • The user data was scored by joint probability density of multivariate Gaussian and was determined by a threshold. It was observed that abnormal emotion had a low joint probability density.	84.60%
2020,[20] (IEEE)	7368 manually annotated tweets & Airline-Sentiment	Feature ensemble method.		• Revealed that the proposed method based on feature ensemble and CNN models improved the performance in the sentiment analysis of tweets containing fuzzy sentiment.	F1-Score: 0.72

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 27,2021 at 10:20:03 UTC from IEEE Xplore. Restrictions apply.

2020,[7] (Elsevier)	Cryptocurrency tweets, financial data from CoinMarketCap	Lexicon-based , bi-lateral Granger-causality testing.		• Predicted the price returns of Litecoin, Bitcoin Cash and Bitcoin by applying a crypto currency-specific lexicon-based approach and bi-variate Granger-causality tests. • Discovered that bots posted 1-14% crypto currency related tweets using a heuristic approach	The polarity scores are relatively constant Mean polarity= 0.33.
2019,[21] (IEEE)	Live tweets in English only	Bagging, Lexicon approach, SVM, MAXENT, DT, RF, Naive Bayes,		• Used the approach of 4-fold cross validation for training and testing. • For both the available data's of McDonalds and KFC, Testing data of several supervised algorithms proved that Maxent (Maximum entropy) had the highest accuracy.	F-score: Bagging: 0.70 SVM: 0.67 RF:0.62 Maxent: 0.58 DT:0.57 Naive Bayes: 0.51
2019,[8] (Elsevier)	3896 manually annotated tweets	LSTM, SVM, Decision Tree, LR, RF	TF-IDF	• Analyzed the sentiment for various national political parties of India with reference to the Lok Sabha elections, 2019. • Presented that Random Forest and LSTM performed better on various metrics while the performance of SVM was found to be the poorest of all.	F-score LSTM: 0.74 RF: 0.74 DT: 0.70 LR: 0.69 SVM: 0.39
2019,[11] (ACM)	YouTube, Face book	LSTM, Bi-LSTM, GRU	word2vec, GloVe	• Used a DL-based framework to handle slang and special social language.	Bi-LSTM: 87.17% LSTM: 80.83% GRU: 64.92%
2019,[2] (IEEE)	35,000 manually annotated tweets	Random forest	word2vec, Bag of words, TF-IDF	• Showed that word2vec provides the highest accuracy of 86.87% for feature extraction as compared to TF-IDF and BOW over RF classifier.	Word2Vec: 86.8% TF-IDF: 84.4% Bag-of-Words: 83.4%
2019,[3] (IEEE)	1.1M manually annotated tweets	SentiWordNet		• Observed a net negative sentiment using lexicon-based system on the 'US ban Huawei' topic i.e. against US ban huawei. • Visualized the data using geographic vision.	Negative sentiment score = - 1641
2019,[22] (IEEE)	Live tweets	Sentiment dictionary		• A model to anticipate the boom of Indian Major Telecom Companies in terms of subscriber addition was proposed. • Indicated that the overall public incorporates a relatively stronger positive opinion concerning Reliance Jio.	Sentiment score: Jio 7530 Vodafone 3752 Airtel 2841, Idea 1353

IV. EXISTING METHODOLOGY

This section calls attention to machine learning (ML), deep learning (DL), lexicon-based approaches, and other general approaches to detect sentiments from available texts.

A. Rule construction approach[23]*:*

The rule construction approach circumscribes lexical affinity methods, and keyword recognition (KR)[23]. These methods deal with the constructive use of sentiment dictionaries or lexicons. There are many Keyword recognition dictionaries, namely the WordNet, SentiWordNet dictionaries, EmoSenticNet, the National Research Council of Canada (NRC) lexicon, and DepecheMood. The main aim is to find the occurrences of sentiment based search words in a sentence. Once such lexicon is identified in the sentence, polarity and a label are designated to that sentence. Examples of this approach are [17], [21], [7], [3] and [22] respectively.

B. Machine Learning approach:

It solves the analysis problem by categorizing the texts into various labels through ML classifiers' implementation.

- *Naive Bayes (NB):* Belongs to a family of "probabilistic classifiers" supported by applying the Bayes theorem. This is used in seen in, e.g., [15].
- *Support Vector machines (SVM):* It uses a hyper-plane in N-D space that clearly classifies the data points. Here N is the number of features. This model is used in [12], [14], [15], [16], and many more.
- *Logistic regression (LR):* Provides a method for reducing and classifying words existing in the document and special features in the same instance. E.g., in [18], [1].
- *Random Forest (RF) and Decision Trees (DT):* Both are non-linear classical supervised ML classification algorithms based upon iteratively asking partition data questions. This approach was observed in [8], [16], etc.
- *K nearest neighbour (KNN):* This algorithm is based on finding the nearest k neighbor of a decision. It is used for non-linear classification. This could be witnessed in [12], [14], respectively.
- *Maximum Entropy (Maxent):* A probabilistic classifier which depends upon the Principle of Maximum Entropy. It chooses the dataset having the largest entropy amongst all the models that fits our training data. [21]

Other approaches could also be seen in ML such as Tree augmented Naïve Bayes (TAN) [15] , Bayes Factor-TAN (BF-TAN) [15], Multinomial Naïve Bayes (MNB) [17][24], bagging [21], Bernoulli classifier, Stochastic gradient descent (SDG) , Linear SVC, NuSVC[18], Gaussian models [19], and Bayesian network [9].

C. Deep-Learning approaches:

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 27,2021 at 10:20:03 UTC from IEEE Xplore. Restrictions apply.

Deep learning is another refined derivative of machine learning, which is based upon multilayer ANN (artificial Neural network) to mimic the given tasks. The Neuron is the smallest and information processing unit of the neural network, It learns to perform numerous jobs by updating the weights between two neurons, just as working of a living brain. Deep learning models are mainly of two types: CNN [25] and RNN [13]. CNN is presented in [10], [18].There are two more refined models of RNN: (GRU)Gated Recurrent Unit [11], (BI-LSTM) Bidirectional long short-term memory [11], and (LSTM) Long Short-Term Memory [26]. Cheng et al. [11] used all the sophisticated approaches of RNN.

D. Other Techniques:

TextBlob and NLTK are Python in-built libraries for processing the data in the textual form. Textblob provides a simple Application programming interface for doing common NLP tasks such as noun phrase extraction, POS tagging, sentiment analysis, translation, classification, etc. Many authors such as Cheng et al. [11], Muntazar et al. [5], Singh et al. [4], and Sahu et al.[27], used these techniques

V. Phases Of Sentiment Analysis

The sentiment analysis includes five stages: data extraction, data pre-processing, feature extraction, sentiment classification, and finally, analysis of output.

A. Data extraction: In this initial step, various approaches to data collection can be showcased as extracting real-time tweets using twitter APIs, datasets available at multiple sources, or collected by multiple authors. Tweets extracted from Twitter may be unstructured, semi-structured, or structured type.

B. Data Pre-processing: They can be collected using different programming languages such as R or python. The input data set is transformed into a structure that is suitable for feature representation using various NLP tools. Tasks involved in the pre-processing study are:

- Removing the twitter handles (for twitter dataset)
- Removal of Punctuations, URLs, Stop words, Special characters.
- Stemming and Tokenization of dataset.

C. Feature Representation: At this stage, features are extracted to define the polarity of tweets. As observed, the authors used various techniques such as TF-IDF, N-grams, Bag of words, word2vec, and Glove.

D. Sentiment Classification: The fundamental task of sentiment analysis is to segregate the tweets depending upon their polarity. The polarity of a text may be neutral, negative or, positive. Classifiers are used to extract the sentiment of tweets.

E. Analysis of output: Evaluation of the obtained result using different metrics helps to take decisions regarding the model. It also paves the way for comparing other techniques. The result is expressed using pie charts, bar-graphs, and measured parameters.

VI. Discussion And Result Analysis

Analysis of the result is mostly used to test the accuracy of each model and verify. It also helps us to depict the results. This includes accuracy, F-score, recall and Precision, hence collectively called key performance indicators (KPIs). [14].

A. Accuracy: Is defined as the percentage of correctly classified instances.

$$Accuracy = \frac{True\ Positive + True\ Negitive}{All\ Samples} \quad -(1)$$

B. F-score: Is defined as the harmonic mean of recall and precision.

$$F - Score = 2\left(\frac{Precision\ X\ Recall}{Precision + Recall}\right) \quad -(2)$$

C. Precision: Defines the percentage of the relevancy of the obtained results.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad -(3)$$

D. Recall: Defines the percentage of total relevant results correctly classified.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negitives} \quad -(4)$$

Other ways is to visualise the results using bar-graph analysis, histograms, pie-charts, heat maps and creative approaches such as word cloud as in Qing Lim [1], geographic visualisation, e.g., [3].

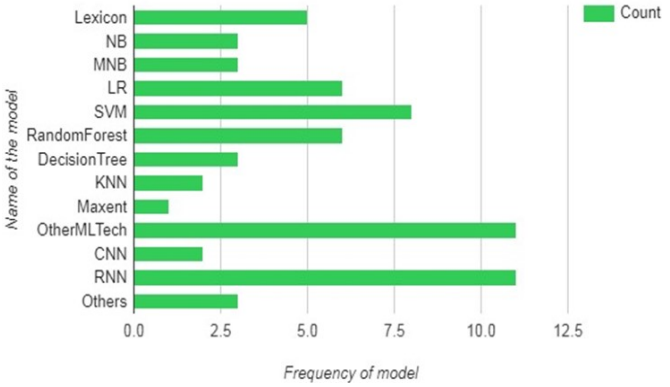


Fig.3 Model Frequency Distribution

VII. Conclusion

This paper introduces the concept of sentiment analysis, 3 main available approaches, highlights some important datasets available for opinion mining. This paper discusses the current available state-of-the-art approaches compared to the various available metrics. It is observed that the most commonly used techniques for sentiment analysis are ML techniques, specifically SVM classifier and Naive Bayes. While RNN and its derived models are most opted classifier in deep-learning *(Fig.3)*. However, amongst the various techniques, deep-learning techniques outperform machine learning techniques and lexicon based approach [17]. Chandra et al. [18] showed that the best machine-learning classifier has an accuracy of 82%, while deep-learning approaches like LSTM and CNN

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 27,2021 at 10:20:03 UTC from IEEE Xplore. Restrictions apply.

witnessed 97% and 95% respectively. Lexicon based technique performance is the poorest amongst all especially when large dataset is available[17]. Highest accuracy observed is 99.4% using Random Forest as a classifier presented in [16]. The performance of each approach heavily depends upon size and type of dataset. Future work includes areas such as crime detection i.e., mitigation by interpreting messages of victims so as to identify the menacing words, analysis of patient messages to determine his level of depression, etc. Finally, the cultural bonding of a person greatly influences his sentiment towards any scenario. Our survey paper highly relies on English literature resources. However, research on sentiment analysis in other languages can also be worked upon.

VIII. REFERENCES

[1] Y. Q. Lim, C. M. Lim, K. H. Gan, and N. H. Samsudin, "Text Sentiment Analysis on Twitter to Identify Positive or Negative Context in Addressing Inept Regulations on Social Media Platform," in 10th Symposium on Computer Applications and Industrial Electronics, IEEE, 2020.

[2] M. S. R. Hitesh, V. Vaibhav, Y. J. A. Kalki, S. H. Kamtam, and S. Kumari, "Real-time sentiment analysis of 2019 election tweets using word2vec and random forest model," in 2nd Int. Conf. on Intelligent Communication and Computational Techniques, 2019.

[3] N. H. Khun and H. A. Thant, "Visualization of Twitter Sentiment during the Period of US Banned Huawei," in 2019 International Conference on Advanced Information Technologies, ICAIT , 2019.

[4] N. A. Singh, "Sentiment analysis on motor vehicles amendment act, 2019 an initiative by government of India to follow traffic rule," in Int. Conf. on Computer Communication and Informatics, 2020.

[5] M. M. Chandio and M. Sah, "Brexit Twitter Sentiment Analysis: Changing Opinions About Brexit and UK Politicians," in International Conference on Information, Communication and Computing Technology ICICCT: Intelligent Computing Paradigm and Cutting-edge Technologies, Springer (LAIS, vol 9), 2020.

[6] V. Prakruthi, D. Sindhu, and S. Anupama Kumar, "Real Time Sentiment Analysis of Twitter Posts," in 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions, CSITSS 2018, IEEE, 2018.

[7] O. Kraaijeveld and J. De Smedt, "The predictive power of public Twitter sentiment for forecasting cryptocurrency prices," J. Int. Financ. Mark. Institutions Money, vol. 65, 2020.

[8] M. Z. Ansari, M. B. Aziz, M. O. Siddiqui, H. Mehra, and K. P. Singh, "Analysis of Political Sentiment Orientations on Twitter," in Procedia Computer Science, ScienceDirect, 2020.

[9] H. Liang, U. Ganeshbabu, and T. Thorne, "A Dynamic Bayesian Network Approach for Analysing Topic-Sentiment Evolution," IEEE Access, vol. 8, 2020.

[10] Z. Penghua and Z. Dingyi, "Bidirectional-GRU based on attention mechanism for aspect-level Sentiment Analysis," ACM Int. Conf. Proceeding ICMLC, pp. 86–90, 2019.

[11] L. C. Cheng and S. L. Tsai, "Deep learning for automated sentiment analysis of social media," in Proceedings of the IEEE/ACM Int. Conf. on Advances in Social Networks Analysis & Mining, 2019.

[12] S. M. Parikh and M. K. Shah, "Classification of sentiment analysis using machine learning," in Lecture Notes on Data Engineering and Communications Technologies, 2020.

[13] M. Khan and A. Malviya, "Big data approach for sentiment analysis of twitter data using Hadoop framework and deep learning," in ic-ETITE, IEEE, 2020.

[14] N. Pawar and S. Bhingarkar, "Machine Learning based Sarcasm Detection on Twitter Data," in 5th International Conference on Communication and Electronics Systems (ICCES), IEEE, 2020.

[15] G. A. Ruz, P. A. Henríquez, and A. Mascareño, "Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers," Futur. Gener. Comput. Syst., vol. 6, pp. 92–104, 2020.

[16] S. Tiwari, A. Verma, P. Garg, and D. Bansal, "Social Media Sentiment Analysis on Twitter Datasets," in 6th International Conf. on Advanced Computing and Communication Systems, 2020.

[17] P. Harjule, A. Gurjar, H. Seth, and P. Thakur, "Text Classification on Twitter Data," in Proceedings of 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things, IEEE, 2020.

[18] Y. Chandra and A. Jana, "Sentiment analysis using machine learning and deep learning," in 7th Int.Conf. on Computing for Sustainable Global Development, INDIACom, IEEE, 2020.

[19] J. Xu and C. Chen, "Smart Abnormal Emotion Analysis from English Sentences," Internet Technol. Lett., no. Special Issue Article, Wiley, p. 9, Aug. 2020.

[20] H. T. Phan, V. C. Tran, N. T. Nguyen, and D. Hwang, "Improving the Performance of Sentiment Analysis of Tweets Containing Fuzzy Sentiment Using the Feature Ensemble Model," IEEE Access, vol. 8, 2020.

[21] S. A. El Rahman, F. A. Alotaibi, and W. A. Alshehri, "Sentiment Analysis of Twitter Data," in ICCIS, IEEE,2019.

[22] S. Ranjan, S. Sood, and V. Verma, "Twitter Sentiment Analysis of Real-Time Customer Experience Feedback for Predicting Growth of Indian Telecom Companies," in 4th ICCS, 2019.

[23] F. A. Acheampong, C. Wenyu, and H. Nunoo-Mensah, "Text-based emotion detection: Advances, challenges, and opportunities," Eng. Reports, vol. 2, no. 7, 2020.

[24] A. Poornima and K. S. Priya, "A Comparative Sentiment Analysis of Sentence Embedding Using Machine Learning Techniques," in 6th ICACCS, IEEE, 2020.

[25] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in ICET 2017, 2018.

[26] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in Proceedings of the Annual Conference of the International Speech Communication Association, 2014.

[27] K. Sahu, Y. Bai, and Y. Choi, "Supervised Sentiment Analysis of Twitter Handle of President Trump with Data Visualization Technique," in 10th Annual Computing and Communication Workshop and Conference, IEEE, 2020.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 27,2021 at 10:20:03 UTC from IEEE Xplore. Restrictions apply.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/348216368

A Review of AI for Urban Planning: Towards Building Sustainable Smart Cities

Conference Paper · January 2021

CITATIONS
0

READS
31

5 authors, including:



Avinash Kumar Jha
Delhi Technological University
6 PUBLICATIONS **2** CITATIONS
SEE PROFILE



Surendrabikram Thapa
Delhi Technological University
8 PUBLICATIONS **2** CITATIONS
SEE PROFILE



Aryan Jha
Delhi Technological University
1 PUBLICATION **0** CITATIONS
SEE PROFILE

All content following this page was uploaded by Surendrabikram Thapa on 05 January 2021.

The user has requested enhancement of the downloaded file.

A Review of AI for Urban Planning: Towards Building Sustainable Smart Cities

Avinash Kumar Jha[1,#], Awishkar Ghimire[2], Surendrabikram Thapa[2], Aryan Mani Jha[3] and Ritu Raj[1]

[1]Department of Civil Engineering, Delhi Technological University
[2]Department of Computer Science and Engineering, Delhi Technological University
[3]Department of Mechanical Engineering, Delhi Technological University
[#]avinash.jha.6696@gmail.com

ABSTRACT

Urban planning, in short, deals with solving the problems of the modern society we live in. The problems are complementary to the growing population in today's society. The problems in society range from mundane tasks like ensuring sanitization in society to more technical tasks like managing infrastructures. The concept of smart cities, lately, has been a topic of great interest for social scientists, engineers, and everyone who wants to integrate technologies into their daily lives. AI and IoT have become an important part of our lives. Data has become ubiquitous with such smart devices that are connected to the internet. Such data can be used to make intelligent systems for smart cities. AI and IoT have promising effects on urban life. Artificial Intelligence is often considered to be the fourth industrial revolution because of its unprecedented potential to change everything. As AI progresses to improve day by day, it has blessed humans with everything from smart healthcare to secured smart cities. Everything has been changed by AI and IoT in smart cities. In this paper, the various use cases of AI and IoT for urban planning in order to build smart and sustainable cities are discussed. We have reviewed various papers that could potentially be used in the development of smart cities.

Keywords— Artificial Intelligence, Internet of Things, Smart Cities, Sustainability, Urban Planning

1. INTRODUCTION

Urban Planning is the process of designing cities or broadly a living structure for human beings like towns or rural areas, in such a way that all the events that happen within this living structure are efficient and proper. Urban planning has been given many synonyms among which are town planning, regional planning, city planning. Urban planning in its core is about designing and developing structures in which human beings can live their day-to-day life effectively and without much hassle. It is a field that requires proper knowledge of various subjects which include political science, civil engineering, human geography, social sciences. Urban planning answers questions such as how people will live, how people will commute, how people will work, how garbages are disposed, how distribution channels function, how materials are transported, how sanitization will be implemented and many more such logistical questions that deal with people and their relation to a particular physical location, primarily the location where people's homes and workplace is situated [1].

Urban Planning is not a new concept. Kings and rulers of ancient cities implemented urban planning when building cities and settlements. However, their idea of urban planning was vastly different from the ideas of modern-day urban planning. Ancient urban planning mostly dealt with how homes, public buildings and streets were structured [2]. However, modern day urban planning deals with a much broader range of subjects such as transport optimization, traffic systems management, water leakage management, safety in streets and other things that are unique to the modern-day era [3].

Artificial intelligence is a field of computer science that deals with intelligent agents that are able to solve a task without the intervention of a human mind. It aims to create machines, robots or simply computational models that solve different problems in the world. The motivation for artificial intelligence comes from trying to mimic or replicate the functioning of the human mind [4]. The major subfield of Artificial Intelligence that is gaining wide recognition today is machine learning, because machine learning as a discipline finds many applications in various different subjects. The heart of machine learning is computational algorithms that are able to learn through data as well as improve over time. It does so using mathematical disciplines such as computational statistics and calculus as well as various other related subjects [5]. Artificial Intelligence and Machine learning finds many applications among a broad range of fields today. It has been estimated that around 47% of today's jobs will be replaced by artificially intelligent agents within a short span of 10 years. Artificial Intelligence and Machine learning finds applications in propelling the growth of businesses and since business is arguably the most important entity in today's economic world, we can infer that artificial intelligence has a

huge impact in running the present-day world as well as building the future [6]. Presently, artificial intelligence and machine learning also finds application in medical diagnosis, easing the workload for medical practitioners [7].

Artificial intelligence is automating today's world and naturally it also finds applications in urban planning and urban design accelerating the development of smart cities. Since the 19th century Urban planning has been mainly concerned with improving the quality of people's lives based on the economical functioning of cities and on improving social equity. In the urban planning process, many functions are routine and automated with human labour but there also processes that need the intervention of a critical human mind. Within the short term, it has been predicted that many routine tasks will be overtaken by artificially intelligent systems whereas those tasks which require critical skills such as design will take more time to be replaced by automated systems [8]. Artificial Intelligence as is used today requires enormous amounts of data for its proper use. Since, today's modern cities are being digitized through the installation of CCTV cameras, sensors and large telecommunication networks it is possible to obtain data about how people in cities live, and how cities evolve over a certain period of time. It is possible to obtain data about land spaces, open spaces, buildings and such. Once this data is fed to the artificial intelligence system, primarily machine learning models, information about the urban fabric i.e., the logistics of the city comes to light and through proper reasoning and design urban planners are able to make use of it. Urban governance can also be implemented through these data obtained through various digital sources. Politicians can implement better policies that benefit the people living in urban areas because of information that comes out of an artificial intelligence model that processes the data. Artificial intelligence has also been used to make policies in urban planning that mitigates the effect of climate change and also policies that help in mitigating climate change, as in the modern world climate change is of paramount importance and designing cities that alleviates negative effects of climate change is quintessential. Even though AI influenced urban planning seems to be humanity's attempt at a futuristic world, it has its own flaws, like breach of privacy. Since large amounts of data need to be collected and stored for an AI model to work, the privacy of citizens living in smart cities is inherently disrupted. Since data has to be stored somewhere, there can be hackers trying to barge into centers and access unauthorized data. So there needs to be certain policies and techniques that safeguard the data of people [9].

2. RELEVANT WORKS

We have reviewed the latest artificial intelligence and machine learning techniques and its applications that could potentially be used in the development of smart cities. In particular we have reviewed Applications of AI in traffic system management, crime detection, air quality monitoring, efficient energy management and water leakage detection systems. There are a lot more applications that are quintessential in developing smart cities but these are the most important ones and hence we have reviewed recent popular research papers on them.

A. Traffic System Management

Yongchangma et al. [10] et al have proposed a framework for assessing the real time condition of a certain highway or a road. In their paper they use the concept of Vehicle Infrastructure Integration (VII). In this concept vehicles have computational units as well as infrastructure in the road have computational units, and these units are able to communicate with one another along a communication network. The computational unit and sensors in the vehicles are able to detect the speed profiles, wheel turning as well lane changing behavior of the vehicles. With all these data, it is important to classify them into classes of information so as to know the condition of the traffic. Yongchangma et al used support vector machines to classify the data into 3 classes of information namely, normal, passed by a possible incident or stooped in a queue. This classification was done for each vehicle infrastructure integration enabled vehicle. They compared support vector machines with artificial neural networks for the classification and found that the support vector machines performed better given the same training set and test set. The false alarm rate detected by their system was also considerably low. Their system was also further able to classify incidents blocking one lane, two lanes and three lanes. It was found that the SVM performed much slower when the density of vehicle infrastructure enabled vehicles increased, however not by much. The experiment done by them was performed via a traffic simulation software and not in a real road network. So trying to extrapolate this research into a real road network might not turn out to be as accurate or efficient. However, this paper shows that with the right technology and the correct use of machine learning models it is possible to have a real time system that is able to accurately predict traffic configurations.

Nallapermua et al. [11] have proposed a system names STMP (Smart Traffic management platform) that is able to harness the power of big data and artificial intelligence algorithms. The system makes use of sensor networks in roadways, Internet of Things (IoT) as well as social media data to make predictions on traffic flow as well as give solutions to traffic management problems. Their system primarily does the following things:

a. Detect concept drifts such as peak hours/non peak hours as well as incidents in roadways such as accidents. This all happens real time. This system is

implemented through an online incremental machine learning algorithm based on Incremental Knowledge Acquisition and Self Learning (IKASL) algorithm.

b. Determine the sentiment and emotion of vehicle users using social media data in a non-recurrent traffic event such as an accident.
c. Use real time traffic data to provide optimal traffic control strategies using deep reinforcement learning.
d. Predict traffic flow and make estimates on impact propagation using deep neural networks.

Their system was run on a smart sensor network traffic data generated by hundreds of thousands of vehicles on the arterial road network in a state in Australia. Their system provided very good results and was also implementable in the real world. The high-level overview of the architecture is as shown in fig. 1.

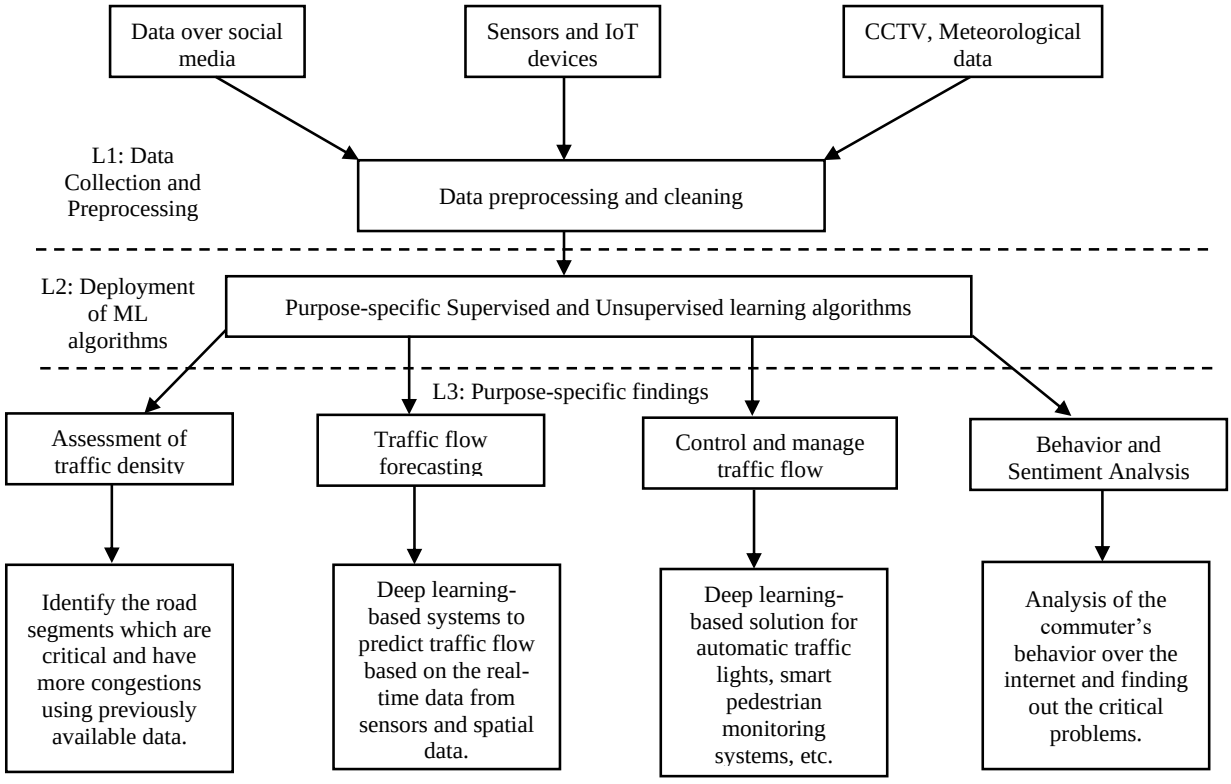


Fig. 1. Smart Traffic Management Architecture

Wang et al. [12] have proposed a system which can detect traffic jams in a freeway using camera video footage and images. This is an essential step towards building a smart city because ideally in a smart city there should be a system that detects and resolves traffic jams in the city. Their system deals with the detection part only and not the resolution part. They use CNN architecture for this system. The various different CNN architectures they use are VCGNET, AlexNET specifically they have tested and compared AlexNet + SVM, Transfer Learning from AlexNet, Transfer learning from VCGNet and VCGNet + SVM architecture. The dataset they used consisted of 30,000 traffic congestion images and 20,000 non congestion images and they obtained the dataset from a surveillance system used in Shaanxi Province Traffic Management Bureau. The accuracy they obtained was 90% using the transfer learning from AlexNet architecture.

B. Crime Detection

It is of paramount importance that surveillance systems in smart cities are able to automatically detect unusual activities such as crimes, robberies, shooting and further alarm the concerned authorities. Much research has been done using artificial intelligence and machine learning techniques to solve this particular problem of crime detection or largely anomaly detection in surveillance systems. Having such automatic systems would greatly reduce human intervention and the need for 24/7 security. Many researchers have worked on this problem and produced models that are able to give high accuracy on certain datasets, however employing these models in the real world do not yet give the same accuracy as they do on simple datasets [13]. Employing

these techniques would harness safer streets for women, decrease crime and theft considerably and improve the well-being of the general people.

Ullah et al. [14] have proposed a deep learning model that is able to detect violence in surveillance systems. Their system consists of the Convolutional neural network architecture and SoftMax layer for output classification as shown in fig. 2. Their system primarily consists of 4 parts. In the first part a CNN model was used for object detection, mainly human detection. If a human was detected a stream of 16 frames was passed onto a 3D CNN that extracted features from the sequence of frames. These features were further passed onto a SoftMax layer for classification. If the SoftMax layer predicted that violence was detected their system alerted a nearby police station and the concerned authorities. They tested their model on three different popular datasets for violence detection and they were hockey fight dataset, violent crowd dataset, and violence in movies dataset. When their model was trained on the violent crowd dataset, testing it against the violent crowd dataset gave an accuracy of 98%, whereas when they tested this model on violence in movies dataset and hockey fight dataset it gave accuracies of 65% and 47% respectively. When their model was trained on the violence in movies dataset, testing it against the violence in movies dataset gave an accuracy of 99.9%, whereas when they tested this model on violent crowd dataset and hockey fight dataset it gave accuracies of 54% and 63% respectively. When their model was trained on the hockey dataset, testing it against the violent crowd dataset gave an accuracy of 96, whereas when they tested this model on violent crowd dataset and violence in movies dataset it gave accuracies of 52% and 49% respectively. Their model gave exceptional accuracy for a particular dataset but it was not able to generalize well into other datasets. For a violence detection surveillance system to work in the real world, it has to generalize into different datasets as the surveillance environment may vary.

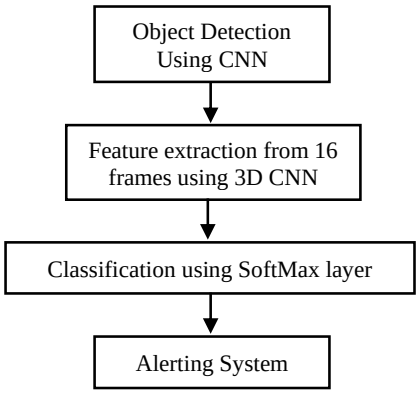


Fig. 2. Model of ullah et al. [14] for violence detection in surveillance systems

Bappee et al. [15] have proposed a way to relate geospatial features to crime. They used features such as the coordinates, postal code, hubs like cafes near the area as well as measures such as distance nearest to the hotpoint. A hot point is referred to as an area where crime happens. Their data consisted of 4 labels namely, alcohol related fights and accidents, assault, property damage and motor vehicle related accidents. The classifiers they used were Logistic regression classifier, support vector machines and random forest classifier. Their research is very important in the context of urban planning because it shows some sort of relation between geospatial features and crimes. Using this information, urban planners can design cities that are less prone to crime.

C. Air quality monitoring

It is important for smart cities to have automatic air quality monitoring systems that are able to detect the quality of the surrounding air and alarm the concerned authorities in case of any anomaly.

Wu et al. [16] have proposed a system that can detect the air quality using computational lens free microscopy and machine learning algorithms. Their system is named C-Air and it is also combined with a smartphone application that can control various parameters as well as display the results. It consists of a device that is able to take microscopic images of the particles in the air and the machine learning model is responsible for predicting what particles are present in the image and the size of the particles. They used a custom-designed machine learning algorithm. The accuracy of the model for predicting the size was found to be around 93%.

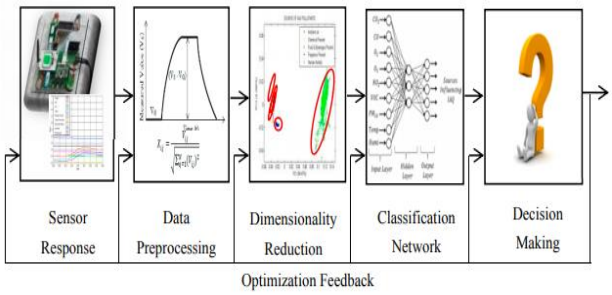


Fig. 3. Indoor air quality detection system as proposed by saad et al. [17]

They did some testing with this system and found that the air quality increased as the distance from the airport increased and hence it is known that the air quality near an airport is much worse than average.

Saad et al. [17] have proposed a system that is able to detect indoor air quality using sensors and supervised machine learning algorithms as shown in fig. 3. In their experiment they used a couple of sensors that were able to

detect 9 different parameters. The parameters were temperature, humidity, ozone(O3), oxygen(O2), volatile

organic compounds (VOC), carbon monoxide (CO), nitrogen oxide (NO2), carbon dioxide (CO2) and particulate matter (PM10). Their system was able to classify 5 different output classes namely, presence of fragrance, presence of food and beverages, ambient air, combustion activity and presence of chemicals. They collected the data themselves and didn't use any popular dataset. They tested their self-collected data with a lot of traditional machine learning models and found that 3 of the models gave an accuracy of up to 100%. The three models were multi-layer perceptron, linear discriminant analysis and k nearest neighbors. When there was a presence of more output classes indoors at the same time their model was unable to correctly classify and gave the output as unknown. Their research is very important in the context of building smart cities because ideally all homes would have an automatic air quality detection system and in the case of some sort of an anomaly like a gas leakage the concerned authorities or the homeowner would be informed.

D. Efficient Energy Management

It is of paramount importance that energy usage is optimized and properly utilized in smart cities. Using energy management schemes, it is possible to distribute energy to various locations in a smart city such that there is no energy loss and energy is properly distributed according to the need of the area. Energy Management is not just concerned with the consumer of energy but also with the supplier of energy. In today's world, there are many sources of energy such as renewable sources and non-renewable sources, so it would be important in smart cities to be able to distribute the energy to households and buildings efficiently. Energy management schemes also come into play when the amount of energy that can be distributed is limited. The energy that we are talking about here is largely electrical energy.

One of the key parts in an electrical management system is electrical load forecasting, that is to be able to forecast how much energy a particular area would use at a particular point of time. This is important so that more electrical energy than needed is not sent to this area, thus effectively managing the distribution of electricity. Khan et al. [18] have proposed a machine learning model that predicts electrical load in time series data based on machine learning models. The dataset they have used is South Korea's hourly energy consumption. The data set contained huge gaps, around 21%. They use several predictive machine learning that predict the values at the gaps to fill in the gaps. The classifier they used was a hybrid classifier that consisted of three machine learning models namely, XGBoost, CatBoost and random forests. They compared their hybrid model with several existing machine learning models and found that their hybrid model gave the best results. In time series forecasting the measure that we use to know the effectiveness of the model is regression score. It was found that the regression score of XGBoost, CatBoost and random forests were 0.8195, 0.8294, 0.7996. The hybrid model proposed by Khan et al had a regression score of 0.9212, thus outperforming the traditional machine learning models.

Kim et al. [19] conducted a research in which they effectively predicted the energy consumption of a particular household based on the features of the house which include number of exterior walls in that house, area of the house, number of years occupied, numbers of household members, occupation of the household members among other features. They firstly divided the data according to the period of the year. The periods were fall, summer, winter and autumn. Among the large feature set they performed regression analysis for each period and selected the input features. The classification model that they used was the artificial neural network model. They used the Household Energy Standing Survey dataset. Their research found that inspiring the building direction was the most influential element in predicting energy consumption. In summer it was the heating method, colling method and the area of the household. IN fall it was the housing area, housing type and occupation of the household members and in winter it was building direction, housing area and occupation of the household members.

E. Water Leakage Detection

Water is one of the most essential elements to life. Smart and efficient distribution of water is a must in cities. It has been known that the water pipeline leakage rate is 22% in Asia [20]. Hence it is important to build systems that are able to detect water leakage and stop them. An automatic water leakage detection, localization and resolution system is very important in the development of truly smart cities. Much research has been done on this topic particularly in water leakage detection and localization systems.

Kang et al. [21] have proposed an artificial intelligence system that is able to detect water leakage in water distribution systems. Their system is also able to localize the area of the water distribution system where the leakage is happening. In their system a leakage monitoring system was installed in a water distribution system. The leakage monitoring system and sensors that captured noise data and other data in the form of signals. The machine learning model they used was an ensemble of 1D CNN and SVM. The localization technique they used was based on graph theory and the water pipeline system was represented as a graph. It was found that their system correctly identified water leakage up to an accuracy of 99.3% and gave a localization error of less than 3m. Using their system, it is possible to build real world water leakage detection and localization systems. Their

work can undoubtedly be used in building water pipeline distribution systems in smart cities.

3. DISCUSSION AND FURTHER APPLICATIONS

Table I mentions some of the works that have been done in the field of urban planning using artificial intelligence. There are many other applications of artificial intelligence on the development of smart cities and in modern urban planning. It is important for urban planners to know the layout of cities. Traditionally, Human labor was used to find the specific layout of cities, but with the advancement of technology it is possible to analyze the layout of cities using geospatial imagery [22]. Using satellite imagery, computational techniques are automatically able to detect and identify different types of objects present in the image, thus making analysis of infrastructure in cities a million times easier and faster, and subsequently making the urban design and planning procedure much quicker and easier.

In smart cities, theoretically every single aspect would be automated. Recent advancement in the field of Artificial intelligence has made it possible to automate the field of road repairing as well. From the detection phase of road cracks and anomalies to the repairing process can be automated. Kumar et al. [24] have done a study on detection of cracks from geospatial imagery of roads. Their model uses Convolutional neural networks, a deep learning framework to solve the problem of crack detection. Principal Component Analysis is also incorporated in their work. We can use their work for building an automatic road repairing system. Their work can be used in the detection phase of road cracks, and further systems can be created that will alert the concerned authorities on the matter. The alerting process can also be optimized wherein cracks in important roads are detected and notified first. Many problems can be solved with AI, IoT and automation. Problems that occur accidentally and are usually difficult to handle like tackling pandemic can also be solved with these technologies [23].

TABLE I. RECENT WORKS IN THE FIELD OF URBAN PLANNING AND SMART CITIES

Name of Author	Year of Publication	Methods used	Use case
Yongchcangma et al. [10]	2009	Vehicle infrastructure integration + support vector machines	Traffic system management
Nallaperuma et al. [11]	2019	Incremental machine learning algorithm, deep neural network and deep reinforcement learning	Traffic system management
Wang et al. [12]	2018	CNN	Traffic system management
Ullah et al. [14]	2019	2D CNN + 3D CNN + SoftMax	Crime Detection
Bappee et al. [15]	2018	Logistic regression classifier, support vector machines and random forest classifier	Crime detection
Wu et al. [16]	2017	Custom designed machine learning algorithm	Air quality monitoring
Saad et al. [17]	2017	Multi-layer perceptron, Linear discriminant analysis and K nearest neighbours	Air quality monitoring
Khan et al. [18]	2020	Ensemble of XGBoost, CatBoost and random forest	Efficient Energy management
Kim et al. [19]	2019	Regression analysis + Artificial Neural Networks	Efficient energy management
Kang et al. [21]	2017	1D CNN + SVM and graph theory	Water leakage detection

It is important for smart cities to ponder upon and implement the concept of sustainability as well. Due to the recent environmental crisis, it has become increasingly important to develop the modern world in a sustainable manner. Urban planning should incorporate sustainable development into its philosophy for a better and healthier world for our children. Artificial Intelligence also sees use cases in the area of sustainable development.

4. CONCLUSION AND FUTURE SCOPE

In this paper, the various frameworks and technologies for smart cities are discussed. The concept of smart cities is ever changing. With engineering and social sciences leaping towards perfection every day, there is always a scope for improvement in existing infrastructures and technology. The research in artificial intelligence is booming which has affected every mundane and advanced thing. It has become self-evident that the integration of IoT and AI in our day-to-day activities has made our lives easier. The complex barriers for smart cities are broken down by AI and IoT. In this paper we have reviewed various recent studies done on artificial intelligence that could potentially be used in the making of smart cities. From various successful use cases of AI and IoT discussed above, it is unanimous that intelligent systems can solve the existing problems in more convenient ways. The upcoming innovation and technologies are certain to change our lives by meeting the concept of smart cities. From smart homes to safe streets for women, AI and IoT has changed each and every prospects of a society and will continue to do so with booming innovations in engineering and technology. In the future, the technology might be able to solve the problems that are even beyond the capacity of engineering and technology today to make the world a better place to live.

REFERENCES

[1] J. Niemelä, "Ecology and urban planning," Biodiversity & Conservation, vol. 8, no. 1, pp. 119-131, 1999.

[2] M. E. Smith, "Form and meaning in the earliest cities: a new approach to ancient urban planning," Journal of planning history, vol. 6, no. 1, pp. 3-47, 2007.

[3] O. Kaiwartya et al., "Internet of vehicles: Motivation, layered architecture, network model, challenges, and future aspects," IEEE Access, vol. 4, pp. 5356-5373, 2016.

[4] S. Thapa, S. Adhikari, A. Ghimire, and A. Aditya, "Feature Selection Based Twin-Support Vector Machine for the diagnosis of Parkinson's Disease," in 2020 IEEE 8th R10 Humanitarian Technology Conference (R10-HTC), 2020.

[5] S. Thapa, S. Adhikari, U. Naseem, P. Singh, G. Bharathy, and M. Prasad, "Detecting Alzheimer's Disease by Exploiting Linguistic Information from Nepali Transcript," in International Conference on Neural Information Processing, 2020: Springer, pp. 176-184.

[6] A. Ghimire, S. Thapa, A. K. Jha, S. Adhikari, and A. Kumar, "Accelerating Business Growth with Big Data and Artificial Intelligence," in 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), 2020: IEEE, pp. 441-448.

[7] S. Thapa, P. Singh, D. K. Jain, N. Bharill, A. Gupta, and M. Prasad, "Data-Driven Approach based on Feature Selection Technique for Early Diagnosis of Alzheimer's Disease," in 2020 International Joint Conference on Neural Networks (IJCNN), 2020: IEEE, pp. 1-8.

[8] M. Batty, "Artificial intelligence and smart cities," ed: SAGE Publications Sage UK: London, England, 2018.

[9] Z. Allam and Z. A. Dhunny, "On big data, artificial intelligence and smart cities," Cities, vol. 89, pp. 80-91, 2019.

[10] Y. Ma, M. Chowdhury, A. Sadek, and M. Jeihani, "Real-time highway traffic condition assessment framework using vehicle–infrastructure integration (VII) with artificial intelligence (AI)," IEEE Transactions on Intelligent Transportation Systems, vol. 10, no. 4, pp. 615-627, 2009.

[11] D. Nallaperuma et al., "Online incremental machine learning platform for big data-driven smart traffic management," IEEE Transactions on Intelligent Transportation Systems, vol. 20, no. 12, pp. 4679-4690, 2019.

[12] P. Wang, L. Li, Y. Jin, and G. Wang, "Detection of unwanted traffic congestion based on existing surveillance system using in freeway via a CNN-architecture trafficnet," in 2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA), 2018: IEEE, pp. 1134-1139.

[13] W. Ullah, A. Ullah, I. U. Haq, K. Muhammad, M. Sajjad, and S. W. Baik, "CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks," Multimedia Tools and Applications, pp. 1-17, 2020.

[14] F. U. M. Ullah, A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, "Violence detection using spatiotemporal features with 3D convolutional neural network," Sensors, vol. 19, no. 11, p. 2472, 2019.

[15] F. K. Bappee, A. S. Junior, and S. Matwin, "Predicting crime using spatial features," in Canadian Conference on Artificial Intelligence, 2018: Springer, pp. 367-373.

[16] Y.-C. Wu et al., "Air quality monitoring using mobile microscopy and machine learning," Light: Science & Applications, vol. 6, no. 9, pp. e17046-e17046, 2017.

[17] S. Mad Saad et al., "Pollutant recognition based on supervised machine learning for indoor air quality monitoring systems," Applied Sciences, vol. 7, no. 8, p. 823, 2017.

[18] P. Waqas Khan, Y.-C. Byun, S.-J. Lee, and N. Park, "Machine Learning Based Hybrid System for Imputation and Efficient Energy Demand Forecasting," Energies, vol. 13, no. 11, p. 2681, 2020.

[19] M. Kim, S. Jung, and J.-w. Kang, "Artificial Neural Network-Based Residential Energy Consumption Prediction Models Considering Residential Building Information and User Features in South Korea," Sustainability, vol. 12, no. 1, p. 109, 2020.

[20] Y. Liu, X. Ma, Y. Li, Y. Tie, Y. Zhang, and J. Gao, "Water Pipeline Leakage Detection Based on Machine Learning and Wireless Sensor Networks," Sensors, vol. 19, no. 23, p. 5086, 2019.

[21] J. Kang, Y.-J. Park, J. Lee, S.-H. Wang, and D.-S. Eom, "Novel leakage detection by ensemble CNN-SVM and graph-based localization in water distribution systems," IEEE Transactions on Industrial Electronics, vol. 65, no. 5, pp. 4279-4289, 2017.

[22] D. Kit, Y. Kong, and Y. Fu, "LASOM: Location Aware Self-Organizing Map for discovering similar and unique visual features of geographical locations," in 2014 International Joint Conference on Neural Networks (IJCNN), 2014: IEEE, pp. 263-270.

[23] A. Ghimire, S. Thapa, A. K. Jha, A. Kumar, A. Kumar, and S. Adhikari, "AI and IoT Solutions for Tackling COVID-19 Pandemic", in 2020 International Conference on Electronics, Communication and Aerospace Technology, 2020: IEEE, pp. 1083-1092 .

[24] A. Kumar, A. Kumar, A. K. Jha, and A. Trivedi, "Crack Detection of Structures using Deep Learning Framework", in 2020 International Conference on Intelligent Sustainable Systems, 2020: IEEE

View publication stats

AIP Conference Proceedings

A theoretical comparison between $CH_3NH_3PbI_3$ and $CH_3NH_3SnI_3$ based solar cells

Cite as: AIP Conference Proceedings **2265**, 030609 (2020); https://doi.org/10.1063/5.0017201
Published Online: 05 November 2020

Gagandeep, Mukhtiyar Singh, Ramesh Kumar, Vinamrita Singh, and Sunita Srivastava

View Online Export Citation

Meet the Next Generation of Quantum Analyzers
And Join the Launch Event on November 17th



Register now
Zurich Instruments

AIP Publishing

© 2020 Author(s).

A Theoretical Comparison between CH3NH3PbI3 and CH3NH3SnI3 based Solar Cells

Gagandeep[1, a)], Mukhtiyar Singh[2], Ramesh Kumar[1, b)],Vinamrita Singh[3] and Sunita Srivastava[1,4)]

[1]*Department of Physics, Guru Jambheshwar University of Science & Technology, Hisar, India*
[2]*Department of Applied Physics, Delhi Technological University, Delhi, India*
[3]*Department of Applied Sciences and Humanities, Ambedkar Institute of Advanced Communication Technologies & Research, Delhi, India*
[4]*Department of Physics, Punjab University Chandigarh, India*

Corresponding author: a) gagankundu22@gmail.com
b) rameshkumar.bibiyan@gmail.com

Abstract. The perovskite materials based photovoltaic cells play a crucial role in the conversion of sunlight into electric energy. The commonly used perovskite materials are $CH_3NH_3PbI_3$ and $CH_3NH_3SnI_3$. In the present manuscript, a theoretical comparison is done by simulating the structures ITO/TiO_2/$CH_3NH_3PbI_3$/p-Gr/Ag and ITO/TiO_2/$CH_3NH_3SnI_3$/p-Gr/Ag structures in AFORS-HET software under the illumination condition of AM 1.5G. Here, TiO_2 and p-Graphene (p-Gr) play the role of electron transporting material and hole transporting material respectively with ITO and Ag as front and back contacts. The thickness of layers is directly related to the cost of the cells. Therefore, we optimized the $CH_3NH_3PbI_3$ and $CH_3NH_3SnI_3$ layers thickness to observe the performance of the device. The power conversion efficiency of 10.62% is obtained for $CH_3NH_3SnI_3$based solar cell at optimized thickness of 800 nm as compared to 9.74% efficiency for $CH_3NH_3PbI_3$based solar cell at an optimized thickness of 700 nm. The textured nature of front surface also enhances the efficiency 10.56% and 12.38% for $CH_3NH_3PbI_3$ and $CH_3NH_3SnI_3$ respectively.

INTRODUCTION

The largely employed perovskite material is methyl ammonium lead halide perovskite i.e. $CH_3NH_3PbI_3$ for the perovskite solar cells (PSC) [1, 2]. The PSC possess a heterojunction architecture with three layers [3] i.e ETM (electron transporting material), perovskite absorbing layer and HTM (hole transporting material) with front and back contacts as shown in fig. 1a. The commonly employed ETM and HTM are TiO_2 and organic material spiro-OMeTAD respectively [3]. The synthesis of organic material spiro-OMeTAD is expensive process and enhances the cost of device. Therefore, the researchers are developing different types of HTM in PSCs [4-6].

Graphene can be employed as a novel material in PSCs due to its appropriate electrical and optical characteristics [7]. The graphene can be inserted as ETM, HTM or transfer electrode in PSCs because its work function as well as energy band gap can be tuned under the influence of doping [7-9].

The toxicity of lead in PSC is a major obstacle in its development for commercial application. Therefore, tin based PSC (TPSC) [10] is being explored as a substitute of lead based PSC in future. Keeping this in mind, the

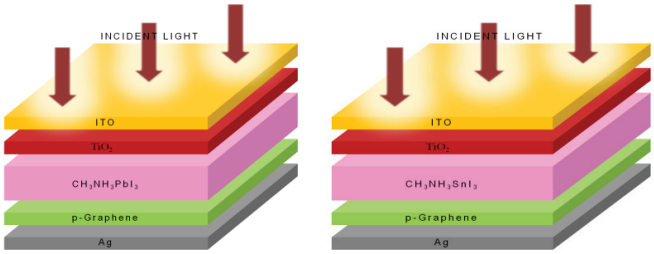
DAE Solid State Physics Symposium 2019
AIP Conf. Proc. 2265, 030609-1–030609-4; https://doi.org/10.1063/5.0017201
Published by AIP Publishing. 978-0-7354-2025-0/$30.00

present work compares the performance of lead-based and lead-free perovskite solar cells in order to develop environmentally friendly technology.

SIMULATION DETAILS

The structures ITO/TiO_2/$CH_3NH_3PbI_3$/p-Gr/Ag and ITO/TiO_2/$CH_3NH_3SnI_3$/p-Gr/Ag are simulated in Automat for Simulation of Heterostructures (AFORS-HET) software [11-13] as shown in fig. 1a and 1b respectively. Both the structures are n-i-p type i.e. light is falling from n-side. The electron-hole pairs are generated in the active layers. The generated electrons are collected by the ETM layer due to the built in voltage developed at the interface between the TiO_2 and perovskite layers. Similarly, generated holes are collected by the HTM layer i.e. p-graphene material.

To check the performance of both the devices, we optimized the active layer thickness. The $CH_3NH_3PbI_3$ and $CH_3NH_3SnI_3$layer thickness are varied in the range of 50-800 nm and 50-700 nm respectively. The variations in open circuit voltage (Voc), short circuit current density (J_{sc}), fill factor (FF) and power conversion efficiency (PCE) of the cell with absorbing layer thickness are observed. The impact of textured front surface is also studied on both the configurations. The detailed device parameters used in the simulation are taken from different theoretical and experimental studies [3, 14-16] and are mentioned in table 1.



(a) (b)

FIGURE.1 n-i-p structure of perovskite solar cell using (a) $CH_3NH_3PbI_3$ and (b) $CH_3NH_3SnI_3$ absorber layer

TABLE 1. Material parameter of each layer

Parameters	TiO_2	$CH_3NH_3PbI_3$	$CH_3NH_3SnI_3$	Graphene
Thickness (nm)	90	50-800	50-700	0.334
Dielectric constant	100	25	6.5	7.1
Electron affinity (eV)	4	3.9	4.17	4.7
Energy Band gap (eV)	3.2	1.5	1.3	0.5
Optical energy band gap (eV)	3.2	1.5	1.3	0.5
Valence band density (cm^{-3})	1×10^{21}	2.5×10^{20}	1×10^{18}	3×10^{19}
Conduction band density (cm^{-3})	2×10^{20}	2.5×10^{20}	1×10^{19}	3×10^{19}
Hole mobility ($cm^2\ V^{-1}\ s^{-1}$)	0.006	50	1.6	1×10^{5}
Electron mobility ($cm^2\ V^{-1}\ s^{-1}$)	0.006	50	1.6	10
Thermal velocity of electron(cm/s)	1×10^{6}	2.4×10^{18}	1×10^{6}	1×10^{7}
Thermal velocity of hole(cm/s)	4×10^{6}	2.4×10^{18}	1×10^{6}	1×10^{7}
Doping concentration(cm^{-3})	5×10^{19}	1×10^{17}	3.2×10^{15}	1×10^{18}
Layer density(g cm^{-3})	2.328	2.328	2.328	2.328

RESULTS AND DISCUSSIONS

With the help of layer parameters listed in table 1, we studied the effect of absorbing layer thickness of both the perovskite on the performance parameters i.e, J_{sc}, V_{oc}, FF and PCE of the solar cell. Fig.2 gives the effect of thickness of $CH_3NH_3PbI_3$ (Pb) and $CH_3NH_3SnI_3$ (Sn) perovskite layers on the different parameters. It is observed that J_{sc} (fig. 2(a)) shows increment in its value with increase in thickness in both the cases ($CH_3NH_3PbI_3$ and

$CH_3NH_3SnI_3$) to a saturation value of 16.99 mA/cm^2 and 19.90 mA/cm^2, respectively. This enhancement results due to comparatively thicker active layer which absorbs more photons. Greater photon absorption creates large electron-hole pair generation in active layer. The better J_{sc} behavior is demonstrated by $CH_3NH_3SnI_3$ based structure.

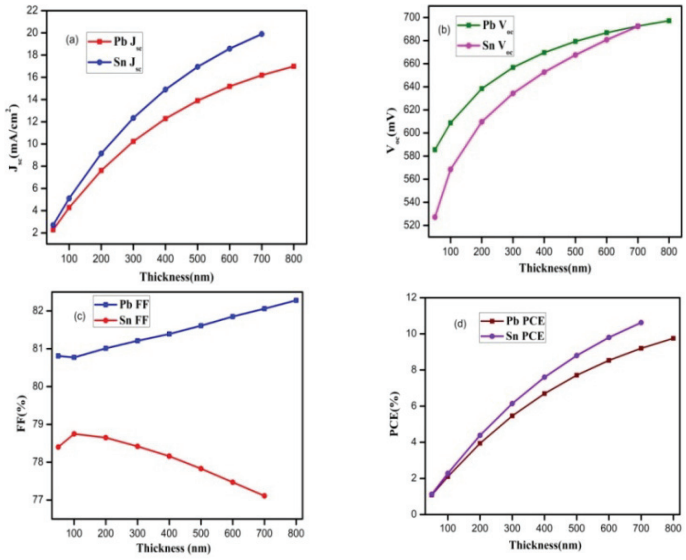


FIGURE 2. Effect of layer thickness of $CH_3NH_3PbI_3$ and $CH_3NH_3SnI_3$ on (a) J_{sc}, (b) V_{oc}, (c) FF and (d) PCE of the perovskite solar cell

The value of V_{oc} varies from 585.5 mV to 697.2 mV and 527.1 mV to 692.2 mV for $CH_3NH_3PbI_3$ and $CH_3NH_3SnI_3$ perovskite respectively and the results are given in fig. 2 (b). However, the FF value slightly reduces from 78.40% to 77.11% in $CH_3NH_3SnI_3$ layer within the thickness range of 50-700 nm. This may probably due to internal recombination in perovskite absorber layer. But in $CH_3NH_3PbI_3$, the value of FF increases from 80.81% to 82.28% with thickness ranging from 50-800 nm and the results are given in fig. 2(c). Fig. 2(d) gives the variation of PCE with thickness for both the perovskite solar cells and a considerable increase in the value of efficiency from 1.07% to 9.74% and 1.12% to 10.62% is noted for $CH_3NH_3PbI_3$ and $CH_3NH_3SnI_3$ respectively. The performance for $CH_3NH_3SnI_3$ based device is better as compared to $CH_3NH_3PbI_3$based device.

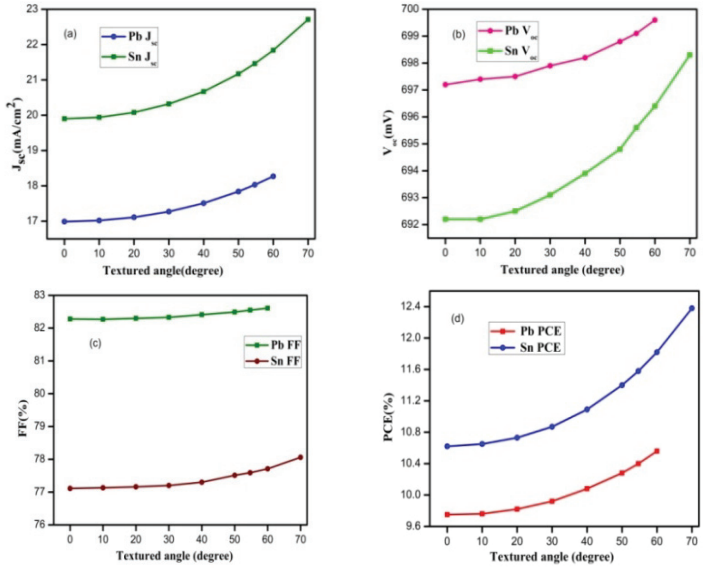


FIGURE 3. Effect of angle of incidence of light on $CH_3NH_3PbI_3$ and $CH_3NH_3SnI_3$ on (a) J_{sc}, (b) V_{oc}, (c) FF and (d) PCE of the perovskite solar cell

The performance of solar cell is also influenced by the nature of the front contact surface because it may be plane or textured. The textured front surface absorbs more photons as compared to the plane front surface. This is due to the fact that when incident light falls on the plane surface of the cell, it may bounce back but if the surface is textured the light undergoes multiple reflections. Hence, in effect, large area is used for the absorption of incident light improving the cell performance. We varied the angle of incidence from 0° to 60° and 0° to 70° for $CH_3NH_3PbI_3$ and $CH_3NH_3SnI_3$ respectively and the results are shown in fig. 3. It is observed that all the cell parameters improve with incidence angle giving J_{sc} values 16.99 mA/cm^2 to 18.27 mA/cm^2 and 19.90 mA/cm^2 to 22.71 mA/cm^2, V_{oc} = 697.2 mV to 699.6 mV and 692.2 mV to 698.3 mV, FF = 82.28% to 82.61% and 77.11% to 78.06% and PCE varying from 9.74% to 10.56% and 10.62% to 12.38% for $CH_3NH_3PbI_3$ and $CH_3NH_3SnI_3$ perovskite solar cell.

CONCLUSIONS

The present manuscript is devoted to the comparison between the $CH_3NH_3PbI_3$ and $CH_3NH_3SnI_3$based solar cells. The behavior of both the cells is discussed under the influence of perovskite layer thickness. We also replaced the complex material spiro-OMeTAD by graphene as hole transporting material. The PCE for both $CH_3NH_3PbI_3$ and $CH_3NH_3SnI_3$ solar cells are 9.74% and 10.62% at optimized active layer thickness of 800 nm and 700 nm respectively. The textured nature of front surface also improved the efficiency to 10.56% and 12.38% for $CH_3NH_3PbI_3$ and $CH_3NH_3SnI_3$solar cells respectively. Therefore, based on the values of performance parameters for both $CH_3NH_3PbI_3$ and $CH_3NH_3SnI_3$ solar cells, the tin based solar cells has found to exhibit better performance for the above suggested structure.

ACKNOWLEDGMENT

One of the authors (**Ramesh Kumar**) is grateful to the **University Grants Commission, INDIA** for providing the financial support through START-UP GRANT under FRPS (No. F.30-364/2017 (BSR), Dated 15/12/2017). The authors gratefully acknowledge the AFORS-HET software developer for providing free access of the software.

REFERENCES

1. A. Kojima, K. Teshima, Y. Shirai and T. Miyasaka J. Am. Chem. Soc. **131**, 6050 (2009).
2. N. J. Jeon et. al., Nat. Energy **3**, 682 (2018).
3. H. J. Du, W. C. Wang and Y. F. Gu, Chin. Phys. B **26**, 028803 (2017).
4. K. Mahmood, B. S. Swain and H. S. Jung, Nanoscale.**6**, 9127 (2014).
5. S. Lu et. al., Chem. Commun.**50**, 6931(2014).
6. P. Qin et. al., Nat. Commun.**5**, 3834 (2014)
7. E. L. Lim et.al., Nano-Micro Lett. **10: 27** (2018).
8. T. Mahmoudi, Y. Wang and Y. B. Hahn, Nano Energy.**47**, 51 (2018).
9. M. F. Bhopal et.al., J. Mater. Chem. C. **5**, 10701 (2017).
10. MH. Kumar et.al., Adv. Mater. **26**, 7122 (2014).
11. R.Varache et. al., Sol. Energ. Mat. Sol. C.**141**, 14 (2015).
12. R. Stangl, J. Haschke and C. Leendertz,Solar Energy, Intech (R.D. Rugescu (Ed.), Croatia, 2009), pp. 432.
13. R.Stangl, M. Kriegel and M. Schmidt, Proceeding of the WCPEC-4, Fourth World Conference on Photovoltaic Energy Conversion (Hawaii, USA), PP. 1350 (2006).
14. S.M. Iftiquar, J. S. Kim and J. Yi, Optik**148**, 54(2017).
15. K. Patel and P.K. Tyagi. Carbon**116**, 744 (2017).
16. Gagandeep, M. Singh and R. Kumar, AIP Conference Proceedings, **2115**, 030548 (2019).

Applied Intelligence
https://doi.org/10.1007/s10489-020-02057-9

A unified approach for detection of Clickbait videos on YouTube using cognitive evidences

Deepika Varshney[1] · Dinesh Kumar Vishwakarma [1]

Accepted: 31 October 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract
Clickbait is one of the form of false content, purposely designed to attract the user's attention and make them curious to follow the link and read, view, or listen to the attached content. The teaser aim behind this is to exploit the curiosity gap by giving information within the short statement. Still, the given statement is not sufficient enough to satisfy the curiosity without clicking through the linked content and lure the user to get into the respective page via playing with human psychology and degrades the user experience. To counter this problem, we develop a Clickbait Video Detector (CVD) scheme. The scheme leverages to learn three sets of latent features based on User Profiling, Video-Content, and Human Consensus, these are further used to retrieve cognitive evidence for the detection of clickbait videos on YouTube. The first step is to extract audio from the videos, which is further transformed to textual data, and later on, it is utilized for the extraction of video content-based features. Secondly, the comments are analyzed, and features are extracted based on human responses/reactions over the posted content. Lastly, user profile based features are extracted. Finally, all these features are fed into the classifier. The proposed method is tested on the publicly available fake video corpus [FVC], [FVC-2018] dataset, and a self-generated misleading video dataset [MVD]. The achieved result is compared with other state-of-the-art methods and demonstrates superior performance.

Keywords Clickbait detection · Cognitive evidence · Dataset · User profiling · YouTube

1 Introduction

In the era of instant gratification, people mostly used to communicate with each other via social media platforms. As social media platforms like Twitter, Facebook, YouTube, etc. providing ease of posting user opinions, lots of misleading and unreliable multimedia content are often posted and widely disseminated via popular social media platforms. Among all other platforms, YouTube is one of the leading ones for sharing videos, it has billions of users covering almost one-third of the internet population and reaches billions of views per day,[1] due to which it is plagued with clickbait videos that doesn't faithfully represent the situation that it refers to. Clickbait's are purposely designed to attract the user's attention and make them curious to follow the link and read, view, or listen to the attached content. In 1994, George Loewenstein has explained clickbait, *"as the information gap theory of curiosity"* [1]. We followed this definition and defined clickbait "*as the information gap theory of curiosity, that play with human psychology, to lure the user to view a content that does not faithfully represent the claim it presents and eventually degrades the user experience*". Whereas, "*Non-click baits can be defined as the content that is presenting the real news and faithfully giving the same picture of content to the viewer, that it is claiming for*". The paper provides a detailed description of clickbait video detection mechanisms in online social media platforms. Detecting clickbait videos is an intelligent task, as it analyses the video content automatically using clickbait video detection frameworks/tool/plugins, as well as in the future it can also be used as an intelligent warning system that can help to automatically report the credibility of video content to the user. Figures 1 and 2 shows the example of clickbait and non-clickbait video in brief. A recent example is of COVID- 19 pandemic, which affects the worldwide badly, and there is no shortage of people who are taking this crisis as an opportunity for malicious activities/gaining profit. A lot of health-related

[1] https://www.youtube.com/yt/about/press/

✉ Dinesh Kumar Vishwakarma
dvishwakarma@gmail.com

Deepika Varshney
deepikavarshney06@gmail.com

[1] Biometric Research Laboratory, Department of Information Technology, Delhi Technological University, Delhi 110042, India

Published online: 02 January 2021

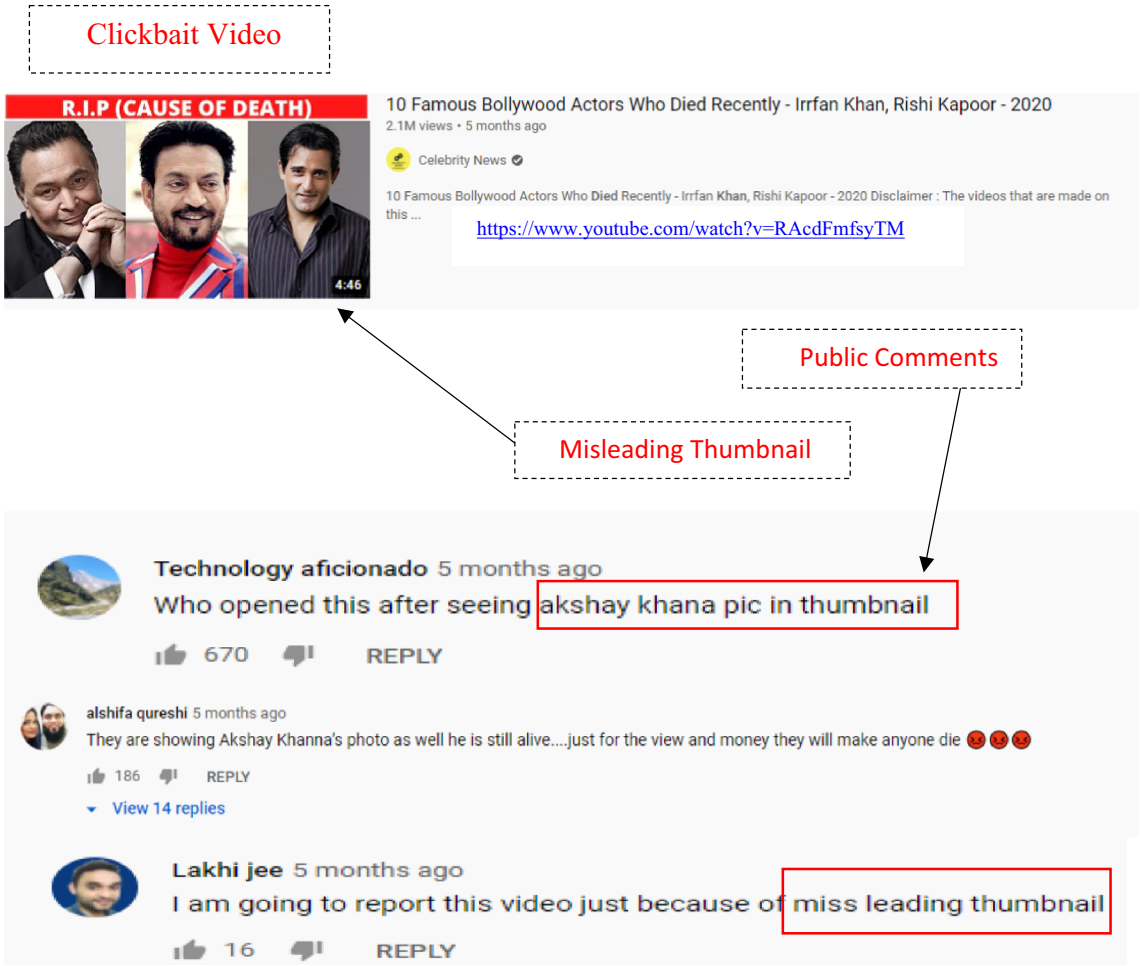


Fig. 1 Example of Clickbait Video

misleading information, some of the fake cures are suggested for COVID-19 have been posted by the malicious user, by adding catchy coronavirus headlines to increase their chance of a click, download, or purchase.

During this pandemic, people have their eye on any news announcement from the government official or some news that can help to get rid of COVID-19. The bad guy uses this opportunity to gain more views on their post, by adding catchy headlines with the news content that does not faithfully represent the event that it refers to and in this way it is spreading false information.[2] One of the fake YouTube video, gone viral having nearly half a million views falsely said that inhaling hot water from a hairdryer can help to cure the coronavirus,[3] which later turned out to be false. The presence of such misleading content over social media makes it more difficult for the user to discriminate the credible information from false stories, and it leads making it a challenging area in research. As the spread of clickbait videos not only degrades user experience, but also decreases the trustworthiness of video-sharing platforms. Few works have been reported in detecting clickbait on the video platform. There is a careful analysis required among the features extracted from the video. The current research has not addressed this problem fully, as they focus only on the content-based solution like the content of the video [2, 3], the image of the thumbnail [4, 5] or text of the title. Most of the text-based clickbait detection methods have adopted linguistic features [6], or word embeddings for the detection of clickbait news headlines, but those solutions cannot be employed to address the clickbait videos, as the only title may not be a reliable indicator, because two videos can share the same title with different content. In the same way, another sort of image-based approaches are employed that focus on thumbnail based features and are not found to be effective in solving the video clickbait detection problem.

In this work, we have proposed a novel mechanism by introducing three sets of evidential clues, identified and retrieved concerning each video, so that one can easily discriminate unsubstantiated information. The recent work addressed various text-based and comment-based features, however neglecting video speech-based features, as well as user profile

[2] https://www.buzzfeednews.com/article/janelytvynenko/coronavirus-fake-news-disinformation-rumors-hoaxes
[3] https://www.bbc.com/news/52124740

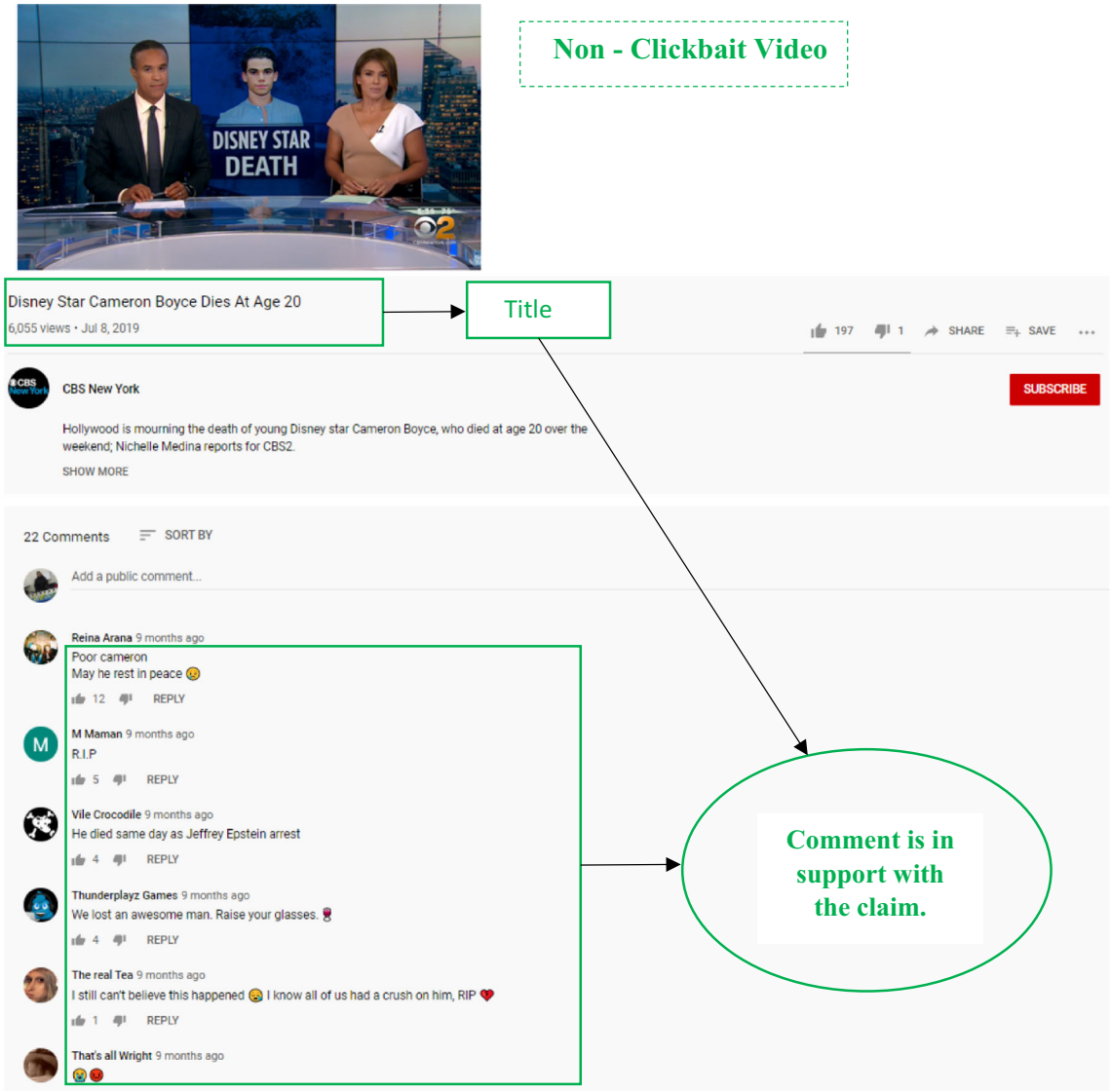


Fig. 2 Example of Non-Clickbait

features that are also not explored well and found to be very effective in detecting clickbait videos. To the best of our knowledge, speech-title similarity based features have not been explored well by the previous research, which can be an important clue to solve the problem. When we have two videos with a similar title but different content, the speech is converted into text, and then comparison has been made with the title to check how faithfully the speech is representing the title. Along with it, we have also addressed the problem, when the comments based features are not retrieved, as the uploader doesn't allow comments from the viewers. In that case, relying only on a certain set of features is not effective. To address this case, we have introduced another clue i.e. credible sources, through which we will be able to predict the credibility of the video, the detailed description has been given in the later sections. The proposed work makes the following significant contributions for the clickbait detection task:

- The proposed work gives a significant contribution in providing a novel methodology for the collection and annotation of YouTube videos of different categories. It builds a self-generated dataset[MVD] of Clickbait's and Non-Clickbait videos, as few and small datasets are publically available related to this area of research.
- To the best of our knowledge, our work is first to provide three sets of cognitive evidence for the clickbait's detection task, with the specification of desirable measures that are required for each specific possible detection case.
- The paper firstly includes speech-title similarity measures that haven't been included in the previous research. Some of the previous research is using thumbnail image and title of the video, which are not effective in the case when the two videos [clickbait and Non-Clickbait] having the same title and may also have the same thumbnail, then in that situation, it is critical to discriminate a clickbait and a Non-

Clickbait one. That's why to tackle this problem, the speech is also analysed, which gives a significant clue for discrimination [7].

- In some of the scenarios, the uploaders are not allowing the comments from the viewers, in that case, the comments part is not contributing, and we are losing some crucial features for the reliable prediction. To tackle this problem, we have also introduced other evidence that gathers important clue from the credible web links, when user responses are not there.
- We investigate the model performance with different classifiers, and comparative analysis reveals that our proposed method outperforms other state-of-the-art on the same dataset.

The remaining 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, where Section 3 gives a detailed description of how the data collection and annotation have been employed, where Section 4 describes the strategy/ method that we have used for the clickbait detection task, which 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 this section, the previous work on Clickbait, Rumour, Hoax, and Fake news detection are analyzed, and their performances are discussed. These are different prominent forms of misleading content that are available on social media [8] and used interchangeably concerning different contexts. For example, hoaxes can be prominently seen in the context of false celebrities' death stories. While the concept of Fake-news widely comes into picture during the US presidential election. The clickbait is also one of the eminent forms of misleading content available over various social media platforms. Initially, the research is going on the detection of clickbait's headlines that lure the user to view the webpage, now the concept is also moving towards detecting clickbait's videos and becoming an emergent area of research. In the following section, we briefly describe each of these categories.

2.1 Clickbait's detection

Detection of clickbait over the social media platforms, several techniques have been developed [[9, 10]]. For the automatic detection of clickbait, a browser extension has been developed by [10], which allows users with an option to block clickbait, as well as facilitate with the warning mechanism. The authors of [6] incorporate several set of handcrafted features like bag-of-words, n-grams, etc., to train the classifier and firstly introduce an automatic clickbait detector. In [6], authors have used the random forest for the prediction of clickbait tweets. Whereas, the authors of [11], proposed a technique to address the issue by employing SVM and Naïve Bayes classifier, while the authors of [12] address the problem of detecting clickbait's in the news articles, by incorporating Gradient Boosted Decision Trees. The authors of [1] have proposed an ensemble learner-based classification system for the prediction of clickbait and Non-Clickbait ones. From the analysis, it has been observed that random forest found to be the best classifier, with an accuracy of 91.16%. Whereas, [7] developed a novel content-agnostic scheme for effectively detecting the clickbait video by exploring and analyzing the comments from the viewers. Along with the machine learning technique, some of the researchers have also explored the deep learning methods. The authors of [13], firstly introduced the use of deep learning methods to counter the problem of detecting clickbait news articles. The recurrent neural network, in conjunction with word embeddings, has been employed as the proposed technique. In [[13, 14]], a deep neural network method has been used, however, the method considered text content for analysis i.e. the title and the descriptor, which is not quite effective in the case when the clickbait and Non-Clickbait videos are sharing same title and descriptor. From the survey, it has been observed that very few works have been reported clickbait detection solution that has considered video-based clickbait's. In recent work, the authors of [9], employed crowdsourcing based technique, where it has been predicted that the thumbnail of the video is clickbait or not. However, the technique found to be time-consuming, when the dataset is large. The further enhanced work has been proposed by [15], where they have used a combination of thumbnails and the statistical features of the user's comments and our work have not only introduced comment and user profile based features to retrieve useful evidence but also incorporate speech data from the videos for analysis, that further be compared with the title, to identify whether the title faithfully represents the video content or not.

Along with this, there are also some other forms of false information present over social media, like Rumours, Hoaxes, Fabricated, Conspiracy theories, Satire where research is going on [8].

2.2 Rumours detection

Rumours are also one of the forms of false information, which can be defined as the content/ post whose veracity is not verified at the time of posting and whose truthfulness is ambiguous or not confirmed. Many of the work has been done to detect whether a given post is a rumour or not [[16–21]]. The authors of [16], proposed two novel features, client-based, and location-based features. The classification result shows that SVM performs best in detecting fake news. The

authors of [22], proposed linguistic, temporal, and structural features from the tweet post and employed random forest, decision-tree, and SVM for the detection of rumours. The evaluation results reveal that the model achieves precision and recall scores of 87% and 92% respectively. Where in [23], user, structural, linguistic, and temporal features have been explored for the rumour classification, by analysing the task over varying time windows on twitter. In contrast, the authors of [24], learn a hidden representation of the input, by employing a recurrent neural network, without the need for hand-crafted features for rumour classification.

2.3 Hoax detection

Hoax is the news reports, whose facts are either false or incorrect, and they are representing it as a legitimate fact. Many times, Hoaxes have been seen in the context of false celebrities' death stories. The authors of [25], have proposed user interaction based features, and employ logistic regression for classification, that able to identify hoaxes with an accuracy of 99%. Whereas, in [26] the author proposes a technique using a random forest classifier to distinguish if an article is a hoax or not. The experimental results reveal that the proposed model achieves an accuracy of 92%. Whereas, the authors of [27], proposes a hoax detection scheme by employing user feedback features for news verification using a Naïve Bayes algorithm. However, Hoax detection is the less explored area, where few works have been reported and need further research.

2.4 Fake news detection

Fake news is another type of false information present over social media, and fake news can be defined as "a news article that is intentionally and verifiably false" [28]. Various textual and visual features have been employed for the detection of Fake news. The authors of [29], proposed a similarity-aware fake news detection method, that employed multimodal data(textual and visual) to investigate the relationship between the extracted features across modalities. The result reveals that the multimodal features and cross-modal similarity relationships are effective in detecting fake news. Whereas, in [30] authors proposed a technique for fake news detection by combining text mining techniques and supervised artificial intelligence algorithms, where the result shows that the best mean values in terms of precision, accuracy, recall, and f-measure have been obtained from the decision-tree, CVPS, ZeroR algorithms. In [31], the authors adopted a deep neural network(Convolution and Recurrent neural network) for the feature extraction process to predict fake news. Whereas, credible web sources are analyzed by [32] for the fake news prediction. They proposed a novel Rp (Non-Clickbait parameter) parameter, in which if the given threshold value exceeds, the event is classified as fake otherwise not. While many of the authors [[33–36]] provide a good literature survey by exploring techniques, features, datasets, and other analyses for further research exploration.

The original contribution of the work can be seen in Fig. 3, where the existing work and the proposed work has been demarcated. The main contribution of the work is as under:

- The major challenge in the area of clickbait detection is that few public datasets are available for the detection of clickbait videos, and the majority of the datasets are available incorporating clickbait headlines. The self-generated dataset[MVD] has been proposed in this paper, which further helps to explore the research in this field.
- Few methodologies have been proposed that aims to detect clickbait video [7]. This is an emerging field and largely unsolved problem, due to which still very few works have been reported in this area. This gives novelty to our work and also motivates us to provide an efficient solution for clickbait detection. Three sets of novel features (Video-based, Comment based, and Channel-based) are reported in this work that are found to be efficient and outperforms other state-of-the-art methods on the same dataset.
- To the best of our knowledge, no work has been reported on the concept of tackling videos having commenting off, where the uploader is not allowing users to comment on their video due to which no important clues can be retrieved from the comments section to predict the video as clickbait or real. Top 15 web headlines are fetched and analyzed to get some important evidence about a video and helps in efficient prediction.
- It has been observed that many of the existing works have considered video metadata instead of extracting some clues from the video transcripts. To the best of our knowledge, we have firstly included the video transcript based feature to get some informative evidence out of it as explained in section 4.2.1.
- The comparative analysis has been done with the other existing algorithms. The results clearly show that the proposed model is superior and outperforms the other existing state-of-the-art.

3 Dataset creation

One of the significant contributions of this work is the dataset creation since few datasets are available. Hence, a dataset of clickbait's and the Non-Clickbait videos have been developed by collecting a diverse set of videos using YouTube REST Data API v3. The details include video content (title, likes, dislikes, views, etc.), Number of comments, Channel details (Number of subscribers, registration date, video count, and view count). In the field of misleading video detection, very

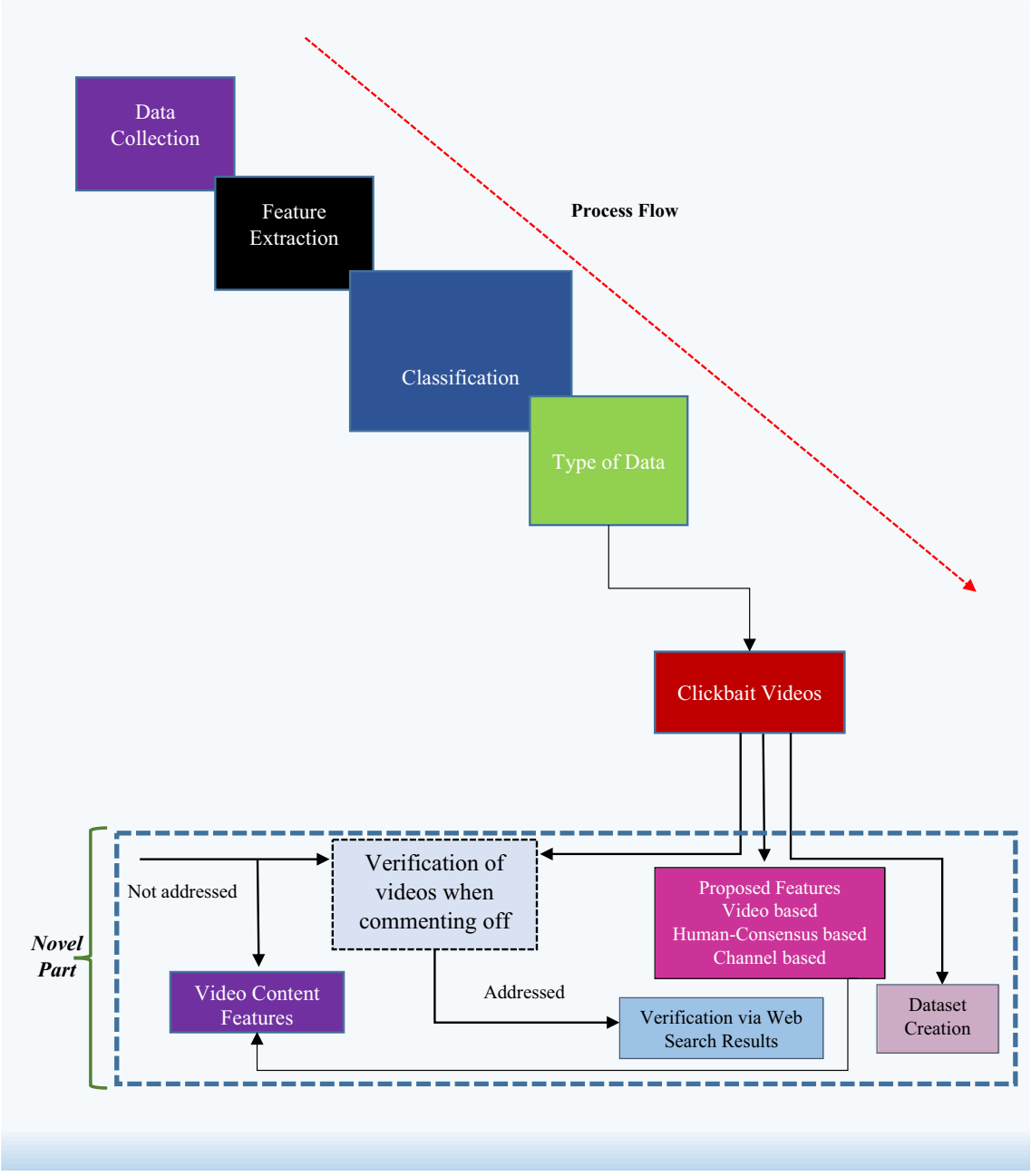


Fig. 3 Flow of Proposed Methodology

few datasets are available, which leads to give the aim to build a generalized dataset, incorporating various categories. From the list of 16 most popular videos category defined by YouTube,[4] we have collected 987 videos(474 Clickbait and 513 Non-clickbait). To collect clickbait videos, we manually crawled and annotated each of the 474 videos. We have analyzed some of the channels as well as their following channels that have prominently posted clickbait's, to lure the user to visit their video. Some of the channels[5] that are posting claim make users curious to visit a link for getting more impression on their video. YouTube has a good check algorithm, for detecting fraudulent videos and also have a blocking mechanism, then also most of the videos are still in their active stage and not removed. From the study, it has been analyzed that most of the clickbait's or hoaxes are posted concerning celebrity's death, which later found to be false and degrades the user experience. As no correct verification is provided, this posted news hamper the public emotions as well. That's why detecting clickbait's video is one of the prominent areas of research. To reduce the time of collecting fake videos, the strategy that we follow is the viral videos, because they are

[4] https://mediakix.com/blog/most-popular-youtube-videos/
[5] https://www.youtube.com/watch?v=zDa-HzCFoIo&t=8s

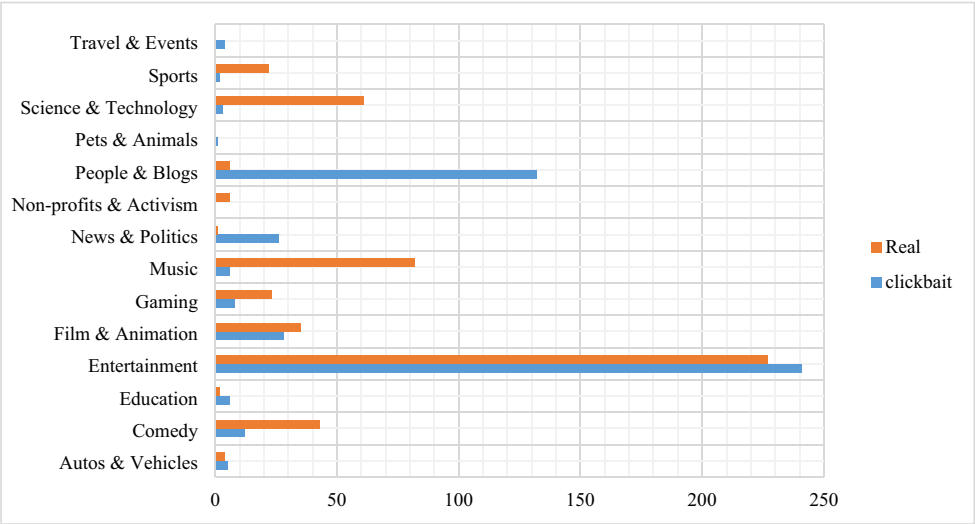


Fig. 4 Number of Videos by Category and Class

more likely to generate fake content, with having catchy headlines also prone to make them viral. We manually analyzed the channels/source for generating these videos.[6]

Additionally, the titles are also scrapped and analyzed to get some clickbait's phrases like some of them are "Shocking", "OMG", "Sad News", "Bad News"," Dukhd Khabar". To direct our search in the correct direction, we try to find out those channels that are following these channels, because it is more likely that the following channel is also posting fake content. Video response also plays an important role while segregating fake/clickbait videos, some of the phrases like "fake video", "bullshit", "galat", "hoax", "clickbait"," Alive"," fake news"," liar", "false", "falsely"," misinformation", "rumor"," clickbait"," hoax" are used, along with it the dislike to like ratio has also been used for further filtering, as it has been observed that clickbait/fake videos received more dislikes compare to likes. For the collection of Non-Clickbait videos, some popular authentic channels have been considered for analysis such as "TEDx Talks", "Harsh Beniwal", "Marvel Entertainment" etc. We call this dataset as "MVD" (Misleading Video Dataset), and the distribution of dataset is as given in Fig. 4 and Table 1. It can be seen from the Fig. 4 that 14 different categories have been considered for the dataset creation, whereas in the dataset, the number of clickbait is found more from the "Entertainment" and secondly with "people and blogs" category, as from the manual analysis it has been observed that most of the clickbait's are prominently available in these categories, while very few videos are considered from the categories("Pets and Animals", "Auto and Vehicles") where the possibility of clickbait generation is quite low.

[6] https://www.youtube.com/channel/UC_UdS7tWCwgBDoaM-Hmkzxg/videos

4 Cognitive evidence for Clickbait's detection

In this section, we present the problem definition for the clickbait detection task and describe the proposed method (CVD), a clickbait video detector to address the problem.

4.1 Problem definition

We define the clickbait's detection task as for a given set of videos, the system has to determine which of the videos are reporting clickbait's and are not faithfully representing the event it refers to. The identification of clickbait videos is ultimately meant to warn users that the given video content is not faithful about the claim it representing, and helps in countering the spreading of false content. In this paper, we have considered the following three detection cases shown below in Table 2.

A set of evidence is required to justify and verify the above cases and warn the user to think twice while believing and spreading false information. If these three cases are identified, then there is a possibility that the video is clickbait and does not faithfully represent the event that it refers to. Formally, the task takes a set of video ids $V_{ID} = V_{ID1}, V_{ID2}, V_{ID3}\ldots. V_{IDN}$ as an input, and the classifier has to determine whether each of these videos V_{IDi} is a clickbait or Non-Clickbait by assigning a label from $Y = \{C, R\}$. Hence, we formulate the task as a binary class classification problem, whose performance is analyzed and evaluated by computing the various performance measures like precision, recall, and F1 score for the target class, i.e., Clickbait. There are three types of cognitive evidence that have been considered for the detection of clickbait. Each of these evidence gives a significant contribution in finding the clues in predicting a video as clickbait or not. The three sets of cognitive evidence for a video are as follows.

Table 1 Detailed Description of the Self-Generated Dataset (MVD)

S.No.	Category	Number of Videos	Class
1	Autos & Vehicles	5	clickbait
2	Autos & Vehicles	4	Non-Clickbait
3	Comedy	12	clickbait
4	Comedy	43	Non-Clickbait
5	Education	6	clickbait
6	Education	2	Non-Clickbait
7	Entertainment	241	clickbait
8	Entertainment	227	Non-Clickbait
9	Film & Animation	28	clickbait
10	Film & Animation	35	Non-Clickbait
11	Gaming	8	clickbait
12	Gaming	23	Non-Clickbait
14	Music	6	clickbait
15	Music	82	Non-Clickbait
16	News & Politics	26	clickbait
17	News & Politics	1	Non-Clickbait
18	Non-profits & Activism	6	Non-Clickbait
19	People & Blogs	132	clickbait
20	People & Blogs	6	Non-Clickbait
21	Pets & Animals	1	clickbait
22	Science & Technology	3	clickbait
23	Science & Technology	61	Non-Clickbait
24	Sports	2	clickbait
25	Sports	22	Non-Clickbait
26	Travel & Events	4	clickbait
Total		474 clickbait 513 Non-Clickbait	

4.2 Detecting Clickbait's videos

In this section, we describe the proposed method, the CVD (Clickbait Video Detector) to address the problem formulated previously. The technique consists of three major pieces of evidence, retrieved using three feature components based on Video, Human-consensus, and User-Profile. The first feature component is used to extract video related features (e.g. speech-title similarity, number of likes, number of dislikes, dislike-like ratio). The second feature component is based on human-consensus. This module learns from individual human cognition and combined from the consensus response. The output has been retrieved, which gives the agreement of individuals towards the posted content. The third component is the user-profile feature extraction (e.g. video-age-ratio, channel views, registration age). This is directly related to the reputation of the video uploader. Lastly, we finish with the classification model, which finally does the binary classification (clickbait and Non-Clickbait) using features extracted from the first three components. The overview of CVD is shown in Fig. 5 where the three sets of features are extracted from the video, comments, and channel information.

Table 2 Possible cases for the clickbait's detection

S.No.	Detection Cases
1.	Title faithfully represents the video speech content and comments are not in contradiction
2.	Title does not faithfully represent the video speech content and both are in contradiction.
3.	Title faithfully represents the video speech content and comments are in contradiction.

4.2.1 Video-content based features

Video-Content based feature is the first component of our proposed method. This component is responsible for the extraction of video-content based feature (e.g. speech-title similarity, number of likes, number of dislikes, dislike-like ratio). The speech-title similarity is one of the crucial features and plays a major role in retrieving Evidence 1. The speech-title similarity is the similar to the speech text with respect to the title of the video, which identifies whether the given claim attached to the video, faithfully represents the event that it refers to. To identify how faithfully the video is representing a claim, speech has been extracted from each video. The Google speech to text API has been used for the speech part, that later be converted into text. The cosine similarity has been applied in between the text extracted from the speech part of the video and the title, to measure the similarity among them. Google's speech to text API has been used for speech recognition. The speech to text has three main methods to perform speech recognition (Synchronous, Asynchronous, and Streaming Recognition).[7] Here we have used the synchronous recognition method, as in our case we need to process the data of less than one 1 min and synchronous recognition requests are limited to audio data of 1 min or less in duration. The detailed description of how the complete process is followed is shown in Fig. 6. It shows the retrieving process of the speech-title similarity for Evidence 1. In the first step, the video is given as an input, which is translated into its audio format. The audio is converted into text format using Google Speech API. So, to get better similarity results, the audio is processed in parts. For each video, the 1 min segments have been analyzed, as we are getting enough information within this duration to predict a video is bogus or credible. The audio transcripts of 1 min are subdivided into 4 parts of each 15 s. The four text parts that are incorporated in Fig. 6 is to split the 1-min audio transcripts into smaller subparts to identify how similar the title/claim with respect to the content that is

[7] https://cloud.google.com/speech-to-text/docs/basics

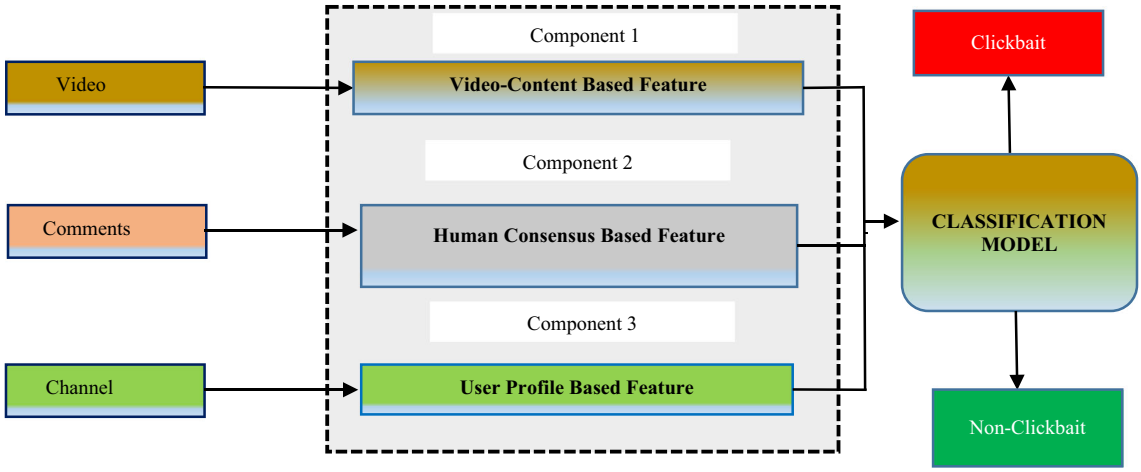


Fig. 5 An Overview of CVD

presented in the video. It has been noticed that in YouTube videos the given titles are too short in length, containing very few words and a 1-min audio transcript considered for analysis is quite lengthy and is not so effective to apply cosine similarity between them for reliable prediction. To address this problem, the audio transcript is split into small subpart (4 text parts of 15 s each) (Text part 1, Text part 2, Text part 3 and Text part 4), each text subpart contains 15-s audio transcripts in sequence after which the individual text similarity with respect to the title has been calculated, later the average value has been considered as the final similarity value.

The video-content based features are shown in Table 3. These features are evaluated to represents the statistical characteristics of the video content.

4.2.2 Human consensus-based feature

Individual human cognition can play an important role and gives a significant contribution in forming evidence for the detection of Clickbait's and the second component of our proposed approach. The individual viewer has their own cognition that has been come out in the form of expression/ emotions given as a video response. Many of the malicious users have not allowed the comments on their video, because the human consensus gives an initial clue about a video, and if any new user visiting the page, they can make an initial thought about video credibility via reading how an individual responds to a video. That's why most of the time, it can be seen that commenting is not enabled on videos created with

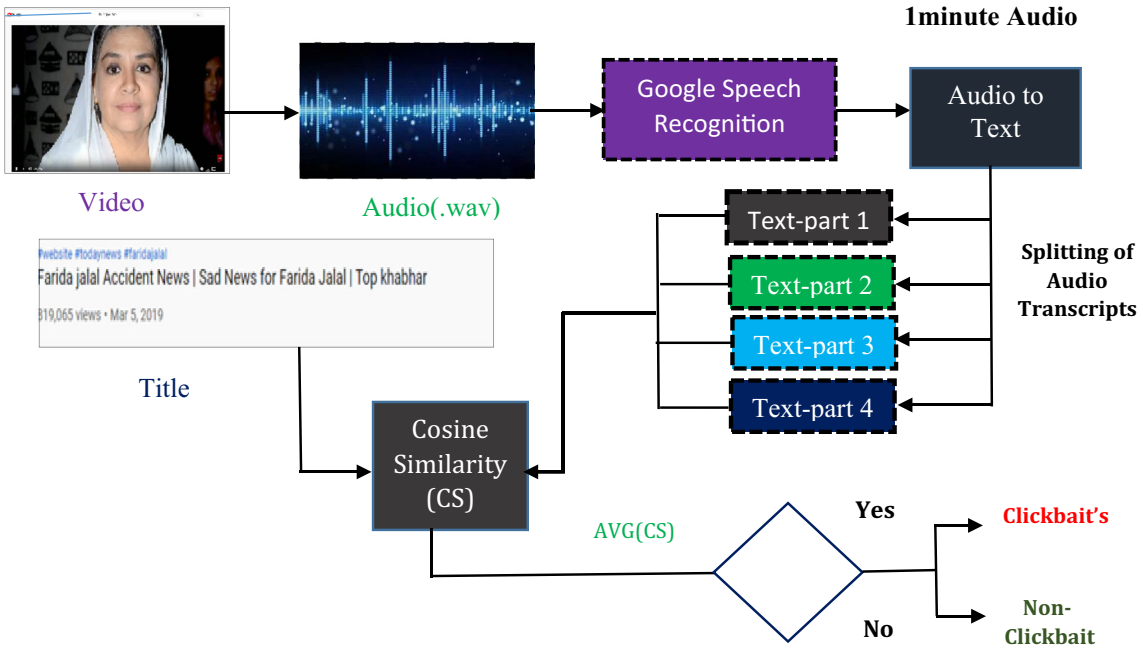


Fig. 6 The Figure represents the process of retrieving Evidence 1

Table 3 Video-Content based feature

Feature	Description
Audio Transcript based Features (Avg_cs)	The average cosine similarity measure between audio transcripts and the title of the video. This is one of the novel features and very few studies incorporate it.
Number of Likes $L(x)$:	This feature represents the number of likes on a video.
Number of Dislikes $D(x)$	This feature represents the number of dislikes on a video.
Dislike to Like Ratio $DL(x)$	The ratio of the number of dislikes to like count on a video. It has been observed that clickbait videos received more dislikes compared to likes. $DL(x)=\frac{D(x)}{L(x)}$
Number of Views	The number of views Received by a video.

some malicious intent. Figure 7 shows the process of retrieving Evidence 2. Here multilingual content has also been addressed. If any of the content is found to be in a different language, it is translated into English text using google translator, then further be used for analysis. As a result, we have also addressed the multilingual content.

A total of 6 Human consensus-based features are extracted. The Human-consensus based features are shown in Table 4. These features are evaluated to represent the statistical characteristics of the responses of the viewers.

4.2.3 User-profile based features

The reputation of the individual channel also plays an important role in identifying the credibility of the uploaded video. (e.g. video-age-ratio, channel views, registration age). A total of 7 User-Profile based features are extracted. The User-Profile based features are shown in Table 5. These features are evaluated to represent the statistical characteristics of the responses of the viewers.

From the all given set of features there are some important findings, it has been observed that the Non-Clickbait video channel has more number of subscribers with respect to registration age as compare to clickbait's video channel. The findings reveal that the average cosine similarity for Non-Clickbait videos is too less than that of the clickbait's, the one of the reason may be, because the clickbait videos probably repeating the same sentence as mentioned in the title, many of the times to extend the length of the video with redundant and bogus content. It has also been found that the number of dislikes is more than the number of likes for many of the videos belongs to clickbait's as compared to non-clickbait's.

4.3 Credible sources

Credible sources are the third sort of evidence, which plays an important role in news verification. The two sets of evidence that we have discussed earlier give significant information; however, they fail in certain situations. The first situation is when the speech-title similarity perfectly matches as well as

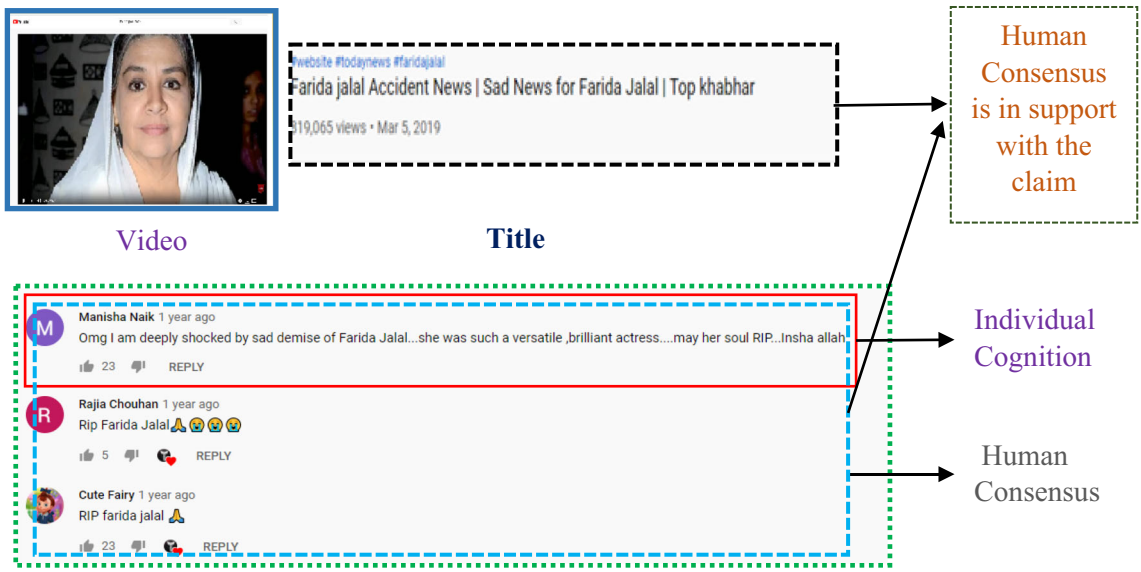


Fig. 7 The Figure represents the process of retrieving Evidence 2

Table 4 Human-Consensus based feature

Feature	Description
Number of Comment c(x)	This feature represents the number of comments received on a video. To restrict our search analysis, in total maximum of 200 comments have been considered. The below equation represent the comment count. $c(x) = \sum_{i=1}^{200}(\text{c i})$
Positive Polarity p(x)	This is the feature that indicates, how many comments showing a positive opinion towards a video. $p(\boldsymbol{x}) = positive\ polarity\ count/c(x)$
Negative Polarity $n(x)$	This is the feature that indicates, how many comments showing a negative opinion towards a video. $n(x) = negative\ polarity\ count/c(x)$
Positive-Negative Polarity Ratio $pn(x)$	This is the ratio of positive to negative comment polarity count. $pn(x) = \frac{p(x)}{n(x)}$
Fake_Comment_Count $FCC(x)$	The fake comment count is the number of comments having clickbait phrases. Clickbait's Phrases(CP) = {fake, bullshit, hoax, wrong… etc.} $FCC(x) = \sum_{i=1}^{200}(\text{CP})$
Fake_Comment_Count Ratio $FCCR(x)$	It is the ratio of the number of fake_comment_count to the total number of comments encountered. $FCCR(\boldsymbol{x}) = FCC/c(x)$

comments are also in supports. So, are these measures sufficient enough? It may happen that, these two pieces of evidence are in support, even though the information is false as shown in Fig. 8 it clearly shows that evidence 1 and evidence 2 are in support, but the information is false. So what is the breaking point here, through which we can reliably conclude about the credibility of news? None other but the credible news sources, we need to scrap the news headlines related to the specified claim by searching it on google, and searching for clickbait phrases in the headlines like {*'misleading', 'misinformation', 'not known', 'no proof', 'no known', 'no scientific evidence', 'no evidence', 'not verified',' hoax', 'clickbait', 'not proven', 'denied', 'deny', 'unverified', 'false', 'fake', 'fake news', 'falsely', 'myth', 'ridiculous', 'rumour',' not dead',' death rumours'*} for justifying the claim. There can be another case of it where it can be applied when the uploader doesn't allow any comments and make the commenting off, where it is quite needful to retrieve Evidence 3. The query is build using the video title concatenated with the fake news keywords *query= "title+ fake news",* that goes as a search query to google. Top 15 URLs concerning the specific claim are scrapped and analyzed. These 15 web titles are considered as a replica of video comments, and the same measures are identified here, as evaluated over video comments (Human-consensus based features) to get the informative clues when comments are not available (Table 6).

Algorithm 1 gives the detailed procedure of the retrieval process of all three pieces of evidence, and from all sets of features, some of the features that play a crucial role in retrieving the evidence are identified. The given algorithm takes a set of videos as an input and returns the status that whether a video is clickbait or not as an output, the given algorithm will further be used to create a Non-Clickbait-time prediction of clickbait. The crucial video-based features encountered are having Average cosine similarity (Avg_cs) and Dislike-like ratio (DL) for retrieving Evidence 1, and from the analysis

Table 5 User-Profile based feature

Feature	Description
Registration Age $r(x)$	The age of the user is an indicative measure of the rounded number of days that the user has spent on YouTube, i.e. from the day account was created up to the day of the current post.
Channel Views $CV(x)$	The total number of views received by the channel.
Total_no_of_Videos $V(x)$	The total number of videos has been posted by the channel till date.
Subscriber Count $SC(x)$	The total number of subscribers count on the channel.
Video_to_Age_ratio $VA(x)$:	This is the ratio of the total number of videos uploaded by the channel to its registration age. $VA(x) = \frac{V(x)}{r(x)}$
Subscribers_to_Age_ratio $SA(x)$	This feature represents the ratio of the number of subscribers on the channel to its registration age. $SA(x) = \frac{SC(x)}{r(x)}$
Channel_Views_to_Subscribers $CS(x)$	It is a ratio of the number of views received by the channel to its subscriber count. $CS(x) = \frac{CV(x)}{r(x)}$

of all set of sample the threshold values have been identified, if the Avg_cs >0.10 OR DL > =0.40 than the video is likely to be clickbait concerning the video content and make the value of Evidence 1 to be true or 1, otherwise 0. On the other hand, FCCR (Fake Comment Ratio) plays an important role in retrieving Evidence 2. After getting the value of Evidence 1 and Evidence 2, three different defined cases have been observed concerning these values and the output status has been predicted as Clickbait or Non-Clickbait.

Algorithm 1.(Clickbait Video Detection)
Input(Video_id) and Output(Status)

```
def. func1(Video_id)
    Evidence1= Video_based_feature ();
    Evidence2= Human_consensus_based_feature ();
    If (Evidence1==0) AND (Evidence2==0):
        Scrapped_urls= Google_search(query)
         Evidence 3= Processing(Scrapped_urls)
         IF (Evidence 3==1):
            Status= Print("Clickbait")
         Else:
            Status= Print("Non-Clickbait")
     Elif (Evidence1==1 and Evidence2==1):
        Status = Print("Clickbait")
def. Video_based_feature ()
  Evidence1 = 0
  Title= Extract_Title ();
  Audio_transcript= Extract_Transcript ();
  number_of_dislike= Count(Dislikes);
  number_of_like= Count(like);
  Averge_cosine_similarity(Avg_cs) = Cosine_similarity (Title, Audio_transcript);
  Dislike_like_ratio(DL)= (number_of_dislike)/ (number_of_like)
    If (Avg_cs>0.10 OR DL>=0.40):
        Evidence1= 1
        return (Evidence 1)
    Else:
        Evidence1=0
        return (Evidence 1)
def. Human_consensus_based_feature ()
    Evidence2= 0
    Number of comment c(x) = ∑_{i=1}^{200}(c i )
    Fake_Comment_Count_Ratio(FCCR)= Fake_Comment_Count /c(x);
    If(FCCR>=0.015)
      Evidence2= 1
      return(Evidence2)
    Else:
       Evidence2= 0
       return(Evidence2)
```

5 Experimental results

This section, we evaluate the performance of the CVD scheme in comparison with the state-of-the-art method. The result shows that the CVD technique significantly outperforms the baseline method with respect to different performance measures.

5.1 Comparison of classifiers on the self-generated dataset [MVD]

Tables 7 and 8 briefly describes the results for different classifiers using all sets of features on the Self-Generated Dataset (MVD) by employing two validation strategies, Cross-validation, and Percentage Split, respectively.

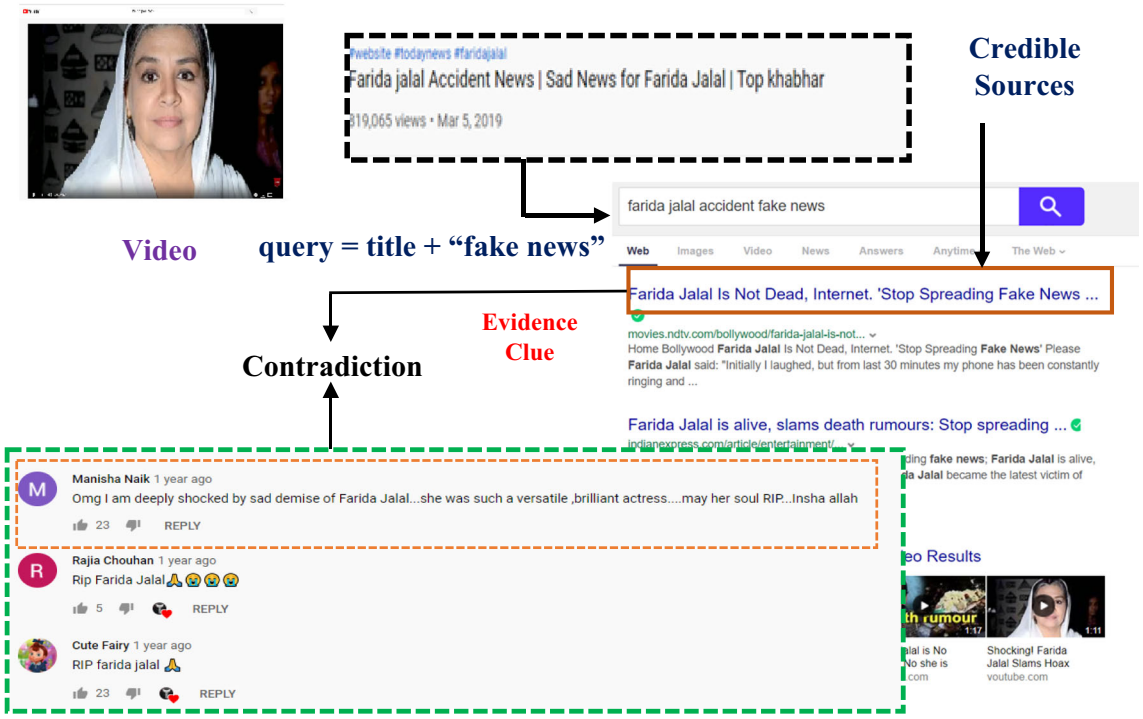


Fig. 8 The Figure represents the process of retrieving Evidence 3

From the rigorous analysis of all the proposed features, it has been observed that Human Consensus and User Profile features give a significant contribution and play a major role in predicting the video as clickbait or non-clickbait. Table 7 shows the performance analysis of the model by employing a cross-validation technique concerning different measures, TP (True Positive), False Positive(FP), PRE(Precision), REC(Recall), FM(F-Measure) and ACC(Accuracy). The analysis has been presented by considering all sets of features, video-based features, human consensus-based features, and user-profile features. The performance of the classifiers (Random-forest, Naïve-Bayes, Logistic, SVM, SGD, k-nearest, and J48 using all set of features and for each independent set of features suggests that J48 remarkably outperforms over the rest of the classifiers with the highest accuracy of 98.28 on both cross-validation and percentage-split mechanism. This can be seen clearly when we look at precision, where J48 performs substantially better than the rest. However, the k-nearest neighbor classifier performs better, only when considering user-profile features. It has been observed that SVM is not performing well when considering only video-based features, while significant improvement in the accuracy, when considering all set of features. Whereas, logistic regression performs worst in comparison to all other classifiers in each scenario. On the other hand, using the percentage split technique, the random-forest, and J48 both performing the same on all sets of features in terms of true positive, precision, recall, f-measure, and accuracy.

Figure 9 shows the comparative analysis of various classifiers (Random-Forest, Naïve-Bayes, Logistic, SVM, SGD, k-

Table 6 The table shows the cases and the needful evidence required for the detection

S.NO	Detection Cases	Essential Measure	Desirable Measure
1.	Title faithfully represents the video speech content and comments are not in contradiction	Evidence 1 Evidence 2 Evidence 3	Evidence 1 Evidence 2 Evidence 3
2.	Title does not faithfully represent the video speech content and both are in contradiction.	Evidence 1	Evidence 1 Evidence 2
3.	Title faithfully represents the video speech content and comments are in contradiction.	Evidence 1 Evidence 2	Evidence 1 Evidence 2

Table 7 Performance of the various classifier by employing Cross-Validation

All Set of Features								
Classifiers	Fold1	Fold2	TP	FP	PRE	REC	FM	ACC
Random Forest	10	–	0.974	0.025	0.975	0.974	0.974	97.37
Naïve Bayes	10	–	0.964	0.034	0.965	0.964	0.964	96.37
Logistic	10	–	0.909	0.085	0.922	0.909	0.909	90.92
SVM	10	–	0.955	0.042	0.959	0.955	0.955	95.46
SGD	10	–	0.955	0.043	0.957	0.955	0.955	95.46
k-nearest	10	–	0.964	0.037	0.964	0.964	0.964	96.37
J48	10	–	**0.983**	0.017	**0.983**	0.983	0.983	**98.28**
Random Forest	–	20	0.974	0.025	0.974	0.974	0.974	97.37
Naïve Bayes	–	20	0.965	0.033	0.966	0.965	0.965	96.47
Logistic	–	20	0.888	0.105	0.905	0.888	0.887	88.81
SVM	–	20	0.955	0.042	0.959	0.955	0.955	95.46
SGD	–	20	0.957	0.041	0.959	0.957	0.957	95.66
k-nearest	–	20	0.965	0.035	0.965	0.965	0.965	96.47
J48	–	20	**0.983**	0.017	**0.983**	0.983	0.983	**98.28**
Video-based features								
Random Forest	10	–	0.875	0.120	0.885	0.875	0.875	87.5
Naïve Bayes	10	–	0.772	0.243	0.830	0.772	0.760	77.21
Logistic	10	–	0.755	0.255	0.773	0.755	0.749	75.50
SVM	10	–	0.542	0.490	0.757	0.542	0.406	54.23
SGD	10	–	0.746	0.264	0.763	0.746	0.740	74.59
k-nearest	10	–	0.876	0.125	0.876	0.876	0.876	87.60
J48	10	–	**0.898**	0.102	**0.898**	0.898	0.898	**89.81**
Human Consensus-based Feature								
Random Forest	10	–	0.884	0.124	0.904	0.884	0.882	88.40
Naïve Bayes	10	–	0.856	0.154	0.883	0.856	0.852	85.58
Logistic	10	–	0.795	0.218	0.845	0.795	0.786	79.53
SGD	10	–	0.823	0.190	0.868	0.823	0.816	82.25
k-nearest	10	–	0.891	0.111	0.893	0.891	0.891	89.11
J48	10	–	**0.902**	0.104	**0.913**	0.902	0.901	**90.22**
User Profile-based features								
Random Forest	10	–	0.955	0.042	0.959	0.955	0.955	95.46
Naïve Bayes	10	–	0.949	0.049	0.952	0.949	0.949	94.85
Logistic	10	–	0.800	0.187	0.857	0.800	0.794	80.04
SVM	10	–	0.955	0.042	0.959	0.955	0.955	95.46
SGD	10	–	0.949	0.048	0.953	0.949	0.949	94.85
k-nearest	10	–	**0.973**	0.027	**0.973**	0.973	0.973	**97.27**
J48	10	–	0.971	0.030	0.971	0.971	0.971	97.07

nearest, and J48) on a different set of features by applying 10-fold cross-validation. The comparison results in term of accuracy measure, clearly show that the model outperforms when employing all three features combinedly, instead of applying individual features. However, it can also be observed that User-profile features individually perform better in comparison to Human-consensus and Video-content based features. At the same time, Video-content based features do not perform well individually. The experimental results reveal that the user-profile based features significantly improves the overall performance of the proposed model compared to other feature sets. One of the main reason identified from the observation that the reputation of an individual channel/account/user profile plays an important role in identifying the credibility of the uploaded video and gives an efficient clue for the verification of misleading content [16, 37]. Previous studies also reveal that user profile based features are efficient in detecting false content.

Table 8 Performance of the various classifier by employing Percentage Split

Classifiers	Split1	Split2	TP	FP	PRE	REC	FM	ACC
Random Forest	70:30	–	0.977	0.024	0.977	0.977	0.977	**97.65**
Naïve Bayes	70:30	–	0.963	0.037	0.964	0.963	0.963	96.30
Logistic	70:30	–	0.903	0.098	0.907	0.903	0.902	90.26
SVM	70:30	–	0.899	0.102	0.916	0.899	0.898	89.93
SGD	70:30	–	0.956	0.044	0.957	0.956	0.956	95.63
K-nearest	70:30	–	0.970	0.030	0.970	0.970	0.970	96.97
J48	70:30	–	**0.977**	0.023	0.977	0.977	0.977	**97.65**
Random Forest	–	80:20	0.975	0.026	0.975	0.975	0.975	**97.47**
Naïve Bayes	–	80:20	0.929	0.067	0.936	0.929	0.929	92.92
Logistic	–	80:20	0.899	0.105	0.907	0.899	0.898	89.89
SVM	–	80:20	0.949	0.054	0.954	0.949	0.949	94.94
SGD	–	80:20	0.960	0.042	0.961	0.960	0.960	95.95
k-nearest	–	80:20	0.970	0.030	0.970	0.970	0.970	96.96
J48	–	80:20	**0.975**	0.025	0.975	0.975	0.975	**97.47**

The authors of [16, 37–41], reports that the user profile/ account-based features plays a significant role in detecting false information. Whereas, the other two features are not performing well compared to this feature due to some constraints like the video-content based features are relying on the similarity of audio transcripts and the title of the video but there are some cases where the audio is too noisy due to which the clear transcripts can't be retrieved for matching or in a case when the video doesn't have any speech content present. These cases may be liable to degrade the performance of the model. Whereas, human-consensus based features are also

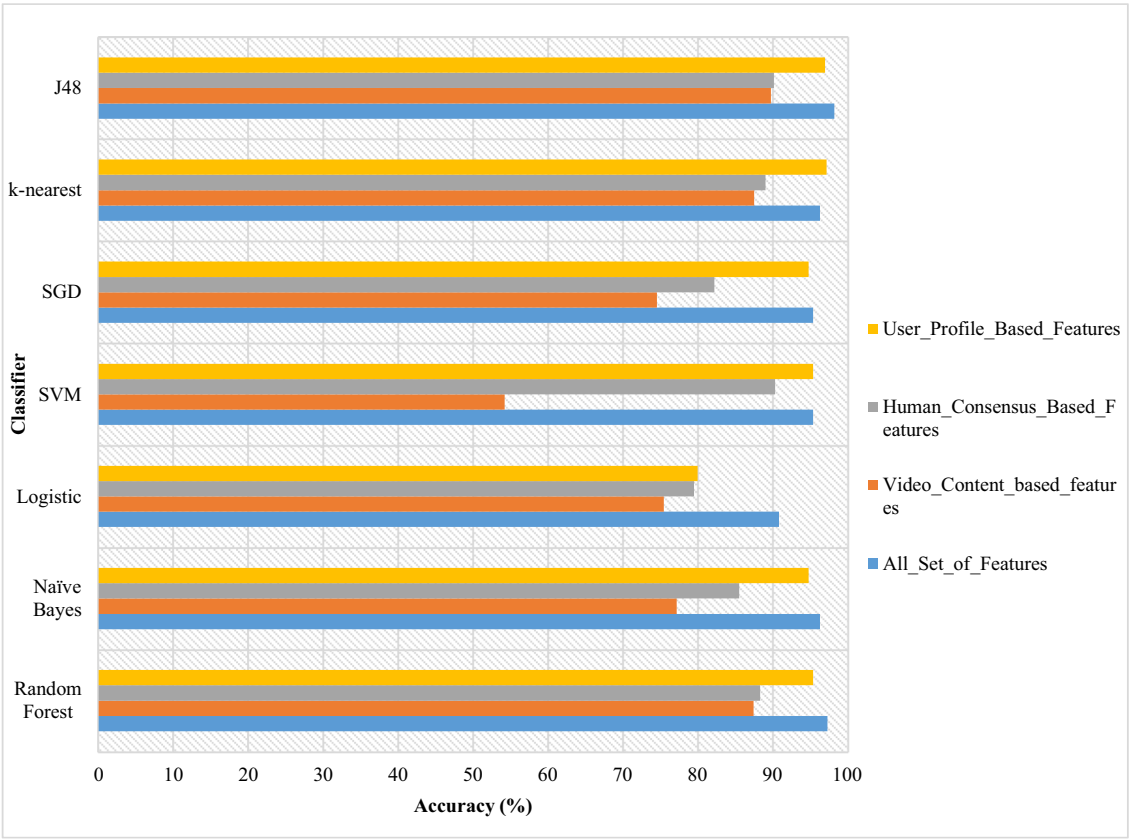


Fig. 9 Comparative analysis of various classifiers on different set of features

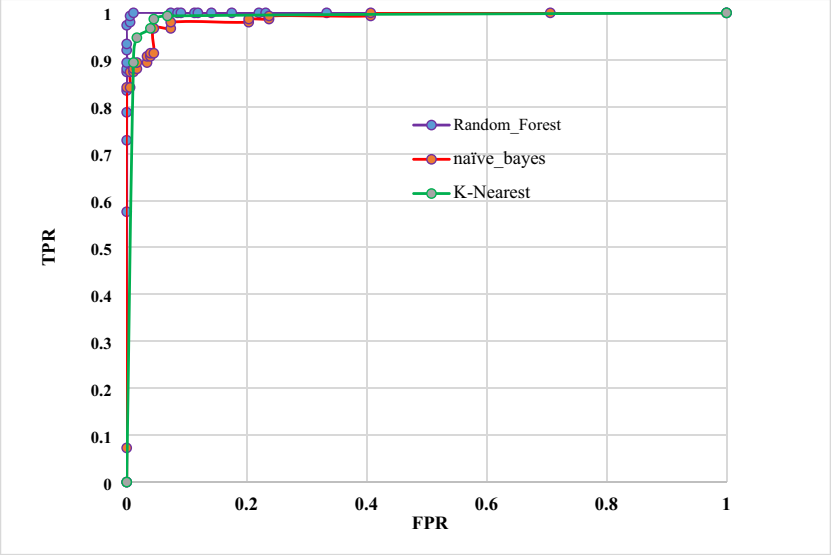


Fig. 10 The AUC-ROC Curve

performing well after user profile features, however in some cases when the sufficient clickbait phrases are not matched/ identified from the user responses or credible link sources in that case this feature may lack in performances.

The AUC-ROC curve using the Random Forest, Naïve-Bayes, and K-nearest neighbor classifier model is shown in Fig. 10. To get a good understanding of the performance of the model, we can also look at their receiver operating characteristics, ROC curves. Figure 10. represents the three ROC curves for the random forest, Naïve Bayes, and K-Nearest Neighbour classifier, trained on all sets of features (video-based, human-consensus, and user-profile based). Here we can see that the peak value for the random forest is achieved at the point (x = 0.01129, y = 1) with having a minimum false positive rate of 0.01. At this point, the model would correctly identify 100% of the true rumors, with only getting around 1% of the false rumors mistakenly classified as true. Whereas, for Naïve Bayes the peak value is achieved at the point (x = 0.401, y = 1), having a minimum false positive rate of 0.406. At this point, the model would correctly identify 100% of the true rumors, with only getting around 40% of the false rumors mistakenly classified as true. Depending upon the application purpose, user can choose or pick up different points on the ROC curve, like with respect to the normal user, who want to find the truthfulness of the specific news, the user could perhaps choose the point on the ROC curve with having minimum false positive rate, or the point where FPR is closer to zero, for getting reliable information.

Along with this, the Scatter plot matrix representation of the features against other features for clickbait and Non-Clickbait data sample is also represented to give a visual qualitative understanding of the correlation. We have created the scatter plot of one feature against another as shown in Fig. 11. Table 9 represents the features on the X and Y-axis of the plot. The plot visually represent the relationship between each of the feature on the set $X = X_1, X_2 \ldots X_9$ on X-axis to the set $Y = Y_1, Y_2 \ldots Y_9$ on Y-axis. The representation is useful, as it is showing the pattern in the relationship between attributes to visually explore the relationship between several numeric values. The dots in the scatter plot are colored by their class value (Clickbait and Non-Clickbait).[8] Like, it can be seen that the scatter plot matrix of subscriber_age_ratio feature on the X-axis against all other features shows an approximately clear separation between two classes (Clickbait's and Non-Clickbait's) and shows how the points are correlated with respect to different classes.

5.2 Comparison of classifiers on publicly available datasets

Along with the self-generated dataset, as we discussed previously, the analysis has also been applied over the other state-of-the-art. It has been observed that some of the work contributed by creating a public dataset of fake/misleading videos [4, 5, 42] . However, still very few datasets are available for comparative analysis, there is small, but some of the datasets of fake videos on YouTube are publicly available called as FVC (Fake Video Corpus) [2] and FVC- 2018 [42]. The dataset[9] is the collection of 381 videos in which 201 are fake, and 180 are Non-Clickbait. After analysis, it has been found that most of the videos are removed from YouTube. Due to

[8] https://machinelearningmastery.com/better-understand-machine-learning-data-weka/

[9] https://github.com/MKLab-ITI/fake-video-corpus/blob/master/FVC.csv

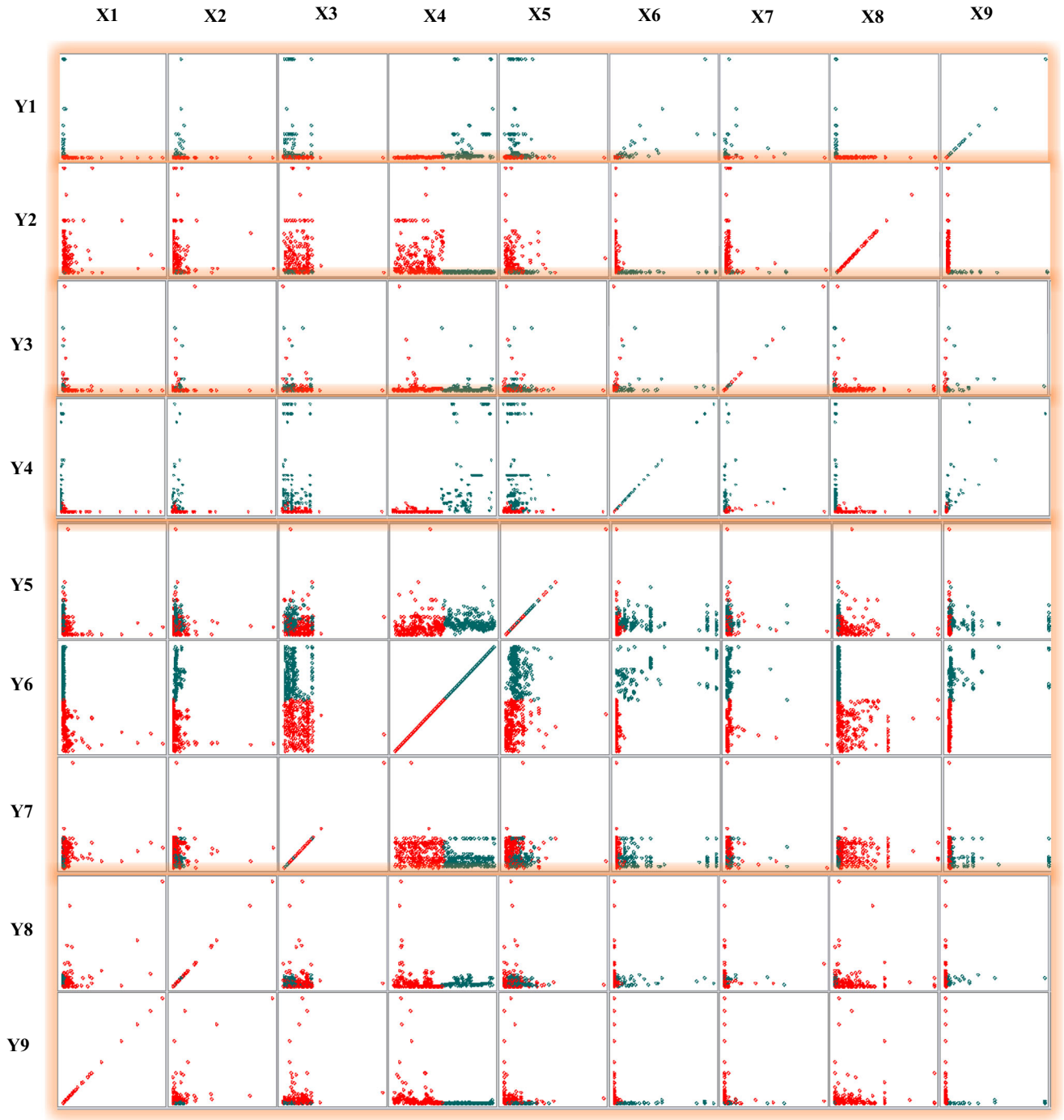


Fig. 11 Plot Matrix representation of proposed features against all other features for Clickbait's (Red) and Non-Clickbait's (Green)

which, we are only able to crawl 84 fake and 90 Non-Clickbait videos, we divide these videos into two disjoint sets, FVC (70:30), with having 70 videos for training and 30 videos for testing. The comparative analysis with the state-of-the-art on the same dataset is shown in Table 10. Whereas, the other dataset that is considered for analysis is FVC 2018. The FVC-2018 is the extended version of the FVC dataset i.e., the samples in the FVC-2018 is an order of magnitude larger than that of FVC, and much more varied. The dataset was extended with 3729 additional fake videos and 2283 real videos, published on YouTube, Facebook or twitter, and considering the time period of April 2006 and June 2018. However for analysis purpose we have considered only YouTube(YT) Videos i.e. 1675(Fake) and 993(Real), comparison analysis has been performed on the same.[10] From

[10] https://github.com/MKLab-ITI/fake-video-corpus/blob/master/FVC_dup.csv

Table 9 Description of features used for the plot of Scatter matrix representation, as shown in Fig. 11

Feature No	Feature	Feature No	Feature
X1	DL(x)	Y1	CV(x)
X2	CS(x)	Y2	FCCR(x)
X3	Avg_cs	Y3	VA(x)
X4	Video_id	Y4	SA(x)
X5	PN(x)	Y5	PN(x)
X6	SA(x)	Y6	Video_id
X7	VA(x)	Y7	Avg_cs
X8	FCCR(x)	Y8	CS(x)
X9	CV(x)	Y9	DL(x)

the previous studies, it has been found that a very limited number of baselines are available for clickbait's video detection [7]. The online clickbait video detection problem is an emerging field and is a largely unsolved research problem. Due to which very few works have been reported yet. Along with this, very limited datasets are publicly available for the evaluation of the proposed algorithm. Many of the works have not released the source code as well. Due to all these research constraints, limited baselines are available for comparative analysis. Some of the prominent methods that are closely related to our work are described in the following paragraph.

The authors of [42], proposed a verification algorithm to detect fake videos. They have also created the FVC-2018 dataset to train and evaluate the proposed method. In the verification algorithm, the author applied the same process that was used in [5], along with it two model variants: a concatenation of the two feature sets(videos metadata and comments feature) and the agreement-based approach given in [43] was used. Their proposed algorithm has been evaluated using 10-fold cross-validation on the dataset proposed by Papadopoulos [5]. and the FVC 2018 dataset with an F1 score of 0.85 and 0.69 respectively. Whereas, the authors of [5], build a classification model using two sets of features(video metadata and comments). Video metadata that specifically considered linguistic features extracted from the video description text and statistics extracted from the video channel. Whereas, the second feature is based on the comments by incorporating a two-level approach. In the first level, features

Table 10 Comparative Analysis with the State-of-the-art

Method	Classifier	Split/ fold	PRE	REC	FM	Dataset
[42] 2019	SVM(Video feature)	10 fold	0.88	0.79	0.82	FVC
	SVM(Comment Feature)	10 fold	0.88	0.74	0.79	FVC
	SVM(Fusion)	10 fold	0.88	0.82	0.85	FVC
	SVM(Video feature)[YT]	10 fold	0.87	0.59	0.70	FVC-2018
	SVM(Comment Feature)[YT]	10 fold	0.91	0.53	0.67	FVC-2018
	SVM(Fusion)[YT]	10 fold	0.79	0.61	0.69	FVC-2018
[5] 2017	SVM (Video Feature)	10 fold	0.88	0.79	0.82	FVC
	SVM (Comment Feature)	10 fold	0.88	0.74	0.79	FVC
	SVM RBF(Fusion)	10 fold	1.00	0.83	0.90	FVC
[44] 2019	Random Forest	70:30	0.74	0.73	0.73	FVC
	Decision- Tree	70:30	0.73	0.67	0.67	FVC
	SVM	70:30	0.56	0.55	0.54	FVC
	Logistic Regression	70:30	0.53	0.53	0.53	FVC
	UCNet	70:30	0.82	0.82	0.82	FVC
Our Method	Random Forest	70:30	0.84	0.78	0.77	FVC
	Decision- Tree	70:30	0.77	0.75	0.73	FVC
	SVM	70:30	0.65	0.63	0.63	FVC
	Logistic Regression	70:30	0.65	0.65	0.65	FVC
	SVM (Video Feature)	10 fold	0.87	0.83	0.83	FVC
	SVM (Comment Feature)	10 fold	0.87	0.83	0.83	FVC
	SVM (Fusion)	10 fold	0.87	0.85	0.85	FVC
	SVM(Video Feature)[YT]	10 fold	0.57	0.57	0.57	FVC-2018
	SVM(Comment Feature)[YT]	10 fold	0.57	0.57	0.57	FVC-2018
	SVM (Fusion)[YT]	10 fold	0.69	0.69	0.69	FVC-2018

are extracted from the individual comment, later the credibility of each comment is evaluated independently using a pre-trained model proposed by the authors of [43]. These two sets of features are used to train the support vector machine classifier. The algorithm is evaluated using 10 fold cross-validation on their proposed dataset with an F1 score of 0.90 on the fusion of both the features. The other comparative analysis has been done concerning the algorithm proposed by the authors of [44], the method is evaluated on 70:30 percentage split scheme, where it has been observed that our proposed work outperforms the method given in [44], considering all three measures(Precision, Recall, and F-Measure) by employing FVC dataset. In [44], the author has proposed an algorithm to counter misleading videos as a supervised classification task. A deep learning-based approach UCNet has been developed, along with it, some simple features are used for the detection of fake videos. It can be seen that the Decision-Tree, SVM, and Logistic Regression classifiers on the proposed approach outperforms the state-of-the-art, except the Random forest classifier. From the above study, it has been observed that most of the reported work mainly employed video metadata and comments based features for the prediction of clickbait videos, however to the best of our knowledge, none of the above-mentioned work has included video related features including video transcript as well as not discussed the similarity among the video title and its content (video transcript), due to which unable to identify how faithfully the video representing the text it claiming to. It has been found from the results outcomes that transcript based features are efficient in improving the model performance and also helps in identifying certain clues about the video credibility that whether it is faithfully representing the same as it claims to. Along with this some of the crucial and novel features are proposed concerning to different feature categories video, comments, and channel that jointly helps in achieving efficient results as described in Section 4.2. The comparative analysis is shown in Table 1; from Table 1 it can be observed that the proposed algorithm outperforms the existing state-of-the-art methods. The proposed approach outperforms the method proposed by [5, 42], concerning both Recall and F-score on 10 fold cross-validation scheme over the FVC dataset. The comparison has been applied to video, comment, and all set of features(fusion), where the proposed model outperforms existing work. On the other hand, while considering the FVC-2018 dataset, it performs better with respect to recall. From the comparative analysis shown in Table 10, the SVM classifier performs better than the previous approach with respect to recall and f-measure and approximate similar with respect to precision when considering the FVC dataset. In addition to this, the SVM also performs well on our proposed dataset (MVD) with an accuracy of 95.4% on 10-fold cross validation as shown in Table 7. However, in case of FVC-2018 dataset, it is less effective on video-based and comment based features individually, whereas by applying fusion of features the classifier performs better with respect to recall and f-score compared to previous approaches. The reason for worse

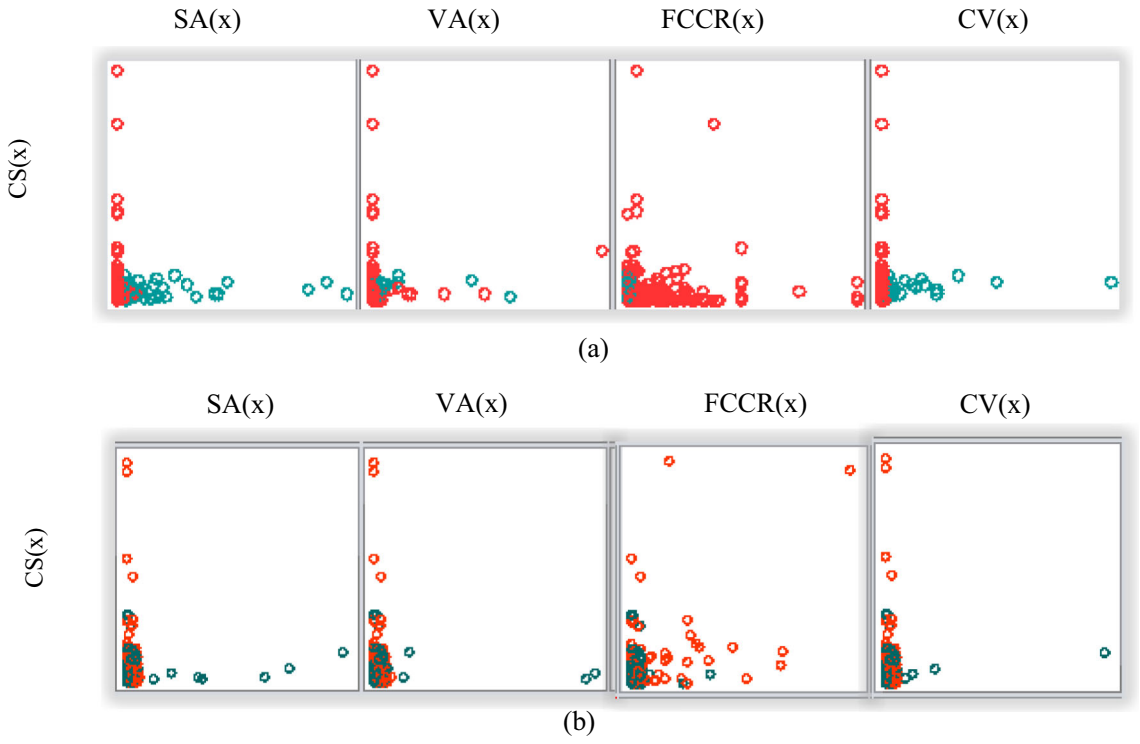


Fig. 12 Plot Matrix representation of features CS(x) against four other features on **a** MVD and **b** FVC-2018 dataset for Clickbait's (Red) and Non-Clickbait's (Green)

performance of the SVM classifier on the FVC-2018 dataset is existence of noise in the feature data. The study also reports that the SVM doesn't perform very well, when the data set has more noise i.e. target classes are overlapping that leads to misclassification of samples. The visiulaization of feature data is as shown in Fig. 12, where scatter plot matrix is presented as an example for some features to show the distribution of target samples on MVD and FVC-2018 Dataset.

The channel view to subscriber ratio(CS(x)) feature has been visualized w.r.t to four other features (subscriber_to_age_ratio SA(x), video_age_ratio VA(x), fake_comment_count_ratio FCCR(x) and channel views CV(x)) to see the distribution of target sample points. From the Fig. 12, it can be clearly noted that the samples of target class are noisy and overlapping in case of FVC-2018 dataset reported in Fig. 12b in comparison to MVD dataset reported in Fig. 12a and due to which we can not be able to get a better decision boundary for classification, and many of the real samples are misclassified as fake, as a result of which it may not end up performing well.

6 Conclusion and future work

In this paper, we develop the CVD scheme to detect clickbait video. The scheme leverages on three components for learning three sets of latent features based on User Profiling, Video-Content, and Human Consensus that further be used to retrieve three sets of cognitive evidence, as an innovative idea for the detection of clickbait videos on YouTube. The set of features are given as an input to machine learning model and performance has been analyzed by considering all set of features and each feature independently by employing a different set of the classifier, and it has been observed that J48 outperforms all other with an accuracy of 98.89% by applying all set of features using cross-validation technique, while 97.47% using percentage split technique on the self-generated dataset [MVD]. The proposed method also performs well on the FVC, FVC-2018 dataset, and outperforms the state-of-the-art with an improved Recall and F score. From the analysis, it has been observed that non-clickbait's video channel has more number of subscribers with respect to registration age as compare to clickbait's video channel. The findings reveal that the average cosine similarity for Non-Clickbait videos is too less than that of the clickbait's, the one of the reason may be because the fake video probably repeating the same sentence as mentioned in the title, of many times to extend the length of the video with redundant and bogus content.

Further work can be enhanced by generating large datasets and employing more video-related features like image frames from the video to get more efficient clues, as well as the clickbait's headlines, can also be analyzed for different other applications like at the time of natural disasters, political elections, healthcare, etc. Along with this, we further extend the work by creating a Non-Clickbait-time application of it.

References

1. Sisodia DS (2019) Ensemble learning approach for Clickbait detection using article headline features. Informing Sci Int J an Emerg Transdiscipl 22:31–44
2. Papadopoulou O, Zampoglou M, Papadopoulos S, Kompatsiaris Y, & Denis Teyssou. (2018). InVID Fake Video Corpus v2.0 (Version 2.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.1147958
3. Zhang DY, Song L, Li Q, Zhang Y, Wang D (2018) Streamguard: A bayesian network approach to copyright infringement detection problem in large-scale live video sharing systems. In 2018 IEEE International Conference on Big Data, Seattle, pp 901–910. https://doi.org/10.1109/BigData.2018.8622306
4. Papadopoulos SA (2019) Towards Automatic Detection of Misinformation in Social Media. arXiv:1909.01543. https://export.arxiv.org/pdf/1909.01543
5. Papadopoulou O, Zampoglou M, Papadopoulos S, Kompatsiaris Y (2017) Web video verification using contextual cues. In Proceedings of the 2nd International Workshop on Multimedia Forensics and Security, pp 6–10. https://dl.acm.org/doi/10.1145/3078897.3080535
6. Potthast M, Köpsel S, Stein B, Hagen M (2016) Clickbait detection. In: Ferro N. et al. (eds) Advances in Information Retrieval. European Conference on Information Retrieval (ECIR) 2016. Lecture Notes in Computer Science, vol 9626. Springer, Cham, pp 810–817. https://doi.org/10.1007/978-3-319-30671-1_72
7. Shang L, Zhang DY, Wang M, Lai S, Wang D (2019) Towards reliable online clickbait video detection: a content-agnostic approach. Knowledge-Based Syst 182:104851
8. Zannettou S, Sirivianos M, Blackburn J, Kourtellis N (2019) The web of false information: rumors, fake news, hoaxes, clickbait, and various other shenanigans. J Data Inf Qual 11(3):1–37
9. Qu J, Hißbach AM, Gollub T, Potthast M (2018) Towards Crowdsourcing Clickbait Labels for YouTube Videos. In HCOMP (WIP&Demo). http://ceur-ws.org/Vol-2173/paper11.pdf
10. Chakraborty A, B. Paranjape, S. Kakarla, and N. Ganguly (2016) Stop clickbait: Detecting and preventing clickbaits in online news media. In IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. pp 9–16. https://ieeexplore.ieee.org/abstract/document/7752207/
11. Chen Y, Conroy NJ, Rubin VL (2015) Misleading online content: recognizing clickbait as false news. In Proceedings of ACM on workshop on Multimodal Deception Detection. pp 15–19. https://dl.acm.org/doi/10.1145/2823465.2823467
12. Biyani P, Tsioutsiouliklis K, Blackmer J (2016) 8 amazing secrets for getting more clicks: detecting clickbaits in news streams using article informality. In Thirtieth AAAI Conference on Artificial Intelligence. https://dl.acm.org/doi/10.5555/3015812.3015827
13. Anand A, Chakraborty T, Park N (2017) We used neural networks to detect clickbaits: You won't believe what happened next! In European Conference on Information Retrieval, pp 541–547. https://doi.org/10.1007/978-3-319-56608-5_46
14. Thomas P (2017) Clickbait identification using neural networks. arXiv preprint arXiv:1710.08721. https://arxiv.org/pdf/1710.08721.pdf
15. Zannettou S, Chatzis S, Papadamou K, & Sirivianos M (2018) The good, the bad and the bait: Detecting and characterizing clickbait on youtube. In IEEE Security and Privacy Workshops, pp 63–69. https://doi.org/10.1109/SPW.2018.00018

16. Yang F, Liu Y, Yu X, Yang M (2012) Semantics, pp 1–7. https://doi.org/10.1145/2350190.2350203
17. Wu K, Yang S, Zhu KQ (2015) False rumors detection on sina weibo by propagation structures. In IEEE International Conference on Data Engineering, pp 651–662. https://doi.org/10.1109/ICDE.2015.7113322
18. Zhang Q, Zhang S, Dong J, Xiong J, Cheng X (2015) Automatic detection of rumor on social network. In Natural Language Processing and Chinese Computing, pp 113–122. https://doi.org/10.1007/978-3-319-25207-0_10
19. Qin Y, Wurzer D, Lavrenko V, Tang C (2016) novelty detection. arXiv preprint arXiv:1611.06322. https://arxiv.org/abs/1611.06322
20. Liu X, Nourbakhsh A, Li Q, Fang R, Shah S (2015) Conference on Information and Knowledge Management, pp 1867–1870. https://doi.org/10.1145/2806416.2806651
21. Wu L, Li J, Hu X, Liu H (2017) Gleaning wisdom from the past: Early detection of emerging rumors in social media. In Proceedings of the 2017 SIAM international conference on data mining, pp 99–107. https://doi.org/10.1137/1.9781611974973.12
22. Kwon S, Cha M, Jung K, Chen W, Wang Y (2013) Conference on Data Mining, pp 1103–1108. https://doi.org/10.1109/ICDM.2013.61
23. Kwon S, Cha M, Jung K (2017) Rumor detection over varying time windows. PLoS One 12:1–19
24. Ma J, Gao W, Mitra P, Kwon S, Jansen B J, Wong K F, Cha M (2016) Detecting rumors from microblogs with recurrent neural networks. In Proceedings of the 25th International Joint Conference on Artificial Intelligence, pp 3818–3824. https://ink.library.smu.edu.sg/sis_research/4630
25. Tacchini E, Ballarin G, Della Vedova ML, Moret S, De Alfaro L (2017) Some like it hoax: Automated fake news detection in social networks. arXiv preprint arXiv:1704.07506. https://arxiv.org/abs/1704.07506
26. Kumar S, West R, Leskovec J (2016) Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In Proceedings of the 25th international conference on World Wide Web, pp 591–602. https://doi.org/10.1145/2872427.2883085
27. Zaman B, Justitia A, Sani KN, Purwanti E (2020) An Indonesian hoax news detection system using reader feedback and Na{\"\i}ve Bayes algorithm. Cybern Inf Technol 20(1):82–94
28. Cao J, Qi P, Sheng Q, Yang T, Guo J Li J (2020) Exploring the Role of Visual Content in Fake News Detection. arXiv preprint arXiv:2003.05096. https://arxiv.org/abs/2003.05096
29. Zhou X, Wu J, Zafarani R (2020) SAFE: Similarity-aware multi-modal fake news detection. arXiv preprint arXiv:2003.04981. https://arxiv.org/abs/2003.04981
30. Ozbay FA, Alatas B (2020) Fake news detection within online social media using supervised artificial intelligence algorithms. Phys A Stat Mech its Appl 540:123174
31. Agarwal A, Mittal M, Pathak A, Goyal LM (2020) Fake news detection using a blend of neural networks: an application of deep learning. SN Comput Sci 1:1–9
32. Vishwakarma DK, Varshney D, Yadav A (2019) Detection and veracity analysis of fake news via scrapping and authenticating the web search. Cogn Syst Res 58:217–229. https://doi.org/10.1016/j.cogsys.2019.07.004
33. Meel P, Vishwakarma DK (2020) Fake news, rumor, information pollution in social media and web: a contemporary survey of state-of-the-arts, challenges and opportunities. Expert Syst Appl 153:1–26. https://doi.org/10.1016/j.eswa.2019.112986
34. Bondielli A, Marcelloni F (2019) A survey on fake news and rumour detection techniques. Inf Sci (Ny) 497:38–55. https://doi.org/10.1016/j.ins.2019.05.035
35. Kumar S, Shah N (2018) False information on web and social media: A survey. arXiv preprint arXiv:1804.08559. https://arxiv.org/abs/1804.08559
36. Zhou X, Zafarani R (2018) A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. ACM Comput Surv 53(5):101–109. https://doi.org/10.1145/3395046
37. Qazvinian V, Rosengren E, Radev DR, Mei Q (2011) misinformation in microblogs. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp 1589–1599. https://dl.acm.org/doi/10.5555/2145432.2145602
38. Shu K, Sliva A, Wang S, Tang J, Liu H (2017) Fake news detection on social media: a data mining perspective. ACM SIGKDD Explor Newsl 19(1):22–36
39. Shu K, Wang S, Liu H (2018) Information Processing and Retrieval, pp 430–435. https://doi.org/10.1109/MIPR.2018.00092
40. Shu K, Zhou X, Wang S, Zafarani R, Liu H (2019) The role of user profiles for fake news detection. In Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp 436–439. https://dl.acm.org/doi/10.1145/3341161.3342927
41. Shu K et al (2018) Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining 8:430–435
42. Papadopoulou O, Zampoglou M, Papadopoulos S, Kompatsiaris I (2019) A corpus of debunked and verified user-generated videos. Online information review 43(1):72–88. https://doi.org/10.1108/OIR-03-2018-0101
43. Boididou C, Papadopoulos S, Zampoglou M, Apostolidis L, Papadopoulou O, Kompatsiaris Y (2018) Detection and visualization of misleading content on twitter. Int J Multimed Inf Retr 7(1):71–86
44. Palod P, Patwari A, Bahety S, Bagchi S, Goyal P (2019) Misleading Metadata Detection on YouTube. In European Conference on Information Retrieval, pp 140–147. https://doi.org/10.1007/978-3-030-15719-7_18

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



Deepika Varshney is a Research Scholar inthe Department of Information Technology,Delhi Technological University India. She has done M.Tech. from IndiraGandhi Delhi Technical University for Women, NewDelhi, India 2017. Her research interest includes OnlineSocial Media Privacy and Security, Machine Learning, and DataScience. She has been awarded with a commendable researchaward in the year 2020 by the Delhi Technological University.



Dinesh Kumar Vishwakarma received the B.Tech. degree from Dr. Ram Manohar Lohia Avadh University, Faizabad,India, in 2002, the M.Tech. degree from the Motilal NehruNational Institute of Technology, Allahabad, India, in 2005,and the Ph.D. degree in the field of Computer Vision and Machine Learning from Delhi Technological University (Formerly Delhi College of Engineering), New Delhi, India, in 2016. He iscurrently an Associate Professor with the Department ofInformation Technology, Delhi Technological University, New Delhi. His current research interests includeComputer Vision, Machine Learning, DeepLearning, Sentiment Analysis, Fake News and Rumour Analysis, Crowd Behaviour Analysis,Person Re-Identification, Human Action and Activity Recognition. He is a reviewer of variousJournals/Transactions of IEEE, Elsevier, and Springer. He has been awarded with "ResearchExcellence Award" by Delhi Technological University, Delhi, India in 2018, 2019 and2020. He is Senior Member of IEEE and life member of ISTE.

2020 IEEE International Conference for Innovation in Technology (INOCON)
Bengaluru, India. Nov 6-8, 2020

Advancements in RNA Secondary Structure Prediction using Machine Learning Methods

Shubham Mittal
Department of Biotechnology
Delhi Technological University
Delhi, India
mshubham9606@gmail.com

Yasha Hasija*
Department of Biotechnology
Delhi Technological University
Delhi, India
yashahasija06@gmail.com
*corresponding author

***Abstract*—New insights and a greater understanding of role of RNA in diseases, cellular processes and regulation could be achieved by acquiring unmitigated information on RNA secondary structures and their base pairing interactions. In this quest, probabilistic models have proven successful over the past two decades, specifically the machine learning based pipelines. These have improved upon alternative thermodynamic based models, folding algorithms employing stochastic parameters and the comparative sequence analysis techniques. The current study delves into the machine learning techniques which have remarkably augmented RNA secondary structure prediction research. An overview of RNA functions, their types, and the roles they play in diseases, has been supplemented, to equip the reader with the required knowledge for understanding the cruciality of uncovering RNA structural information. Additionally, a revision of previous and alternative models has been included for a cohesive understanding of the history and scope of RNA structure prediction.**

Keywords—Secondary structure prediction of RNA, machine learning methods, probabilistic models

I. INTRODUCTION

Serving a storehouse of genetic information, like the DNA, the single stranded RNA is a cardinal component of the chromosomal machinery in living organisms. The nucleotide bases (adenine, guanine, cytosine, and uracil), ribose sugars and the phosphodiester bonds are essential components that contribute to the chain conformation of RNA. While being the progenitors of proteins, following the central dogma of molecular biology, RNA are also integral to gene regulation, and controlling cellular mechanisms.

Conventionally three primary RNA types describe RNA functions – the mRNA, tRNA and rRNA. The mRNA (messenger RNA), synthesized in the nucleus by transcribing the DNA, is the shipper of genetic information and leads to synthesis of proteins after passing through ribosomes. They travel to the cytoplasm to serve as templated for translation. The rRNA (ribosomal RNA) are molecules that make up the ribosomes – the fundamental units that carry out translation, along with proteins. Ribosomal proteins are synthesized by rRNA in the nucleolus of the nucleus. The ribosomes once fully assembled travel to the cytoplasm and translate the 'codons' in mRNA to amino acid molecules. This conversion is enabled by the tRNA (transfer RNA) molecules also known as activator RNA. A three nucleotide sequence make up a codon; each codon sequence is unique from the view that it links a unique amino acid to the protein chain being synthesized [1].

But going by the new data being produced through research on RNA in the last four decades, RNA have many more essential roles, such as – regulating DNA replication, catalyzing reactions in cells by serving as Ribozymes (RNA as enzymes), facilitating multifarious cellular processes in disease induced state of the body and regulating other diverse and complex mechanisms.

Beside rRNA, tRNA and the mRNA, several other forms of RNA exist. Non-coding RNA (ncRNA) and coding RNA (cRNA) i.e mRNA are two main classes of RNA. ncRNA are categorized into - 'housekeeping ncRNA' that include the rRNA and tRNA; and 'regulatory ncRNA'. Depending on length a ncRNA can be 'long ncRNA' with a sequence that is longer than 200 nucleotides or it can be a 'small ncRNA' having a sequence that is smaller than 200 nucleotides. Small ncRNA are known to be of gret import in terms of functional significance. Five types of small ncRNA exist – small-interfering RNA (siRNa), microRNA (miRNA), small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), and PIWI-interacting RNA (piRNA). Out of these five, the miRNA are of great significance since they regulate genes, are found in most. miRNAs function by inhibiting translation to proteins by binding to specific mRNA sequences, a process known a gene silencing. This enables them to pose as potential biomarkers to diagnose several diseases such as cancer in human beings while dysregulation of RNA can cause Alzheimer's and similar diseases [2], [3]. While mRNA are endogenous, siRNA – another type of RNA causing gene silencing (RNA interference) – are exogenous. The gene expression of transposons or the 'jumping genes' is regulated by piRNAs, which are specific to target transposon sequences and carry out silencing thereby thwarting their transcription into germ cells. Circular RNAs enable protein synthesis, bind to miRNA preventing gene silencing, regulate alternative splicing and are formed due to fusion of 5' and 3' ends of the strand [4], [5].

RNAs have an important role to play in evaluating diseases in human beings and finding cures. Several miRNAs have been found to play a role in upregulating cancer associated genes inducing tumour growth. While dysregulated miRNA induces several brain disorders and Alzhiemer's. Non-coding RNAs such as the tRFs (tRNA-derived fragments) have also been linked with induction of cancer [3]. Furthermore, tRNA have been found to inhibit caspases – proteins that lead to apoptosis

978-1-7281-9744-9/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:03:09 UTC from IEEE Xplore. Restrictions apply.

of cells – thereby inducing cancer implicitly. Presence of RNAs can also confirm a disease, an example being MALATI (metastasis-associated lung adenocarcinoma transcript 1) – RNA transcripts whose increased levels in a cell are indicative of cancerous nature of cells [6]. In addition, some classes of RNA can separate RNA binding proteins that amass in neural tissues leading to neurological diseases like ALS (amyotrophic lateral sclerosis) [7] and myotonic dystrophy.

The single stranded nature of RNA adds to its low stability, high energy and high reactivity, unlike DNA which is double stranded and more stable. Due to this the 2′-hydroxyl group on RNA – which enable linking of ribose sugar to phosphate group forming phosphodiester bond – is more susceptible to getting cleaved due to RNA hydrolysis. However, the self-folding of RNA causing three dimensional structures stabilize it in the cellular milieu. Self-folding of RNA is due to complement base pairing of nucleotides on the same strand following both canonical and non-canonical rules. The types of structures formed include nested and non-nested base pairing, triplet interactions, hairpin loops, stem-loops, internal and external loops, multiloops, bulges, helices, and pseudoknots (refer Fig. 1). The three-dimensional RNA structure besides adding to its stability is also apt for binding to proteins and other molecules, attributing multiple functions to RNA. Easy attachment of methyl group to this 3D conformation further adds to an RNA's stability. An example is tRNA structures with and without methyl group attached to their 58th position. While the former remains stable and functional, the latter gets degraded with time by becoming non-functional. RNA also act as catalysts when in conjugation with ribonucleoproteins (RNPs) [8], [9].

The discovery of RNA complex 3D structures for the first time by physical techniques such as X-ray crystallography and NMR, ushered in the understanding that RNAs owe their interacts with other molecules and performance of various functions to such conformations. Subsequently RNAs were found to exist largely as tertiary structure and secondary structures (refer Fig. 2), however since the prediction of RNA tertiary structures from primary structures proves to be extremely complex, owing to the dynamic cellular milieu and conformations, RNA research focuses on secondary structures prediction from primary structures. Still, majority of RNAs are non-coding and their functions and structures remain unknown. Therefore, research streamlined upon accurate prediction of base-pairing in RNA which would divulge vital information regarding the positions and nucleotides on an RNA where self-complementarity occurs and the sites which are free from it; ultimately this would culminate in developing RNA secondary structure prediction models.

While the pioneer experimental techniques (X-ray crystallography, nuclear magnetic resonance (NMR), cryogenic electron microscopy), developed to ascertain 3D structures of RNA were highly successful, they are time taking, expensive and require expert handling which resulted in very slow progress. Through these methods structures of less than 0.01 percent of 14 million known non-coding RNA were deposited in RNAcentral [10]. Therefore, developing much faster, cheaper and relatively precise computational methods to

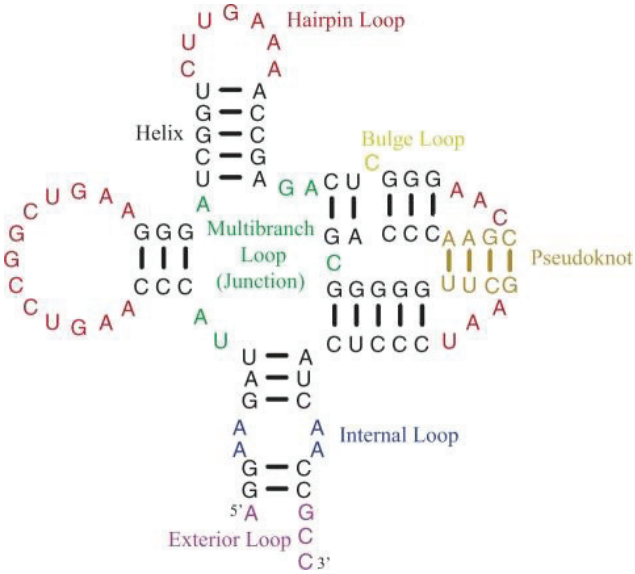


Fig. 1. Diagrammatic representation of self-folded RNA Secondary Structure with numerous types of loops (Image Source: Turner DH. et al.,

explicitly predict RNA secondary structures from sequences serve as more lucrative methods.

Other physical techniques developed to predict RNA secondary structure experimentally were – PARS [11], DMS [12], SHAPE model [13]. The computational methods for RNA secondary structure prediction can be classified into - comparative sequence analysis methods and methods involving folding algorithms with scoring schemes of statistical, thermodynamic, or probabilistic nature [10]. Comparative sequence analysis methods determine conservation of base pairs in homologous sequences; some well-known CSA methods are Sankoff method [14], Knudsen & Hein method [15] and Alali & Sagot method [16]. The success of these methods is based on the availability of wide number of homologous sequences but only about 3000 families are known in Rfam [17] which is very low. As a result, alternative techniques were developed such as creation of folding algorithms which utilize appropriate scoring functions to predict RNA secondary structures. The minimum free energy (MFE) model employs the nearest-neighbor algorithm and dynamic programming to estimate global minimum by creating substructures on the lines of loops and stems [10]. RNAshapes [18], RNAstructure [19], RNAfold [20] models are variations of this method based upon thermodynamically determined parameters. ContextFold [9], CONTRAfold [21] CentroidFold [22] models are all famous examples of another method for secondary structure that employs probabilistic techniques to obtain parameters to base the model upon. These scoring functions can either be attained through random probability-based statistical schemes, using stochastic context-free grammars (SCFGs), or they can be obtained through machine learning techniques (deep learning).

Machine Learning can be defined as the division of Artificial Intelligence in which intricate mathematical models are designed by training machines on provided data (known

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:03:09 UTC from IEEE Xplore. Restrictions apply.

examples). The objective of the models is to make predictions upon unknown data, by recognizing patterns. Modeled after the Biological Neural Networks in the human brain ANN (Artificial Neural Networks) are machine learning models; width of an ANN signifies maximum number of neurons in a layer while the number of hidden layers are indicated by the length of an ANN. 'Deep neural networks' indicate ANN which are capable of being trained with a greater number of hidden layers. Convolution Neural Networks employ 'convolution' (linear mathematical operation) in lieu of general matrix multiplication; CNN are widely used for analyzing images and they are a category of deep neural networks. Recurrent Neural Networks – another type of ANN – retain memory of previous steps in time. They are multilinear and their main advantage is that each layer is akin to observations made in a time series; output of previous time steps determines output of present layer and each layer performs the same work as every time step, which is why they are called 'recurrent'.

II. RNA SECONDARY STRUCTURE PREDICTION METHODS USING MACHINE LEARNING

Machine learning methods are increasingly finding use in predicting RNA secondary structures due to their ability to predict several candidate secondary structures computationally, in less time and with relatively significant accuracy. On a fundamental level these models computationally formulate parameters for scoring substructures of RNA, which prove less strenuous than thermodynamically derived scoring parameters. Described below are landmark machine learning algorithm-based RNA secondary structure prediction tools.

A. CONTRAfold Model

Earlier, the minimum free energy (Mfold [24]) methods or the stochastic context-free grammar (SCFGs) methods which came later, were popular parameter based estimation methods for predicting RNA secondary structure due to relatively higher accuracy of these approaches. The MFE methods were based on experimentally derived thermodynamic parameters which provided for short sequences, the structure that was most optimal. The SCFG methods employed joint probability dependent and statistical algorithms which gave modest

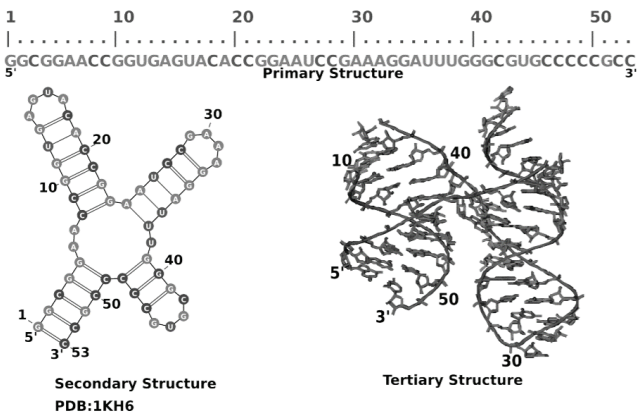


Fig. 2. Representation of primary, secondary and tertiary structures of RNA; PDB:1KH6 (Image Source: Kim N. et al., 2013)

accuracy though lesser than the empirical techniques. But with subsequent research it has come to notice that RNA does not necessarily prefer the most optimal conformation i.e. the conformation with MFE. Besides, there was also a growing need for higher accuracy delivering methods that also accounted for prediction of structures of longer RNA sequences.

In 2006, CONTRAfold became the most popular tool to predict secondary structures of RNA; it primarily used CLLMs (conditional log-linear models) and proved more accurate than experimental MFE methods. CLLMs were relatively more flexible probabilistic models which also made use of feature-abundant scoring and discriminative training including parameters that were modeled after MFE dependent thermodynamic models, and combined these with learning and optimization parameters of SCFGs. CONTRAfold incorporates a newly developed estimator known as 'Maximum Expected Accuracy' (MEA). At the time of the model's release it showed unprecedented performance, better than all other probabilistic models, although without considering pseudoknots [21].

B. CentroidFold Model

The previous Minimum Free Energy (MFE) dependent models which used Maximum likelihood (ML) estimators for prediction, were characterized by low probabilities of prediction and an absence of optimal predictions [24]. Consequently, a troupe of possible RNA structures were considered by the SCFG models for better accuracy of prediction resulting in the proposal of novel estimators. These comprised of the 'centroid' estimator, from Sfold [25] and MEA estimator from CONTRAfold [21]

In view of this approach, in the current model, the authors tried to upgrade preceding estimators, all the while aiming for higher accuracy; this they then proved both experimentally and theoretically. A new estimators known as γ-centroid estimator was proposed as a result [22]. The tool further accommodates for input of RNA in the form of both – a multiple sequence alignment with other RNA (CLUSTAL-W format) and as a single sequence (FASTA format). The tool further boosted the importance of latest probabilistic techniques and novel estimators. However, like the previous model, pseudoknots were also knot accounted for in CentroidFold.

C. ContextFold Model

With an increasing focus on machine learning-based algorithms for parameter estimation for RNA secondary structure prediction, a constant endeavor to improve these methods endured. It was observed that the number of parameters in such models remained almost the same over the years and across several tools, even as the algorithms for prediction improved. For this the authors proposed 'rich parameterization' model in ContextFold, where the information utilized for structure prediction was increased, implying an increase in the number of parameters (about 70,000 free parameters).

The added advantage of advanced RNA folding algorithms, increasing availability of RNA sequential [26], [27] and experimentally determined structural data (larger training sets) [28] and progressions in machine learning techniques [29], [30] all contributed to the increasing parameterization. The structural elements defined by the ContextFold tool are

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:03:09 UTC from IEEE Xplore. Restrictions apply.

founded on those in Turner99 model [31]. This, in addition to analyzing long sequences, and more structural elements resulted in a novel folding prediction model with a notably higher quality of predictions [9].

D. SPOT-RNA Technique

Machine learning-based probabilistic models before SPOT-RNA were successful in predicting, with good accuracy, the canonical and nested base pairs towards RNA secondary structure prediction. However, the accuracy of prediction could not cross 80% ceiling in any of these. This was because these models ignored base pairing which was stabilized by tertiary interactions such as – unstacked (solitary), non-canonical (that do not follow Watson-Crick base-pairing of A-U, G-C and G-U), and non-nested base pairs in addition to triplet interactions [4], [32].

However, SPOT-RNA is a more holistic model inclusive of all types of base pairing including those specific to tertiary structures. The software employs deep learning techniques with convolutional neural network (CNNs) layers, two-dimensional Bidirectional Long Short-Term Memory (2D-BLSTM) layers and merging these with the use of 'transfer learning' from a model trained with bpRNA dataset of higher quality, containing more than 10,000 nonredundant RNA obtained through comparative analysis; this improved upon the limitation of having <250 RNA structures of high resolution – a limitation that deterred previous studies [10]. The model improved not just the accuracy of RNA structure prediction, but also of sequence alignments and functional annotations. The model also improves upon the computation time, even though the sequences used as inputs in training are of limited length.

III. Supplementary Methods

A. Tertiary structure prediction of RNAs

Less attention was allotted to nucleotide interactions that contributed to the tertiary structure of RNA, in the standard models for secondary structure prediction (except SPOT-RNA). This resulted in a performance ceiling of 80% in those models, which in turn motivated the research community to develop models for secondary structure prediction based on tertiary interactions as additional parameters [10].

These models are divided into those that predict secondary structures with pseudoknots (e.g. pknotsRG [33], Probknot [34], IPknot [35], and Knotty [36]), and those that predict secondary structures with non-canonical base pairs (e.g. CycleFold [37], MC-Fold [38], and MCFold-DP [39]). However, the accuracy of these methods is lower than popular models and none of them considers both pseudoknots and non-canonical base pairs for prediction, and neither does any model consider base triplets and lone base pairs.

B. Combined Techniques

Popular predictive models are demarcated according to algorithms – comparative sequence analysis, and folding algorithms (thermodynamic, stochastic or probabilistic). Few attempts have also been made to predict RNA structures by merging two algorithms. For example, CDP-Fold [41] algorithm employs both convolutional neural networks (CNNs) and dynamic programming (MFE). While suggesting modest accuracy, such algorithms require further optimization and incorporation of more parameters to predict RNA secondary structures satisfactorily.

C. Inverse Folding of RNA

Efforts have also been made towards designing RNA sequences based upon target secondary structures that are required to perform certain functions. Several advanced algorithms for inverse folding of RNA exist, such as downhill simplex algorithm, multi-objective genetic algorithm, local search algorithm, etc. RNA inverse [42], Frnakestein [43], EteRNABot [44], IncaRNAtion [45], RNAiFold [46], and MODENA [47] are among popular inverse folding tools.

IV. Discussion

Some limitations of machine learning based probabilistic models cannot be discounted while evaluating their advantages over minimum free energy (MFE) and stochastic context-free grammar (SCFG) methods, in attempting to predict RNA secondary structures. Overfitting is caused when the model trains excessively on the training data, which tarnishes the authenticity of the predictions. ContextFold [48] is a classic example of such a model with overfitting. Low diversity in terms of types of RNA data available in databases can also contribute to overfitting. Statistical models for secondary structure prediction would see an improvement with the creation of curated single-sequence RNA secondary structures databases with multifarious RNA data.

TABLE I. Comparison of popular RNA Secondary Structure Prediction algorithms

Name	Prediction of Pseudoknot	Machine learning method	Type of Algorithm	Ref[a]
Sankoff (1985)	No	No	Comparative Sequence Analysis	[14]
Knudsen & Hein (1999)	No	No	Comparative Sequence Analysis	[15]
Alali & Sagot (2005)	No	No	Comparative Sequence Analysis	[16]
RNAshapes (2014)	No	No	Thermodynamic	[40], [18]
RNAstructure (2010)	Yes	No	Thermodynamic	[19]
RNAfold (2011)	No	No	Thermodynamic	[20]
CONTRAfold (2006)	No	Yes	Probabilistic	[21]
CentroidFold (2009)	No	Yes	Probabilistic	[22]
ContextFold (2011)	No	Yes	Probabilistic	[9]
SPOT-RNA (2019)	Yes	Yes	Probabilistic	[10]
Knotty (2018)	Yes	No	CCJ 2.0	[36]
IPknot (2011)	Yes	No	MEA and Heuristic Integer Programming	[35]
MC-Fold and MC-Sym Pipeline (2008)	Yes	No	MC-Fold and MC-Sym	[38]
CDP-Fold (2019)	Yes	Yes	Probabilistic and Dynamic Programming	[41]

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:03:09 UTC from IEEE Xplore. Restrictions apply.

While the idea behind employing computational methods for structure prediction is based upon their advantages over physical methods, the foundation of such former is experimentally determined RNA structures, upon which a machine learning model build. Therefore, an increase in the number of significant experimentally identified RNA structures with high-resolution is expected to boost veracity and precision of probabilistic models. This brings to light a drawback of computational models in general – they are not completely independent of experimental techniques of identifying RNA structures. A model with adequate experimentally determined structural data and a vigorous probabilistic model to predict structures of unknown input RNA sequences would seem ideal. Lastly, the 'black box' problem in machine learning impairs an insight into the process of actual feature extraction in a model prior to the output layer. This prevents an improvement of the models from a biological standpoint. Specific employment of LSTMs and GRUs (types RNNs) in probabilistic models, may result in higher accuracy of RNA secondary structure prediction, since they effectively address the issue of long-term dependency.

V. Conclusion and Scope

RNA secondary structure prediction has much to contribute to biotechnological research, of which it would certainly find applications in genomic studies and disease prevention. Sites on mRNA where ribosomes can effectively bind could be identified, which would open the gates for gene regulation, gene, and potential gene therapy for treating diseases. RNA binding proteins play significant roles in post-transcription, splicing, and translation; identifying protein binding sites on RNA would open avenues to regulate these processes. Furthermore, binding sites for miRNA in Eukaryotes and siRNA in Prokaryotes could be identified through prediction of accessible sites on secondary structure of mRNA. This could then provide the opportunity to hijack RNA interference, regulate it and also foster research into 'RNA therapy', which is aimed at designing drugs to target specific RNA. A better healthcare system could be developed, more skilled at treating diseases, through a detailed understanding of RNA and disease, established through RNA secondary structure prediction. Finally, an enhancement of current machine learning models through incorporation of more biologically oriented features and detection of nucleotides involved in self folding in RNA would steer us into age of advanced healthcare through novel technologies.

References

[1] J. Brosius and C. A. Raabe, "What is an RNA? A top layer for RNA classification," RNA Biol., 2016.

[2] K. N. Holohan, D. K. Lahiri, B. P. Schneider, T. Foroud, and A. J. Saykin, "Functional microRNAs in Alzheimer's disease and cancer: Differential regulation of common mechanisms and pathways," Frontiers in Genetics. 2013.

[3] C. Battaglia et al., "Candidate genes and MiRNAs linked to the inverse relationship between cancer and alzheimer's disease: Insights from data mining and enrichment analysis," Front. Genet., 2019.

[4] [4] J. Nowakowski and I. Tinoco, "RNA structure and stability," Semin. Virol., 1997.

[5] L. Chen and G. Wong, "Transcriptome Informatics," in Encyclopedia of Bioinformatics and Computational Biology, Elsevier, 2019, pp. 324–340.

[6] N. Amodio et al., "MALAT1: A druggable long non-coding RNA for targeted anti-cancer approaches," Journal of Hematology and Oncology. 2018.

[7] B. Swinnen, W. Robberecht, and L. Van Den Bosch, "RNA toxicity in non‐coding repeat expansion disorders," EMBO J., 2020.

[8] K. Darty, A. Denise, and Y. Ponty, "VARNA: Interactive drawing and editing of the RNA secondary structure," Bioinformatics, 2009.

[9] S. Zakov, Y. Goldberg, M. Elhadad, and M. Ziv-Ukelson, "Rich parameterization improves RNA structure prediction," J. Comput. Biol., 2011.

[10] J. Singh, J. Hanson, K. Paliwal, and Y. Zhou, "RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning," Nat. Commun., 2019.

[11] M. Kertesz et al., "Genome-wide measurement of RNA secondary structure in yeast," Nature, 2010.

[12] Y. Ding, Y. Tang, C. K. Kwok, Y. Zhang, P. C. Bevilacqua, and S. M. Assmann, "In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features," Nature, 2014.

[13] K. A. Wilkinson et al., "High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states," PLoS Biol., 2008.

[14] D. Sankoff, "SIMULTANEOUS SOLUTION OF THE RNA FOLDING ALIGNMENT AND PROTOSEQUENCE PROBLEMS.," SIAM J. Appl. Math., 1985.

[15] B. Knudsen and J. Hein, "RNA secondary structure prediction using stochastic context-free grammars and evolutionary history," Bioinformatics, 1999.

[16] J. Allali and M. F. Sagot, "A new distance for high level RNA secondary structure comparison," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2005.

[17] I. Kalvari et al., "Non-Coding RNA Analysis Using the Rfam Database," Curr. Protoc. Bioinforma., 2018.

[18] S. Janssen and R. Giegerich, "The RNA shapes studio," Bioinformatics, 2015.

[19] J. S. Reuter and D. H. Mathews, "RNAstructure: Software for RNA secondary structure prediction and analysis," BMC Bioinformatics, 2010.

[20] R. Lorenz et al., "ViennaRNA Package 2.0," Algorithms Mol. Biol., 2011.

[21] C. B. Do, D. A. Woods, and S. Batzoglou, "CONTRAfold: RNA secondary structure prediction without physics-based models," in Bioinformatics, 2006.

[22] K. Sato, M. Hamada, K. Asai, and T. Mituyama, "CentroidFold: A web server for RNA secondary structure prediction," Nucleic Acids Res., 2009.

[23] S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics," Briefings in bioinformatics, vol. 18, no. 5. pp. 851–869, 2017.

[24] M. Zuker, "Mfold web server for nucleic acid folding and hybridization prediction," Nucleic Acids Res., 2003.

[25] Y. Ding, C. Y. Chan, and C. E. Lawrence, "Sfold web server for statistical folding and rational design of nucleic acids," Nucleic Acids Res., 2004.

[26] Y. Wexler, C. Zilberstein, and M. Ziv-Ukelson, "A study of accessible motifs and RNA folding complexity," in Journal of Computational Biology, 2007.

[27] R. Backofen, D. Tsur, S. Zakov, and M. Ziv-Ukelson, "Sparse RNA folding: Time and space efficient algorithms," in Journal of Discrete Algorithms, 2011.

[28] S. Griffiths-Jones, S. Moxon, M. Marshall, A. Khanna, S. R. Eddy, and A. Bateman, "Rfam: Annotating non-coding RNAs in complete genomes," Nucleic Acids Res., 2005.

[29] M. Collins, "Discriminative training methods for hidden Markov models," 2002.

[30] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive-aggressive algorithms," J. Mach. Learn. Res., 2006.

[31] R. Kierzek, M. E. Burkard, and D. H. Turner, "Thermodynamics of single mismatches in RNA duplexes," Biochemistry, 1999.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:03:09 UTC from IEEE Xplore. Restrictions apply.

[32] E. Westhof and V. Fritsch, “RNA folding: Beyond Watson-Crick pairs,” Structure. 2000.

[33] J. Reeder and R. Giegerich, “Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics,” BMC Bioinformatics, 2004.

[34] S. Bellaousov and D. H. Mathews, “ProbKnot: Fast prediction of RNA secondary structure including pseudoknots,” RNA, 2010.

[35] K. Sato, Y. Kato, M. Hamada, T. Akutsu, and K. Asai, “IPknot: Fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming,” Bioinformatics, 2011.

[36] H. Jabbari, I. Wark, C. Montemagno, and S. Will, “Knotty: Efficient and accurate prediction of complex RNA pseudoknot structures,” Bioinformatics, 2018.

[37] M. F. Sloma and D. H. Mathews, “Base pair probability estimates improve the prediction accuracy of RNA non-canonical base pairs,” PLoS Comput. Biol., 2017.

[38] M. Parisien and F. Major, “The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data,” Nature, 2008.

[39] C. Höner zu Siederdissen, S. H. Bernhart, P. F. Stadler, and I. L. Hofacker, “A folding algorithm for extended RNA secondary structures,” Bioinformatics, 2011.

[40] Z. J. Lu, J. W. Gloor, and D. H. Mathews, “Improved RNA secondary structure prediction by maximizing expected pair accuracy,” RNA, 2009.

[41] H. Zhang et al., “A new method of RNA secondary structure prediction based on convolutional neural network and dynamic programming,” Front. Genet., 2019.

[42] I. L. Hofacker, “Vienna RNA secondary structure server,” Nucleic Acids Res., 2003.

[43] R. B. Lyngsø, J. W. J. Anderson, E. Sizikova, A. Badugu, T. Hyland, and J. Hein, “Frnakenstein: Multiple target inverse RNA folding,” BMC Bioinformatics, 2012.

[44] J. Lee et al., “RNA design rules from a massive open laboratory,” Proc. Natl. Acad. Sci. U. S. A., 2014.

[45] V. Reinharz, Y. Ponty, and J. Waldispühl, “A weighted sampling algorithm for the design of RNA sequences with targeted secondary structure and nucleotide distribution,” in Bioinformatics, 2013.

[46] J. A. Garcia-Martin, P. Clote, and I. Dotu, “RNAiFOLD: A constraint programming algorithm for rna inverse folding and molecular design,” J. Bioinform. Comput. Biol., 2013.

[47] A. Taneda, “Multi-objective optimization for RNA design with multiple target secondary structures,” BMC Bioinformatics, 2015.

[48] E. Rivas, R. Lang, and S. R. Eddy, “A range of complex probabilistic models for RNA secondary structure prediction that includes the nearest-neighbor model and more,” RNA, 2012.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:03:09 UTC from IEEE Xplore. Restrictions apply.

Multimedia Tools and Applications
https://doi.org/10.1007/s11042-020-10341-6

Affective state recognition from hand gestures and facial expressions using Grassmann manifolds

Bindu Verma[1] · **Ayesha Choudhary**[2]

Received: 24 May 2020 / Revised: 30 September 2020 / Accepted: 22 December 2020 /
Published online: 20 January 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC part of Springer Nature 2021

Abstract

The emotional state of a person is important to understand their affective state. Affective states are an important aspect of our being "human". Therefore, for man-machine interaction to be natural and for machines to understand people, it is becoming necessary to understand a person's emotional state. Non-verbal behavioral cues such as facial expression and hand gestures provide a firm basis for understanding the affective state of a person. In this paper, we proposed a novel, real-time framework that focuses on extracting the dynamic information from a videos for multiple modalities to recognize a person's affective state. In the first step, we detect the face and hands of the person in the video and create the motion history images (MHI) of both the face and gesturing hands to encode the temporal dynamics of both these modalities. In the second step, features are extracted for both face and hand MHIs using deep residual network ResNet-101 and concatenated into one feature vector for recognition. We use these integrated features to create subspaces that lie on a Grassmann manifold. Then, we use Geodesic Flow Kernel (GFK) of this Grassmann manifold for domain adaptation and apply this GFK to adapt GGDA to robustly recognize a person's affective state from multiple modalities. An accuracy of 93.4% on FABO (Gunes and Piccardi 2006) dataset and 92.7% on our own dataset shows that integrated face and hand modalities perform better than state-of-the-art methods for affective state recognition.

Keywords Affective state · Facial expression · Hand gesture · Grassmann manifold · Emotion recognition · Geodesic Flow Kernel

1 Introduction

Understanding affective state of a person is an important aspect of social intelligence, which is the human ability to understand and recognize social signals and social behaviors.

✉ Ayesha Choudhary
ayeshac@mail.jnu.ac.in

Bindu Verma
binduverma67@gmail.com

[1] Delhi Technological University, New Delhi, India

[2] School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi, India

Affective state refers to the experience of feeling the underlying emotional state. Happy, sad, anger, anxiety, boredom, contempt, neutral are the different affective states of a person. Non-verbal behavioral cues such as facial expressions, hand gestures, body postures, etc. form a part of social signals. These social signals are observable behaviors that convey information about feelings, personality, and affective state of an individual as well as the nature of social relationships [1, 2, 34]. Facial expressions and hand gestures are informative behavioral cues that are considered *honest signals* [31] because they are usually produced outside our conscious awareness and provide reliable information about the actual inner mental state, intentions, and feelings of a person [1]. Our subspace learning framework can also overcome the effect of aging and appearance changes in face [5, 6] which can otherwise adversely affect the performance of the models for affective state recognition. Affective computing [32] and social signal processing [45] are important areas that are based on the automatic inference of emotions perceived through machine detectable behavioral cues such as facial expressions, hand and body gestures. Social intelligence is the ability to understand and manage these behavioral cues or *social signals*.

Affective state refers to the experience of feelings, emotions, intentions, and mood of a person. When people communicate with each other, they use multiple non-verbal behavioral cues such as facial expression, hand gesture, and body gestures to interact and these cues help in affective state recognition of a person. Affective computing deals with emotionally intelligent machines that can understand and sense the affective states of humans and take informed decisions. Affective computing aims at providing effective and natural interaction between humans and computers. One important goal is to enable computers to understand the affective states expressed by the human subjects so that personalized responses can be delivered accordingly. Advancements in camera technology and computing power, as well as the reduction in the cost of these devices makes it possible to develop systems for automatic video analytics that is capable of affective state recognition and enables the user/machine to take informed decisions. Human affective state recognition finds application in areas of psychology, neuroscience, cognition studies, and computer science because of its important role in human communication, interaction, advanced driver assistance systems (ADAS), and mental well-being. Human-robot interactions (HRI) would be more natural if the robots could understand and respond to our social signals, thereby understanding our affect [16, 38].

Nowadays, the world is growing towards digital technology and it is very important that we should have intelligent systems that monitor a person's affective state that will help in many areas such as video surveillance, monitoring patient's emotional and depression level, monitoring driver's affective state, and in the advertisement industry, etc. These requirements demand a generic framework that should be error free and works with high-accuracy. Thus, this paper aims to recognize the affective state (emotions) of a person using facial expressions and hand gestures. Psychological studies [26] show that although face expression analysis is important for affective state recognition, the observation of multiple modalities such as hand gestures, body postures improve the recognition of the affective state of the subjects. The study by Gu et al. [18], shows that facial expressions along with hand gestures provide better recognition of the affective state when compared with recognition of affective states by a single modality.

Therefore, in our work, we focus on extracting the dynamic information from videos containing both facial expressions and hand gestures to recognize the affective state of the subject. First, the face is detected using Viola-Jones face detector [46] in each frame and the detected face (Region of Intreset(ROI)) is used to create the motion history images

whichh we call as the face MHI. Similarly, hands are detected in each frame using real-time hand-detection using neural networks (SSD) on Tensorflow [15] and the hand ROI is used to create motion history which we refer to as hand MHI. These motion history images encode the spatio-temporal information of the face and hands as they move across space and time. We use fine-tuned CNN (ResNet-101) to extract the hand and face modality features and these features are concatenated to form a monolithic features vector containing information about both face expressions and hand gestures. Then, these features are used to create subspaces for each affective state. Geodesic flow kernel works on the subspaces on a Grassmann manifold and integrates an infinite number of subspaces from source to target domains. We apply this kernel function to adapt Grassmann Graph Embedding Discriminant Analysis (GGDA) [21] and map the source and target domains to another manifold by creating similarity graphs such that the inter-class discriminatory power increases. The advantage of using subspaces and Grassmann manifold based recognition is that it captures the variations in different instances of the same expression and gesture carried out by the same person as well as the variations across different persons. Experimental results on publicly available datasets show that our method outperforms most of the state-of-the-art methods.

The main highlights of our paper are:

- By creating the motion history image of face and hands, we extract the temporal dynamics of both modalities. MHI helps to overcome the challenges involved in robustly tracking the moving modalities across frames and also reduces the computation time.
- We propose a new hybrid model in which features are extracted using a convolutional neural network and affective states are recognized using Grassmann manifold framework.
- We have adapted Grassmann Graph Embedding Discriminant Analysis and applied the Geodesic Flow Kernel (GFK) to create the similarity graphs in GGDA. The use of GFK kernel reduces the problem of data mismatch between test and training data.

The rest of the paper is organized as follows. In Section 2, we introduce related work on affective state using physiological signals and visual modalities. In Section 3, we describe the details of our proposed method. The experimental settings, results are discussed in Section 4 and experimental discussions are presented in Section 5. Finally, the conclusion of this paper is discussed in Section 6 with future directions.

2 Related work

In this section, we discuss the related work in the area of affective state (emotions) recognition using physiological signals. Electro-dermal activity, Electro-cardiogram, and skin temperature variation are the physiological signals that replicate the influence of affective states on the autonomous nervous system [24]. Dellaert et al. [14] proposed a framework to classify human affective state using vocal features. Similarly, Roy et al. [35] used features such as measures of pitch and energy to classify the emotions using a Fisher linear classifier. EEG signal are used in many areas to recognize the emotions [25], to control the robot or wheelchair task [41, 42], and in affective state recognition [27].

Kim et al. [24] proposed an affective state classification framework using physiological signals such as skin temperature variation, electro-dermal activity, and heart rate to extract the features, and Support Vector Machine (SVM) is used to classify the emotions. Similarly, Koelstra et al. [25] used EEG signal frequency to detect the emotions of a person. Kim

et al. [23] used physiological signal electromyogram, electrocardiogram, skin conductivity, and respiration change to extract the features and linear discriminant analysis to classify the emotions. Similarly, Lisetti et al. [27] used physiological signals such as skin temperature, pulse esteems, galvanic skin reaction, and EEG signs to categories the basic emotion using k-NN and discriminant classifier.

Affective state recognition using speech and physiological signal has been an active area of research. However, there are certain limitations of using physiological and speech signals. Speech-based methods require the person to speak and it may not be possible for the user to be vocal under certain emotional conditions, or may not work for people with speech disorders. Moreover, spoken words and tones are easier to control, and may not always reflect the true feelings of a person. In physiological signal based systems, the user needs to wear some devices to record the signals that may make them uncomfortable and hinder natural behavior. However, vision-based systems for emotion recognition are unobtrusive and do not require the cooperation of the user.

In general, in video based affective state recognition, face expression and hand gesture features are extracted for further analysis. Bastanfard et al. [7] proposed an e-appearance system to predict the essential effects of facial images of different appearances using wrapping technique. Shan et al. [37] proposed affective state recognition using facial and body gestures only. They extracted spatio-temporal features and fused the body and facial features using Canonical Correlation Analysis (CCA). Kapoor et al. [22] used multi-model information such as facial expression, postural shift, and learn activity to classify the interest and disinterest of children solving the puzzle problem in the computer. Cohen et al. [13] used a multi-level Hidden Markov Model (HMM) to capture information and classify the expressions based on the probability of the next stage. Their model also segments long videos into different expression classes. Gunes et al. [19] show the comparative study of mono-model and bi-modal affect recognition using face and upper body features and classification using the Weka tool. In another paper, Gunes et al. [20] proposed an affective state recognition using fiducial points tracking and tracking the shoulder, arms, and head of the subjects. They use selective fusion techniques to fuse the features from multiple modalities using a Support Vector Machine (SVM). Similarly, Chen et al. [11] used both local motion and appearance motion features to classify the affective state using SVM. They also did a comparative analysis with temporal normalization and bag of words. However, both the methods use manual feature fusion which may not use an inner correlation between face and body gestures.

Yang et al. [48] used a weighted geodesic flow kernel for internal emotion recognition in dyadic interaction. Similarly, in another paper, Yang et al. [47] used GFK only in emotion recognition in dyadic interactions. Although, both the works have used GFK for emotion recognition, however, they focus on emotion recognition in case of interaction between a pair of individuals, while we focus on affective state recognition of an individual. They have considered body gestures and acoustics as multiple modalities, where the body gestures are in the form of "Motion Capture Markers", while we work on videos of people wherein a frame has both the face and the gesturing hands in most cases. Although, traditional methods set a benchmark and achieve good accuracy in affective state classification, but extracting hand-crafted features while tracking the modalities may not result in robust recognition.

In recent times, deep learning-based models have worked tremendously in various areas such as nuclear waste object detection and classification [39], EEG classification [28], players affective recognition [33], affective state recognition [40], depression monitoring [44] and solving puzzles [22]. Barros et al. [3, 4] proposed a hierarchical visual representation of a multi-channel Convolutional Neural Network (CNN) for multi-model emotional state

recognition. Tzirakis et al. [43] used CNN for affect recognition using audio and visual modalities. Features are fused by long short-term memory networks and the affective states are classified. Similarly, Chen et al. [12] proposed a multi-convolutional deep learning neural network to recognize the emotion using face and body gestures. Nguyen et al. [30] proposed a framework for affective state recognition. They used 3D convolutional neural network (C3D) and a deep belief network (DBN) to extract the features from videos and audios. They used bilinear pooling theory to fuse the audio and visual features extracted from C3D and DBN to classify the recognize emotions. Bo et al. [40] used hierarchical deep spatio-temporal features over handcrafted features with bilateral long short-term memory recurrent neural networks to recognize the affective states.

Affective state monitoring is actively applied in many areas such as children's activity monitoring, pain management, players' emotions identification, depression analysis, etc. In children's activity monitoring Kapoor et al. [22] used the multi-model affective state to classify the interest and disinterest of children solving the puzzles on a computer. Other applications proposed by Ellgring et al. [36] is to analyze the emotions of the professional actors using facial expression, body movements, vocal expression, and body gestures. Psaltis et al. [33] proposed a framework in gaming to identify the affective state of players using a deep network model. Joy et al. [8] developed a framework for automatic chronic-pain assessment from human-pain behavior and expressive behaviors. Azher et al. [44] used spatio-temporal features of facial motions and used bi-directional long short-term memory (Bi-LSTM) to predict the depression level of a person. Similarly, [9] used EEG signals to detect the depression in a person.

3 Proposed framework

In this work, we detect the face and gesturing hands of the preson (subject of interest) in the video and create separate motion history images of both the face and hands. We then extract the features from each of the MHIs using two deep networks, to overcome the issues associated with handcrafted features. These features are concatenated to form a monolithic feature vector containing information of both the facial expressions and hand gestures in the video. These feature vectors are then used to create subspaces for each state. These subspaces capture the variations in both facial expressions and hand gestures and lie on a Grassmann manifold. The subspaces formed by the labeled training data form the source domain. Since the test data is unlabeled and of different subjects (not the same as training data) we consider the corresponding feature vector as a subspace and these subspaces are elements of the target domain. Ee use unlabelled data of different subjects in the target domain. We use the Grassmann Geodesic Flow Kernel between the source domain and the target domain. This situation is an example of unsupervised domain adaptation since there are no labels associated with the target domain which is taken from unseen data that is not present in the training set. We compute the Geodesic Flow Kernel (GFK) [17] and use this kernel to adapt the Grassmann Graph embedding Discriminant Analysis (GGDA) [21] framework for more robust and accurate recognition. The intuition behind using GFK for improved GGDA is that since the geodesic flow captures the shortest path between the two points on a Grassmann manifold, GFK will increase the inter-class and intra-class discriminatory powers when mapping the subspaces from one manifold to the other in GGDA. We recognize the affective state from multi-modal data using this form of GGDA. Figure 1 shows the flowchart of our proposed framework for affective state recognition from face and hand gesture modalities. We will now describe each step in detail.

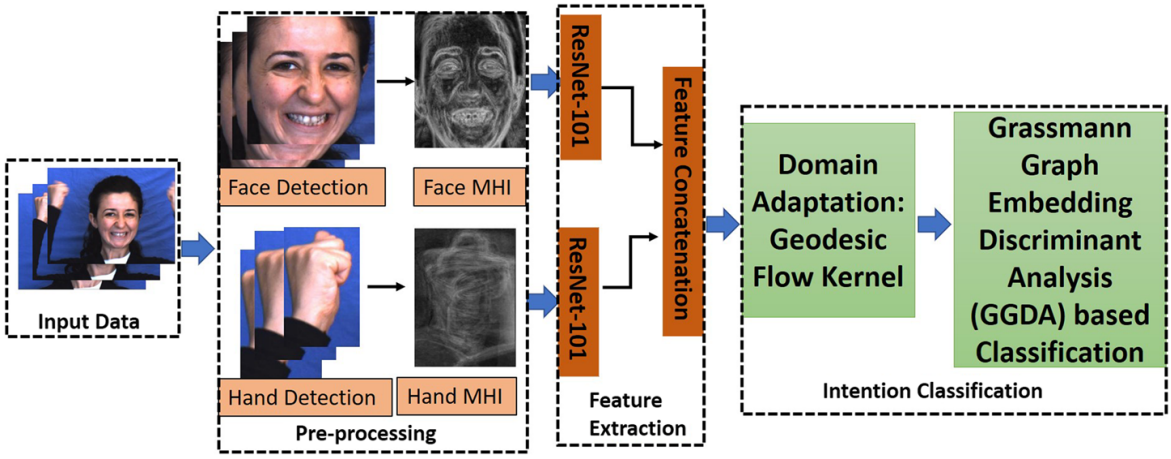


Fig. 1 Flowchart of our proposed approach. Proposed module is divided into three parts **a** face and hand detection **b** feature extraction **c** domain adaptation for unlabeled data and GGDA based classification

3.1 Face detection

To learn the features corresponding to the facial expressions, we first detect the face in each frame using Viola-Jones face detector [46] for each video. The face ROI is cropped and saved for further processing. This helps in focusing on the face and reduces the computation cost in both space and time for further processing. Figure 2a shows frames from different videos and (b) shows the corresponding detected face.

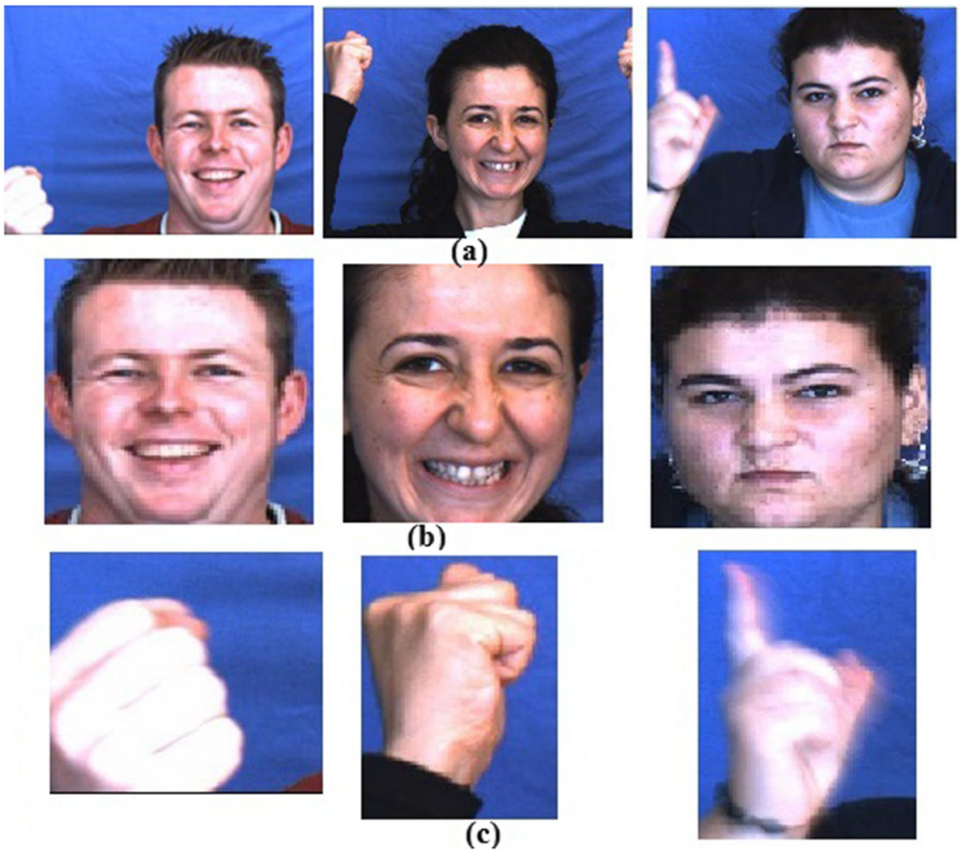


Fig. 2 **a** Input image for face and hand detection **b** Detected face image (Face RoI) **c** Detected hand image (Hand RoI)

3.2 Hand detection

Similarly, hand detection helps to focus on the hand movements across the video and reduces the computation cost in both space and time for further processing. For learning the features corresponding to hand gestures, we first detect the hands in each frame using real-time hand detection using neural networks (SSD) on TensorFlow [15] (this model trained the TensorFlow Object Detection API using the neural network on Ego hands Dataset). After the hands are detected, we crop the detected hand ROI from each frame and save them for further processing. Figure 2a shows frames from different videos and Fig. 2c shows the detected hand ROI.

3.3 Motion history calculation

We use the detected face ROI in each frame to create a motion history image (MHI) of the facial expression in each video, which we refer to as Face MHI. Similarly, for the detected hand ROI in each frame, we create an MHI of the hand movements. In general, to create the motion history of a video, the absolute difference between current and previous frames is computed and summed for all the frames of the video as given by (1). Examples of Motion history images in our experiments on FABO dataset [19] are shown in Fig. 3 for face expression and in Fig. 4 for hand gestures.

$$MHI = \sum_{i=1}^{N} \parallel Frame_{i-1} - Frame_i \parallel \theta_s \tag{1}$$

where N is the number of frames in the video, $Frame_i$ and $Frame_{i-1}$ are current and previous frames, respectively and θ_s is a weighted shadow and is defined as $\theta_s = i/t$, where, t is the constant that defines the number of frames required to represent the MHI. The absolute difference between the current and previous frames removes the background and the regions that are not part of the moving object. Summing up the differences across the frames of the video represents the motion pattern, called the Motion History Image (MHI).

3.4 Network Fine-tuning and Feature Extraction

In our proposed approach, we use the pre-trained Convolutional Neural Network (CNN) (ResNet-101) to extract high-level features. ResNet-101 has high performance on the ImageNet dataset and is computationally less expensive compared to the VGG16 and VGG19 networks [10]. Therefore, using ResNet-101 helps in to improve the computational cost without compromising on the performance. Moreover, the dimension of the feature vector

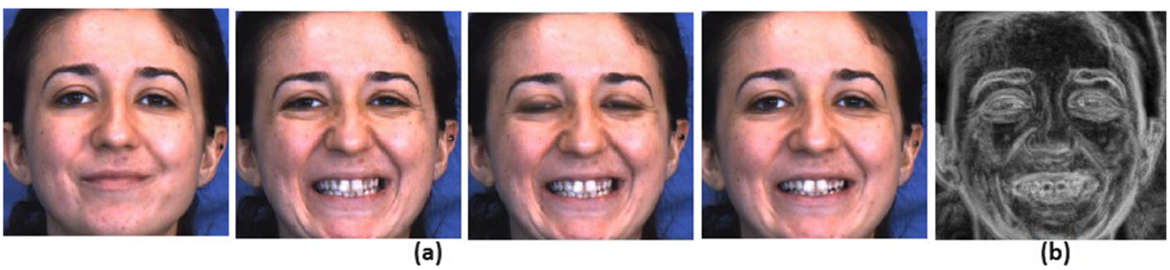


Fig. 3 Face MHI generated by the face RoI in a video

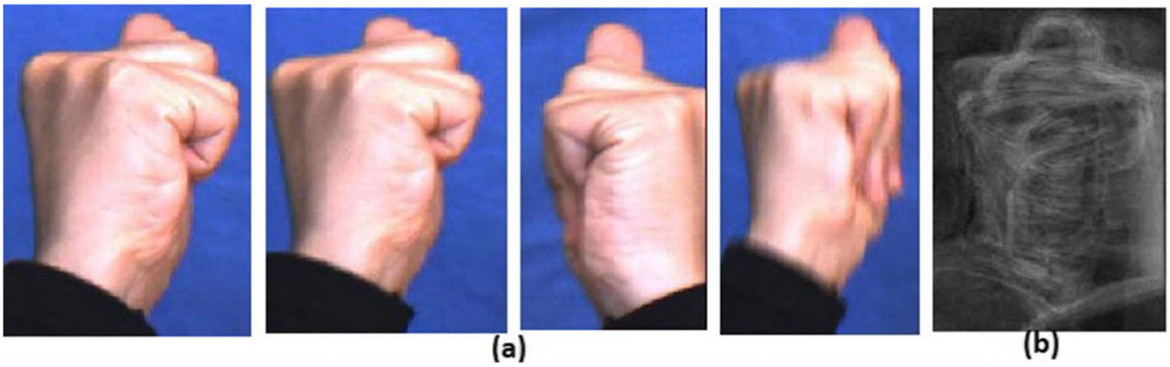


Fig. 4 Hand MHI generated by the hand RoI from the video

obtained from ResNet-101 is 2048, which is half the size of the feature vector obtained from VGG networks. This helps reduce the computation time for GFK and GGDA which are dependent on the size of the feature vector. Thus, we focus on such a network that performs better and gives a low dimensional feature vector.

ResNet-101 network has 101 deep layers (around 28 million parameters) and trained on the ImageNet dataset that contains millions of images and classify the images into 1000 classes. To extract the discriminating features, we first, fine-tune two ResNet-101 on face MHI and hand MHI separately to learn the facial expression features and hand gesture motion features, respectively. Training a ResNet-101 from scratch wiht new dataset is costlier than fine-tuning a pre-trained model. To configure the ResNet-101 network for the new classification problem, we replace the last three layers with a new fully-connected (FC) layer, a soft-max (SM) layer, and a classification output layer for our dataset. Then, we train the one ResNet-101 for 1000 epochs with all the face MHIs and another ResNet-101 for 1000s epoch with all the hand MHIs. We have used stochastic gradient descent with momentum (SGDM) optimizer with a learning rate of .0001 to fine-tune the ResNet-101 network.

We extract the features of both face and hand MHI by using activation on the average pooling layer that is 'pool5'. Finally, we concatenated the features of a face and hand data into a single feature vector to increase the recognition accuracy.

3.5 Creating subspaces of affective state

ResNet-101 on face MHI gives features of size 2048 and ResNet-101 on the hand MHI also gives features of size 2048. We concatenate these features to form one monolithic feature vector of size 4096.

We use these features from each class, to form subspaces of that class for the training data. We apply PCA [29] to all the feature vectors of the same class of affective state to form the subspace of that class for the training data. For the test data, we assume that each feature vector is a subspace by itself. Since all the feature vectors are of the same size, we assume that these subspaces lie on a Grassmann manifold.

A Grassmann manifold $G_r(m, n)$ is a collection of linear subspaces of the same dimension, such that each subspace is of m-dimensional in $\mathbb{R}^n$. Each unique subspace $S^m \subset \mathbb{R}^n$ is a point $P \in G_r(m, n)$, is an orthonormal matrix $O \in \mathbb{R}^n$ of size $m \times n$, such that $S^m = Span(O)$. A Geodesic Flow Kernel (GFK) [17] for domain adaptation is applied on the Grassmann manifold to improve the recognition accuracy.

3.6 Geodesic Flow Kernel (GFK) for domain adaptation

In the real world, it is not always possible that the test data is as perfect as the training data, since the testing images may contain blurred information, varied illumination, and lower image quality that reduces the system's performance. Domain adaptation or transfer learning is a way to overcome the dissimilarity between the training (source domain) and test data (target domain). Generally, a model is trained on the training data or source domain and test data or target domain on which model is being applied for the classification of test data. In our work, we apply the Geodesic Flow Kernel (GFK) [17] to map the target domain to the source domain. The main objective of GFK is to create a new feature space in which the source and target domains have similar characteristics.

Since the GFK captures the incremental changes between the two domains in terms of both the geometric and statistical properties, it is able to robustly discover the similarity between subspaces from both the domains (Fig. 5). In simple terms, GFK creates a feature space by integrating the statistical and geometrical properties between the source and target domain. Assume that $\mathfrak{D}_S$ and $\mathfrak{D}_T$ to be the source (train) and target (test) domains. Therefore, subspaces formed by the labeled training data belong to the source domain, while in the target domain we assume each unlabeled test feature vector to be a subspace by itself and represented as points on a Grassmann manifold G_r.

3.6.1 Constructing the geodesic flow

Let $S_S \in \mathbb{R}^{n\times m}$ denotes the basis of the subspaces of source domain and $S_T \in \mathbb{R}^{n\times m}$ denotes the basis of the subspaces of target domain. Let $P_S \in \mathbb{R}^{n\times(n-m)}$ denote the orthogonal complement to S_S such that $P_S^T S_S = 0$. The geodesic flow is parametrized using canonical Euclidean metric for the Grassmann manifold as $\Theta : t \in [0,1] \longrightarrow \Theta(t) \in G_r(m,n)$ under the constraints $\Theta(0) = S_S$ (source domain subspace) and $\Theta(1) = S_T$ (target domain subspace). For any other $t \in [0,1]$,

$$\Theta(t) = S_S A_1 \Gamma(t) - P_S A_2 \Pi(t) \tag{2}$$

where, $A_1 \in \mathbb{R}^{m\times m}$ and $A_2 \in \mathbb{R}^{(n-m)\times m}$ are orthonormal matrices and can be computed using (3).

$$SVD(S_S^T S_T) = A_1 \Gamma V^T, \quad SVD(P_S^T S_T) = -A_2 \Pi V^T \tag{3}$$

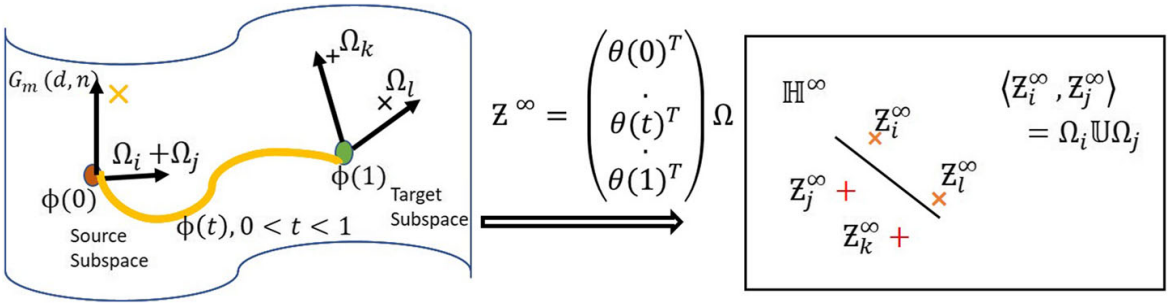


Fig. 5 We create subspaces of the source domain and consider each feature vector in the target domain as a subspace and represent them as points on a Grassmann manifold. We then calculate the geodesic flow between two points from the source and target domains and concatenate the infinite subspaces as $\mathcal{Z}^\infty \in \mathbb{H}^\infty$ along with the flow $\Theta(t)$

where, Γ and Π are diagonal matrices of size $m \times m$ and whose diagonal elements are $cos\theta_i$ and $sin\theta_i$ respectively. Particularly, θ_i are called the principal angles between S_S and S_T:

$$0 \leq \theta_1 \leq \theta_2 \leq \ldots \leq \theta_d \leq \pi/2 \tag{4}$$

They measure the degree that subspaces "overlap". $\Gamma(t)$ and $\Pi(t)$ are also a diagonal matrices whose elements are $cos(t\theta_i)$ and $sin(t\theta_i)$ respectively.

3.6.2 Computing the Geodesic Flow Kernel (GFK)

The geodesic flow allows us to measure how the source domain smoothly changes to the target domain. Assume that $\Theta(t)$ for a $t \in (0, 1)$ is a subspace and find the $\Theta(t)^T \Omega$, that is, the projection of feature vector Ω into the subspace. If Ω is from the source domain and t is close to 1, then projection is more likely to come from the target domain and conversely, for t close to 0.

Thus, using this projection matrix to build a classifier would result in a model using a set of features that have characteristics of both domains. Hence, this classifier would likely perform well on the target domain.

For two original $D-$dimensional feature vectors Ω_i and Ω_j, we find the projection into $\Theta(t)$ for the continuous value of t from 0 to 1 and concatenate all the projections into the feature vectors $\mathcal{Z}_i^\infty$ and $\mathcal{Z}_j^\infty$. The geodesic-flow kernel can be defined by the inner product of projections and subspaces using (5).

$$\langle \mathcal{Z}_i^\infty, \mathcal{Z}_j^\infty \rangle = \int_0^1 (\Theta(t)^T \Omega_i)^T (\Theta(t)^T \Omega_j) dt = \Omega_i^T \mathbb{U} \Omega_j \tag{5}$$

where, $\mathbb{U} \in \mathbb{R}^{n \times n}$ is a positive semi-definite matrix and can be computed in a closed-form using (6).

$$\mathbb{U} = \begin{bmatrix} S_S A_1 & P_S A_2 \end{bmatrix} \begin{bmatrix} F_1 & F_2 \\ F_2 & F_3 \end{bmatrix} \begin{bmatrix} A_1^T S_S^T \\ A_2^T P_S^T \end{bmatrix} \tag{6}$$

where, F_1 to F_3 are diagonal matrices, whose diagonal elements are

$$\alpha_{1i} = 1 + \frac{sin(2\theta_i)}{2\theta_i}, \alpha_{2i} = \frac{cos(2\theta_i) - 1}{2\theta_i}, \alpha_{3i} = 1 - \frac{sin(2\theta_i)}{2\theta_i}. \tag{7}$$

3.7 Adapting Grassmann Graph Embedding Discriminant Analysis (GGDA) using GFK for affective state recognition

Since the GFK captures the incremental changes between the two domains in terms of both the geometric and statistical properties, it is able to robustly discover the similarity between subspaces from both the domains. Therefore, we adapt GGDA by applying the GFK instead of the projection kernel for robust and accurate affective state recognition from multiple modalities. We apply the geodesic flow kernel $\mathbb{U}$ obtained from (6) to map the Grassmann manifold, $G_r(m, n)$, to another Grassmann manifold, $G'_r(m, n)$, to improve the recognition accuracy by maintaining the local properties of same classes and improves the discriminatory power between the different classes.

Let graph (B, E) denotes an undirected weighted graph with vertice's B that are the points on the $G_r(m, n)$ and edges E that connects pair of vertices. Then, E is a symmetric matrix whose value can be obtained by the Geodesic Flow Kernel Equation (6). The diagonal matrix D of the graph is describe as $D(i, i) = \sum_{j \neq i} E(i, j)$ and Laplacian matrix L of the graph is describe as $L = D - E$.

Let $\mathbb{S} = \{(S_i, l_i)\}_{i=1}^{N}$ be a set of subspaces of source domain and represented as a point on Grassmann manifold $G_r(m, n)$, where, $S_i, \forall i = 1, 2, \ldots, N$ is a subspace belonging to $\mathbb{R}^{m \times n}$, $l_i \in \{1, 2, \ldots, C\}$ are the class labels and C is a number of classes and N is the labeled points on the manifold. Then, based on nearest neighbors graph, graphs E_w denotes the with-in similarity graph and E_b denotes the between similarity graph can be created using (8) and (9). Figure 6, explains the mapping of subspaces from one manifold to other manifold with the help of two similarity graphs E_w and E_b that increase the inter-class dissimilarity and intra-class similarity.

$$E_w(i, j) = \begin{cases} 1, & \text{if } S_i \in N_w(S_j) \text{ or } S_j \in N_w(S_i) \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

$$E_b(i, j) = \begin{cases} 1, & \text{if } S_i \in N_b(S_j) \text{ or } S_j \in N_b(S_i) \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

In (8), $N_w(S_i)$ denotes set of neighbors $\{S_i^1, S_i^2, \ldots, S_i^v\}$, having the same label as l_i. Similarly in (9), $N_b(S_i)$ denotes the set of v neighbors having different labels. Mapping from one manifold to another manifold maximizes the discriminatory power to recognize the affective state. In GGDA, a geodesic flow kernel (GFK) in place of projection kernel is applied to map a Grassmann manifold $G_r(m, n)$ to another Grassmann manifold G_r', that is, $\alpha : S_i \rightarrow K_i$ such that the points in with-in class similarity graph E_w are mapped close together while a between-class similarity graph E_b are mapped far apart. Optimizing the two objective functions, η_1, η_2 given in (10) and (11) gives the mapping, $\alpha : S_i \rightarrow Q_i$.

$$\eta_1 = \min \frac{1}{2} \sum_{i,j} (K_i - K_j)^2 E_w(i, j) \tag{10}$$

$$\eta_2 = \max \frac{1}{2} \sum_{i,j} (K_i - K_j)^2 E_b(i, j) \tag{11}$$

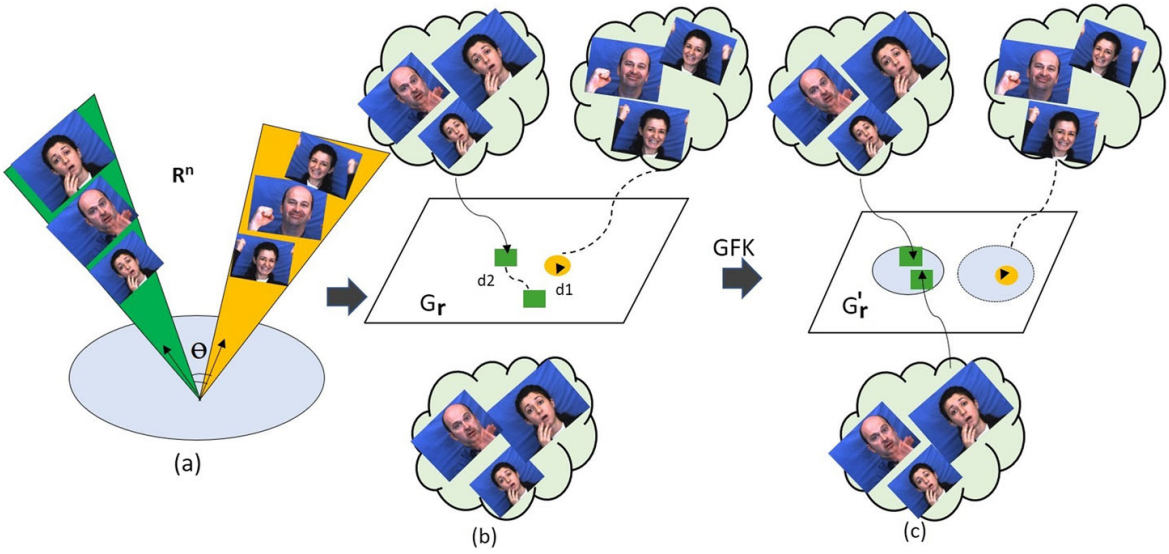


Fig. 6 **a** Features set of "Happy" and "Fear" affective state. **b** Linear subspaces of gestures in $\mathbb{R}^n$ can be represented as points on a Grassmann manifold G_r. d_1 and d_2 are geodesic distance between points on G_r. In **c** using a geodesic flow kernel, points on manifold G_r can be mapped into other manifold G_r', where, the geometrical structure and local information have been preserved by using within-class and between-class similarity graphs, thereby increasing the discriminatory power between classes

The objective function η_1 imposes a penalty if points K_i and K_j are mapped far away in $G_r{}'$ while they belongs to the same class. Similarly, η_2 imposes a penalties if points K_i and K_j are mapped closely together in G_r' while they belong to different class.

Confining the solution to be linear, convert η_1 and η_2 into maximization problem is given by (12).

$$\max\{\mathbb{A}^T\mathbb{U}\{L_b + \gamma E_w\}\mathbb{U}^{\mathbb{T}}\mathbb{A}\} \quad \text{subject to} \quad \mathbb{A}^T\mathbb{U}D_w\mathbb{U}^T\mathbb{A} = 1 \tag{12}$$

where γ is a Lagrangian multiplier that acts as a regularization parameter in the final solution. Solving the generalized Eigenvalue problem, (13) gives the solution of (12).

$$\mathbb{U}\{L_b + \gamma E_w\}\mathbb{U}^{\mathbb{T}}\mathbb{A} = \lambda\mathbb{U}D_w\mathbb{U}^T\mathbb{A} \tag{13}$$

where, the projection matrix $\mathbb{A}$ is formed using r, largest eigenvectors of Rayleigh quotient given by (14).

$$\frac{\mathbb{U}D_w\mathbb{U}^T}{\mathbb{U}\{L_b + \gamma E_w\}\mathbb{U}^T} \tag{14}$$

The mapping $\mathbb{A}$ preserves the overall geometry while maximizing a quotient similar to discriminant analysis. Test data are assigned to the class with which it has minimum similarity score, using k-NN method.

Pseudo-code of our proposed approach:

1. Modalities Detection and Feature Extraction
   - Detect hand using Neural Networks (SSD) on Tensorflow [15] and detect face using Viola-Jones face detector [46], save the ROI for further processing.
   - Find the Motion History Image of each ROI using (1).
   - Extract the features of each modality using fine-tunned Resnet-101 network and finally concatenate the features of both modality into single feature vector.
2. Domain Adaptation: Geodesic Flow Kernel
   - Input data is source and target domain features obtained from the deep network with corresponding class labels in the source domain and target domain does not have any class labels.
   - GFK creates a features space called GFK kernel $\mathbb{U}$ in which both source and target domain shares similar characteristics and can be used in any kernel-based classifier to classify the test data.
   - Compute the GFK flow kernel matrix $\mathbb{U}$ for train and test data using (6) and store them in $T_{Rkernel}$. and $T_{Skernel}$.
3. Grassmann Graph Embedding Discriminant Analysis:
   Training Phase:
   - Using (8) and (9), compute the two graph E_w and E_b respectively. We also compute the diagonal within matrix D_w as $D_w(i,i) = \sum_{j \neq i} E_w(i,j)$ and between Laplacian matrix $L_w = D_w - E_w$. Similarly, matrices D_b and L_b are also computed.
   - We solved the corresponding maximization problem in (13), and obtain the Rayleigh quotient α using (14) and find the matrix $FT_{Rproj} = \alpha \times T_{Rkernel}$.

   Testing Phase:
   - We again compute the corresponding Rayleigh quotient α and find the matrix $FT_{Sproj} = \alpha \times T_{Skernel}$, for testing data.

– Then, the k-nearest neighbor is used to calculate the similarity score between each test subspaces and each training subspaces. Based on the similarity score test data is assigned to the training class with which it has a minimum score.

4 Experimental analysis

4.1 Experimental setup

Experiments are computed on an Intel Core i7-7500U CPU @ 2.70 GHz with 8GB RAM and GeForce 940MX GPU. We perform experiments on publicly available bi-modal face and body dataset FABO [19] and results on the FABO dataset show that our proposed method performs well and is comparable to the state-of-the-art methods. FABO dataset contains RGB videos of face and body gesture recorded by two cameras respectively. Each video contains different expressions for some situations, for example, in the situation "the lecture is the most boring one", the person performs "boredom" expression with the help of facial expressions and hand gestures. Each face video and body gesture videos are labeled independently by 4 to 6 observers. For this work, we choose the videos from one camera that contains face and body gestures together. Figure 7a and b shows two sample videos of "anger" and "boredom". There are a total of 391 videos including basic and non-basic expressions as reported in Table 1. We perform a cross-subject validation to evaluate our proposed model. Each expression is performed by 17 subject from which we choose 2/3rd subjects for training and 1/3rd for testing. Thus, videos of the same subjects are not used for both training and testing. 3-fold cross-subject validation is performed and average accuracy is reported in Fig. 9. Table 1 shows the affective classes and the corresponding number of videos for training and testing.

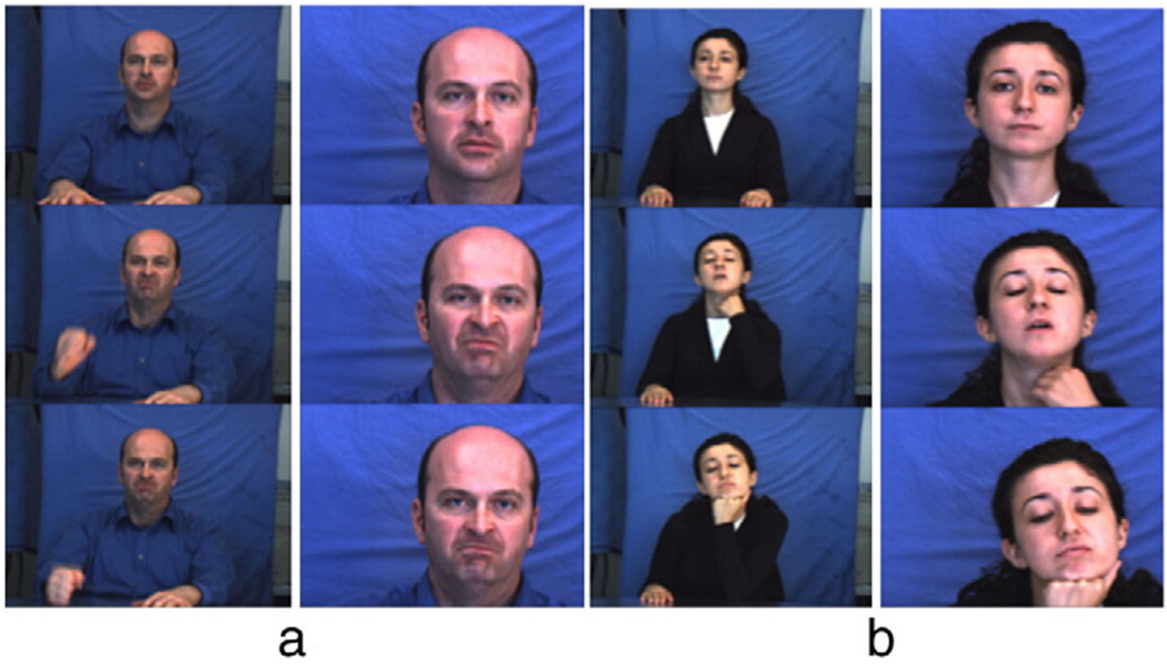


Fig. 7 Sample images of different expressions from FABO dataset [19]. **a** shows few frames of "Anger" expression recorded by body modality (left) and face modality (right) **b** shows few frames of "Boredom" expression by body modality (left) and face modality (right)

Table 1 Number of videos available for each affective state in the FABO dataset [19]

Affective state	Videos	Affective state	Videos
Happiness	25	Uncertainty	25
Surprise	13	Puzzlement	54
Boredom	38	Anxiety	27
Disgust	28	Neutral	70
Fear	25	Sadness	16
Anger	70		

4.2 Affective state recognition using our framework

In our proposed work, we perform affective state recognition through facial expression and hand gesture recognition. Therefore, as the first step, face and hands are detected in each frame in each video. We detect the face using Viola-Jones face detector [46] and select the face as the face region of interest (Face ROI) in each frame of each video. Then, for each video, to incorporate the motion dynamics of a facial expression across frames, we calculate the motion history image of the face ROIs across frames and use the Face MHI to extract the features using ResNet-101 network.

Similarly, we detect the gesturing hands in each frame of each video, using neural networks SSD on TensorFlow. We create the motion history images of the detected hands, in each frame to create the Hand MHI of each video and extract the features using another ResNet-101 network. These features are concatenated, GFK is computed and GGDA is applied for multi-modal affect recognition.

We extract the features of different modalities using ResNet-101 network and perform classification in different scenarios. We consider four cases in which classification is done using (a) ResNet-101 Network; (b) feature extraction using ResNet-101 and then domain adaptation using GFK for unlabeled data and recognition using SVM [17]; (c) feature extraction using ResNet-101 and recognition using GGDA (d) feature extraction using ResNet-101 and then domain adaptation using GFK for unlabeled data and recognition using GGDA. In the first case, (a), we pass the MHI images of both hands and face, extract the features and then fuse the features at the fully connected layer and pass the fused features to softmax for the classification. In this case, we achieved 89.8% accuracy as shown in Fig. 8. In the second case, (b), features are extracted using ResNet-101, then, GFK domain adaptation is carried out to reduce the mis-match in the test and training data. Then, after extracting the geodesic flow kernel, that kernel matrix is used to train the SVM for classification. In this case also, we get an average accuracy of 91.2% accuracy shown in Fig. 8. In the third case, (c), we used subspace learning framework with Grassmann manifold to classify the affective states. In this features are extracted from ResNet-101 and fused, then passed to the Grassmann graph embedding framework. In this case, we achieve an average accuracy of 90.8% shown in Fig. 8. As we see see that in case (b) ResNet-101+GFK performs better than case (c) ResNet-101+GGDA. Therefore, we proposed a novel model wherein GGDA is extended to incorporate the kernel matrix obtained from GFK. In this case also, first the features are extracted using ResNet-101. Then, these features are fused and subspaces ae created that lie on a Grassmann manifold. We then use GFK for domain adaptation and obtain the kernel matrix. We use this kernel matrix in the GGDA framework for classification. In the Fig. 8 we can see that our proposed, novel model (case (d))

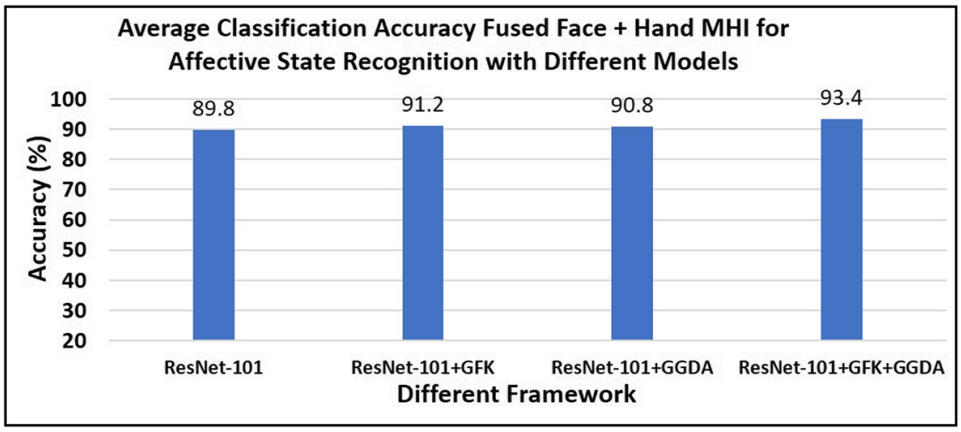


Fig. 8 Figure shows the average accuracy of the face and hands MHI together for intent recognition with different frameworks. We achieve the best accuracy of 93.4% for our proposed framework

of ResNet-101+GFK+GGDA performs best and achieves an average accuracy of 93.4%. Therefore, we carry out all other experiments using our proposed novel model (case (d)).

Further, we compare and show in Fig. 9 the accuracy of affective state recognition using a combination of facial expression and hand gestures, only facial expressions and only hand gestures in different test cases. The average accuracy we achieved in the case of combined modalities is 93.4% which is more than the average accuracy of single modalities. We achieve an average accuracy of 92.4% for affect recognition using only facial expression analysis and 76.8% when using only hand gesture recognition. We get the highest accuracy in case of combined modalities because hand and face modalities together give richer features to recognize the affective state rather than a single modalities. In both these cases, we have applied our framework, that is, extracted features from the respective MHI using ResNet-101 and then computed GFK and applied GGDA with the GFK. These experiments also show that although face expression analysis is a good indicator of a person's affective state, both facial expressions and hand gestures together give a more accurate recognition of the person's affective state. Figure 10 shows the confusion matrix of combined face

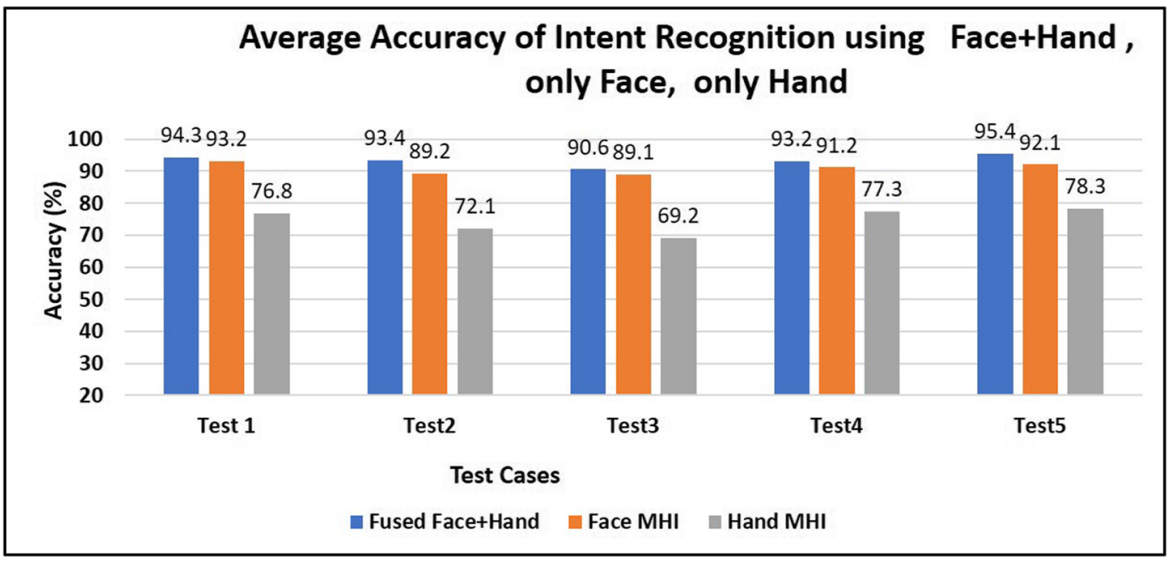


Fig. 9 This figure shows that the average accuracy of fused face and hand modality performing better than each individual modality. We achieved average accuracy of 93.4% in case of fused modalities

CRR	Happiness	Surprise	Boredom	Disgust	Fear	Anger	Uncertainty	Puzzlement	Anxiety	Neutral	Sadness	% Accuracy
Happiness	24	1	0	0	0	0	0	0	0	0	0	96
Surprise	1	12	0	0	0	0	0	0	0	0	0	92.3
Boredom	0	0	36	0	0	1	0	1	0	0	0	94.7
Disgust	0	0	1	26	1	0	0	0	0	0	0	92.8
Fear	0	0	0	0	24	0	0	0	1	0	0	96
Anger	0	0	4	0	0	63	0	0	2	0	1	90
Uncertainty	0	0	1	0	0	0	24	0	0	0	0	96
Puzzlement	0	0	3	0	0	2	0	49	0	0	0	90.7
Anxiety	0	1	0	0	0	0	0	2	24	0	0	88.8
Neutral	0	0	0	0	0	0	0	2	0	63	6	88.7
Sadness	0	0	0	0	0	1	0	0	0	0	15	93.7

Fig. 10 Sample confusion matrix of combined face and body gesture modalities. The average accuracy we achieved is 93.4%

and hand gesture modalities using ResNet-101 features using GFK domain adaptation and GGDA based recognition. In the confusion matrix, the last column shows the true positive rate of each affective state. From the confusion matrix, we can see that "puzzlement" is highly confused with "boredom" and "neutral" is highly confused with sadness.Similarly, "anger" is confused with "boredom" because expressions of boredom and anger are quite similar and hands also do not play a major role in it. In Fig. 7 we can see that face expression of boredom and anger are similar, yields to the similar MHI. In anxiety and neutral state, we get lower accuracy because of similarity in the expressions.

4.3 Experiments on our own dataset

Due to the limited unavailability of bi-model face and body gesture datasets, to generalize our proposed approach we captured our own dataset. We perform a set of expressions such as happiness, disgust, sadness, surprise, anger in front of the camera and record the videos for each. The recording environment is set up in the lab and recorded by the mobile camera. Each expression is performed five times each by three subjects. Thus, the complete dataset contains 15 videos for each expression. Figure 11 shows the sample images of happy and surprise video expression. When we perform cross-data testing, that is, training is on the



Fig. 11 Sample images of different expressions. **a** Sample images of "Happy" expression video of own dataset **b** Sample images of "Surprise" expression video of own dataset

FABO dataset [19] and testing on our own dataset, then we achieve 65.4% average accuracy. In some cases, we get more than 75% accuracy. We also trained the model on our own dataset by including two subject's data in training and testing on one remaining subject. We achieved an average accuracy of 92.7%. Thus, this shows that our proposed method can give good recognition accuracy if the model is trained on the same dataset and moderate accuracy in the case of cross-data testing.

5 Discussion on the experiments

In this section, we compare our proposed framework with different modalities. We show the accuracy of our proposed model on face and hand modalities separately in Figs. 12 and 13. Then, in Section 5.2, we have shown comparison of proposed model with state-of-the-art-methods. In Section 5.3, we discuss the time analysis.

5.1 Comparison of different frameworks with different modalities

To show that face detection and extracted ROI image helps to increase the recognition accuracy, we perform experiments by considering (a) the complete image and (b) only the Face RoI to create the MHIs. We find the motion history image of the full image (without detecting face) and compare the accuracy of the framework in a case where we use only the Face MHI as shown in Fig. 12. From the experimental results, we observe that in our proposed framework, when the Face ROI is used, we get an accuracy of 92.4% while when the complete image is used we get an accuracy of 87.3% . As face detection takes 0.04 s for each frame,it can be easily incorporated in the system to achieve high recognition accuracy.

We also conduct experiments where we perform feature extraction using ResNet-101 network and affective state recognition using (a) ResNet-101 Network; (b) feature extraction using ResNet-101 and then domain adaptation using GFK for unlabeled data and recognition using SVM [17]; (c) feature extraction using ResNet-101 and recognition using GGDA

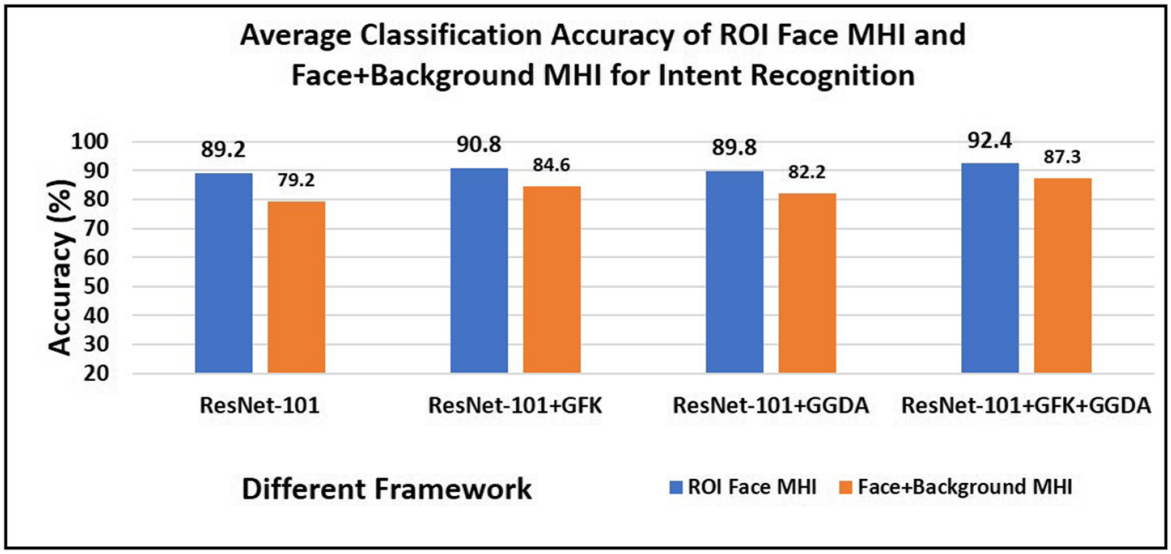


Fig. 12 This shows the average accuracy of ROI face MHI and complete image MHI data separately for affective state recognition with a different framework. We achieve the best accuracy of 92.4% in case of ROI face MHI using ResNet-101 + GFK + GGDA

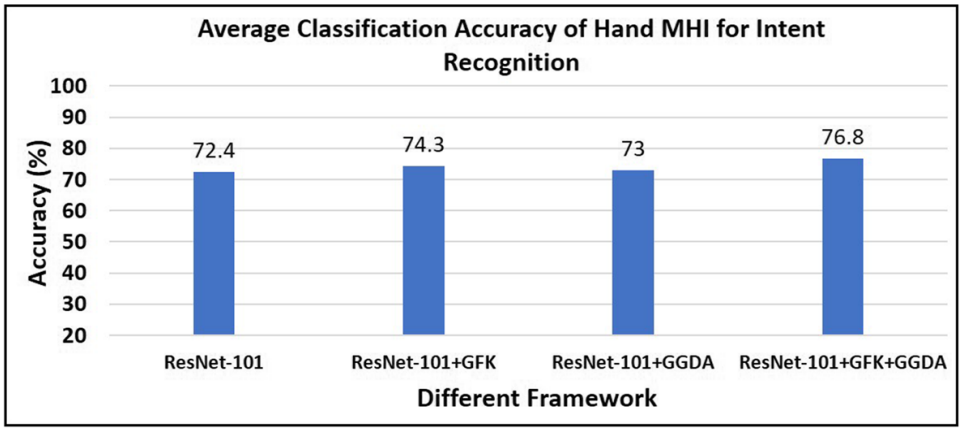


Fig. 13 This shows the average accuracy of hand MHI using different frameworks. It can be seen that the best accuracy acheived is 76.8% when using our proposed framework

(d) feature extraction using ResNet-101 and then domain adaptation using GFK for unlabeled data and recognition using GGDA. In ResNet-101, we pass the MHI images and recognize the input data. Figure 12 also summarizes our experiments on the FABO dataset for affective recognition using an only facial expression with different frameworks. As can be seen, by Fig. 12, our proposed framework ResNet-101+GFK+GGDA gives the best results for affective state recognition even when only the face modality is taken into consideration. Since we achieve better accuracy in the case of face detection, we use the detected face ROI to perform all our experiments. Face detection also reduces the computation time and increases recognition accuracy.

Similarly, we detect the hand in each frame using Neural Networks (SSD) on Tensorflow [15], create the hand MHI and perform experiments using the different frameworks mentioned above. We perform experiments using hand MHI only to recognize the affective state using different frameworks and average accuracy achieved by various frameworks is reported in Fig. 13. In the case of hand MHI, we get 76.8% accuracy using our proposed framework, which is lower compared to using only Face MHI because of fewer data and also very less variation in performing the gesture. In this also we have shown the recognition accuracy of various framework and found that best accuracy is achieved in the case of ResNet-101 + GFK + GGDA of 76.8%, and we have used this proposed framework for all other experiments.

5.2 Comparison with state-of-the-art methods

In the literature, work on affective state recognition based on multiple modalities has focused on facial expressions and body gestures as the modalities. In most cases, where body gestures are taken into consideration, the hands are dynamic and performing the gestures, while the rest of the body is still. Therefore, in our approach, although we focus on hand gesture and facial expression as the two modalities for affective state recognition, we compare our work with those that have taken into consideration face expression and body gestures and performed experiments on the FABO [19] dataset.

Considering only the hand instead of the complete body helps make our system efficient in both space and time also. Chen [11] used MHI-HOG and Image-HOG to extract the features, representing the features using temporal normalization as well as a bag of words (BOW). They achieved 68% and 66% accuracy respectively with the SVM classifier. Gunes [20] used feature extraction techniques such as optical flow, geometric features

with hand tracking, temporal features, and shoulder tracking and perform emotion recognition with random forest classifier (RF) and achieved an accuracy of 78%. Barros [3], uses a multi-channel convolutional neural network with a cubic receptive field for affective state recognition based on facial expressions only, body gestures only and combining both face and body gestures and achieved an accuracy of 72.7%, 57.8% and 91.3% respectively. We have used deep learning for feature extraction and Grassmannian learning framework with Geodesic Flow Kernel and achieved an accuracy of 93.4% when considering both face and hand gestures for affective state recognition. The comparisons are shown in Fig. 14. Feature extraction using a convolutional neural network such as ResNet-101 gives high discriminating features and also solves the problems associated with working with hand-crafted features and their extraction. Representing each feature vector in the form of subspaces on the Grassmann manifold captures all the variations and uncertainty occurs in the facial expressions and hand gesture performed either by the same person or different people. We find that adapting GFK kernel with GGDA for affective state recognition helps in increasing the recognition accuracy. Our experiments have shown that even in case when the training data does not contain any examples from the test dataset, the recognition accuracy is better than some of the methods reported in the literature.

We also compare our framework with the above-mentioned work from the literature, where only the face modality has been used for affective state recognition (using FABO [19] dataset). Figure 15 shows that our proposed framework has 92.4% average recognition accuracy which is higher compared to the Chen [11], Gunes [20] and Barros [3]. The Hidden Markov model (HMM) based method proposed by Gunes [20] is video-based achieved very little accuracy compare to our method.

5.3 Time analysis

Our proposed work is divided into three modules (i) face, hand detection, and ROI extraction; (ii) feature extraction and concatenation of the two feature vectors to obtain a single feature; (iii) domain adaptation using geodesic flow kernel and GGDA based affect

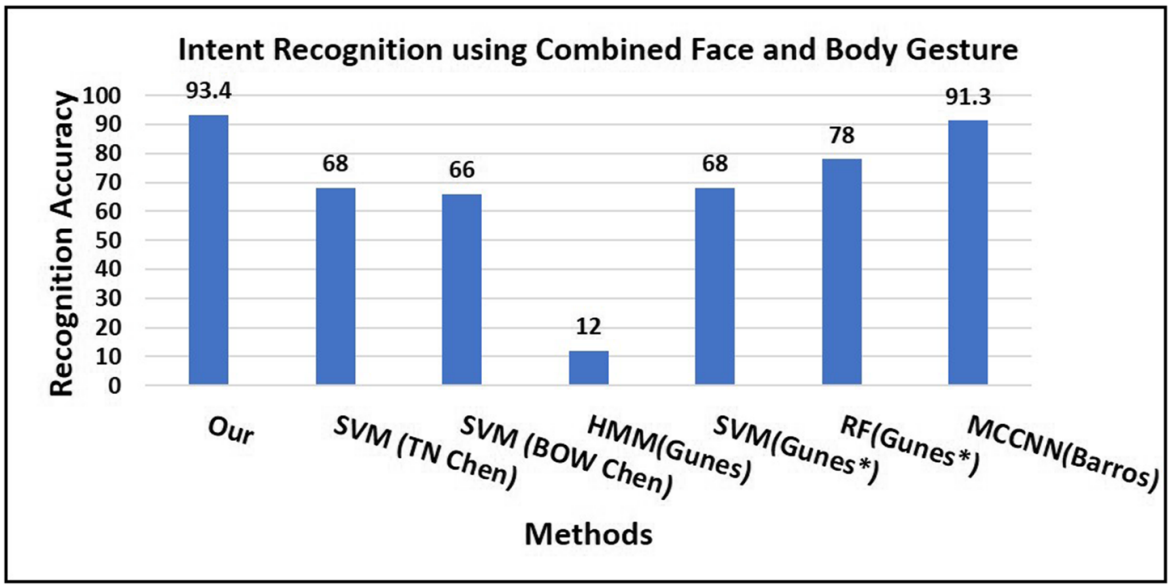


Fig. 14 Comparison of our proposed approach with state-of-the-art-methods on face and body gesture modality. Accuracy of Gunes* [20] is measured per frame

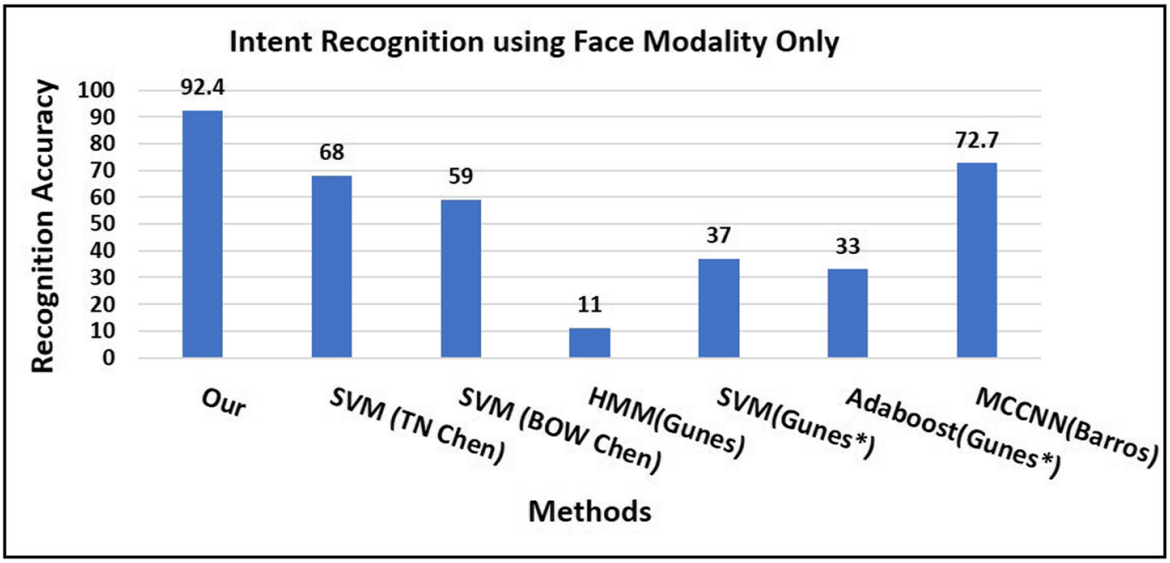


Fig. 15 Comparison of our proposed approach with state-of-the-art methods on face modality only. The accuracy of Gunes* [20] is measured per frame

recognition. Each video contains an average 20 frames of resolution 1024 × 768 pixels. For face detection, we have used a Viola-Jones face detector that takes 25 frames/seconds and for one video it takes 0.8 s. For hand detection, we have used a hand detector using Neural Network (SSD) on TensorFlow that takes 30 frames/seconds and for one video it takes 0.67 second. Both face and hand detection can be performed in parallel. Motion history calculation takes 25 frames/seconds for creating both face MHI and hand MHI and is also done in parallel and takes 0.8 s for one video. MHI calculation time can be reduced to half if we skip every 2nd frame. Therefore, in our proposed work MHI calculation takes 0.4 s. Deep learning-based feature extraction for each MHI takes 0.062 s. GGDA with Geodesic Flow Kernel takes 0.36 s for one video. Thus the total time for one test data modality detection, feature extraction and recognition takes 1.622 s.

6 Conclusion

In this paper, we have proposed a novel framework for vision-based affective state recognition using facial expression and hand gestures as the multiple modalities. We find that performing face and hand detection, helps to focus on the modalities of interest and removes the noise that the background may contain. Therefore, as the first step, we perform face and hand detection and use the face and hand ROIs to create the motion history images of each.

We perform experiments on the publicly available FABO dataset [19] to validate our proposed framework. We also perform several experiments to validate that multi-modal affective state recognition is an improvement in affect recognition using a single modality such as facial expressions. However, we also show that even in the case when the only facial expression is used for affective state recognition, our framework works very well. We also carry out experiments that show that our proposed framework of using features from a deep network and adapting GGDA by using GFK for recognition performs much better when compared with performing recognition using only deep learning, only GFK with features extracted from deep networks, GGDA with projection kernel and features from a deep network. Moreover, detecting the face and hand and using these instead of the

complete framework for creating the motion history images reduces the computational costs in terms of both space and time. Our experimental results show that our proposed framework for affective state recognition using both facial expression and hand gestures performs with a high degree of accuracy and is comparable to the state-of-the-art methods.

References

1. Ambady N, Rosenthal R (1992) Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. Psychol Bull 111:256
2. Ambady N, Bernieri FJ, Richeson JA (2000) Toward a histology of social behavior: judgmental accuracy from thin slices of the behavioral stream. Adv Exp Social Psychol 32:201–271
3. Barros P, Jirak D, Weber C, Wermter S (2015) Multimodal emotional state recognition using sequence-dependent deep hierarchical features. Neural Netw 72:140–151
4. Barros P, Parisi GI, Weber C, Wermter S (2017) Emotion-modulated attention improves expression recognition: A deep learning model. Neurocomputing 253:104–114
5. Bastanfard A, Karam H, Takahashi H, Nakajima M (2002) Simulation of human facial aging and skin senility. In: ITE technical report, the institute of image information and television engineers, vol 26(3), pp 1–7
6. Bastanfard A, Bastanfard O, Takahashi H, Nakajima M (2004) Toward anthropometrics simulation of face rejuvenation and skin cosmetic. In: Computer animation and virtual worlds 15, vol 3, pp 347–352
7. Bastanfard A, Takahashi H, Nakajima M (2004) Toward E-appearance of human face and hair by age, expression and rejuvenation. In: International conference on cyberworlds. IEEE, pp 306–311
8. Berthouze N, Valstar M, Williams A, Egede J, Olugbade T, Wang C, Meng H, Aung M, Lane N, Song S Emopain challenge 2020: Multimodal pain evaluation from facial and bodily expressions. arXiv:2001.07739
9. Cai H, Qu Z, Li Z, Zhang Y, Hu X, Hu B (2020) Feature-level fusion approaches based on multimodal eeg data for depression recognition. Inf Fusion 59:127–138
10. Canziani A, Paszke A, Culurciello E An analysis of deep neural network models for practical applications. arXiv:1605.07678
11. Chen S, Tian Y, Liu Q, Metaxas DN (2013) Recognizing expressions from face and body gesture by temporal normalized motion and appearance features. Image Vis Comput 31(2):175–185
12. Chen L, Wu M, Wanjuan SU, Hirota K (2018) Multi-convolution neural networks-based on deep learning model for emotion understanding. In: Proceedings of 37th international conference on chinese control conference (CCC), pp 9545–9549
13. Cohen I, Sebe N, Garg A, Chen LS, Huang TS (2003) Facial expression recognition from video sequences: Temporal and static modeling. Comput Vis Image Underst 91(1–2):160–187
14. Dellaert F, Polzin T, Waibel A (1996) Recognizing emotion in speech. In: Proceedings of 4th international conference on spoken language processing
15. Dibia V (2017) Real-time hand tracking using ssd on tensorflow, https://github.com/victordibia/handtracking
16. Gadanho SC (2003) Learning behavior-selection by emotions and cognition in a multi-goal robot task. J Mach Learn Res 4:385–412
17. Gong B, Shi Y, Sha F, Grauman K (2012) Geodesic flow kernel for unsupervised domain adaptation. In: Proceedings of international conference on computer vision and pattern recognition (CVPR), pp 2066–2073
18. Gu Y, Mai X, Luo Y-J (2013) Do bodily expressions compete with facial expressions? Time course of integration of emotional signals from the face and the body. PLoS One 8:736–762
19. Gunes H, Piccardi M (2006) A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In: Proceedings of 18th international conference on pattern recognition, vol 1, pp 1148–1153
20. Gunes H, Piccardi M (2009) Automatic temporal segment detection and affect recognition from face and body display. IEEE Trans Syst Man Cybern 39(1):64–84
21. Harandi MT, Sanderson C, Shirazi S, Lovell BC (2011) Graph embedding discriminant analysis on Grassmannian manifold for improved image set matching. In: Proceedings of international conference on computer vision and pattern recognition (CVPR), pp 2705–2712
22. Kapoor RW, Picard A (2005) Multimodal affect recognition in learning environments. In: Proceedings of 13th annual ACM international conference on multimedia, pp 677–682

23. Kim J, André E (2008) Emotion recognition based on physiological changes in music listening. IEEE Trans Pattern Anal Mach Intell 30(12):2067–2083
24. Kim KH, Bang SW, Kim SR (2004) Emotion recognition system using short-term monitoring of physiological signals. Med Biol Eng Comput 42(3):419–427
25. Koelstra S, Muhl C, Soleymani M, Lee JS, Yazdani A, Ebrahimi T, Patras I (2012) Deap: A database for emotion analysis; using physiological signals. IEEE Trans Affect Comput 3(1):18–31
26. Kret ME, Roelofs K, Stekelenburg J, De GB (2013) Salient cues from faces, bodies and scenes influence observers face expressions. Front Hum Neurosci 7:810–850
27. Lisetti CL, Nasoz F (2004) Using noninvasive wearable computers to recognize human emotions from physiological signals. EURASIP J Appl Signal Process 11(929414):1672–1687
28. Liu P, Yu H, Cang S (2019) Adaptive neural network tracking control for underactuated systems with matched and mismatched disturbances. Nonlinear Dyn 98(2):1447–1464
29. Mohammadi E, Fatemizadeh MR, Mahoor MH (2014) Pca-based dictionary building for accurate facial expression recognition via sparse representation. J Vis Commun Image Represent 25(5):1082–1092
30. Nguyen D, Nguyen K, Sridharan S, Dean D, Fookes C (2018) Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition. Comput Vis Image Underst 174:33–42
31. Pentland A (2008) Honest signals: How they shape our world, MIT Press
32. Picard R (1997) Affective computing. The MIT Press, Cambridge, Massachusetts
33. Psaltis A, Kaza K, Stefanidis K, Thermos S, Apostolakis KC, Dimitropoulos K, Daras P (2016) Multimodal affective state recognition in serious games applications. In: Proceedings of international conference on imaging systems and techniques (IST), pp 435–439
34. Richmond VP, McCroskey JC, Hickson M (2008) Nonverbal behavior in interpersonal relations. Allyn & Bacon
35. Roy D, Pentland A (1996) Automatic spoken affect classification and analysis. In: Proceedings of 2nd international conference on automatic face and gesture recognition, pp 363–367
36. Scherer KR, Ellgring H (2007) Multimodal expression of emotion: Affect programs or componential appraisal patterns? Emotion 7(1):158
37. Shan C, Gong S, McOwan PW (2007) Beyond facial expressions: learning human emotion from body gestures. In: Proceedings of international conference on british machine vision conference (BMVC), pp 1–10
38. Spexard T, Hanheide M, Sagerer G (2007) Human-oriented interaction with an anthropomorphic robot. IEEE Trans Robot 23:852–862
39. Sun L, Zhao C, Yan Z, Liu P, Duckett T, Stolkin R (2018) A novel weakly-supervised approach for RGB-D-based nuclear waste object detection. IEEE Sens J 19(9):3487–3500
40. Sun B, Cao S, He J, Yu L (2018) Affect recognition from facial movements and body gestures by hierarchical deep spatio-temporal features and fusion strategy. Neural Netw 105:36–51
41. Tang Z, Yu H, Lu C, Liu P, Jin X (2019) Single-trial classification of different movements on one arm based on ERD/ERS and corticomuscular coherence. IEEE Access 7:128185–128197
42. Tang ZC, Li C, Wu JF, Liu PC, Cheng SW (2019) Classification of EEG-based single-trial motor imagery tasks using a B-CSP method for BCI. Front Inf Technol Electron Eng 20(8):1087–1098
43. Tzirakis P, Trigeorgis G, Nicolaou MA, Schuller BW, Zafeiriou S (2017) End-to-end multimodal emotion recognition using Deep Neural Networks. IEEE J Sel Top Signal Process 11(8):1301–1309
44. Uddin MA, Joolee JB, Lee Y-K Depression level prediction using deep spatiotemporal features and 698 multilayer bi-ltsm. IEEE Transactions on Affective Computing (2020)
45. Vinciarelli A, Pantic M, Bourlard H (2009) Social signal proessing: Survey of an emerging domain. Image Vis Comput 27:1743–1759
46. Viola P, Jones MJ (2004) Robust real-time face detection. Int J Comput Vis 57(2):137–154
47. Yang Z, Narayanan S (2015) Modeling mutual influence of multimodal behavior in affective dyadic interactions. In: Proceedings of international conference on acoustics, speech and signal processing (ICASSP), pp 2234–2238
48. Yang Z, Gong B, Narayanan S (2017) Weighted geodesic flow kernel for interpersonal mutual influence modeling and emotion recognition in dyadic interactions. In: Proceedings of international conference on affective computing and intelligent interaction (ACII), pp 236–241

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

Proceedings of the Third International Conference on Intelligent Sustainable Systems [ICISS 2020]
IEEE Xplore Part Number: CFP20M19-ART; ISBN: 978-1-7281-7089-3

Alignment-Free Cancelable Biometric: A Contemporary Survey, Opportunities & Challenges

Pratima Sharma
Delhi Technological University
Delhi, India
pratimasharma1290@gmail.com

Gurjit Singh Walia
DRDO
Delhi, India
gurjit.walia@gmail.com

Rajesh Rohilla
Delhi Technological University
Delhi, India
rajesh@dce.ac.in

2020 3rd International Conference on Intelligent Sustainable Systems (ICISS) | 978-1-7281-7089-3/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICISS49785.2020.9316054

***Abstract*—This survey aims to review different alignment free recognition approaches inclusive of advantages and limitations of various state-of-the-art biometric template protection schemes. Occurrence of rotational and translational variation while acquiring images followed by one way transformation or cryptographic hashing, results in degradation of biometric recognition performance. The non-invertible transformation and cryptographic methods addresses the security and privacy threats associated with biometric templates. Although, maintaining high recognition accuracy became the foremost challenge for various proposed approaches of biometric template protection. Among largely suggested approaches for performance enhancement, need for registration or alignment of biometric features came out as a difficult requirement to be fulfilled. Hence, performance degradation mainly due to elastic distortion can be confronted using various alignment-free approaches. The alignment-free method involves use of a feature extraction process that extracts invariant attributes found in local geometrical structures of biometric templates. This survey article discusses various techniques of forming local structure accompanying non-invertible methods to generate alignment free cancelable templates or bio-cryptosystems. Further, a comparative analysis of various alignment-free cancelable approaches are compiled in tabular manner. Review also includes a summarization of the research gaps to aid in finding future directions.**

***Keywords*—Cancelable, Bio-Cryptosystem, Alignment-free, Biometrics**

I. INTRODUCTION

In the last two decades, traditional knowledge based authentication has been replaced by an automated biometric based authentication system [2]. The foremost reason is the need for reliable user authentication. Loss of password/token can lead to catastrophic consequences in security critical application. In response, for accurate and secure authentication, biometric technology was recognized as the most accepted solution and has been widely deployed worldwide. The intrinsic characteristic associated with biometrics such as uniqueness, permanence and usability, distinguished this technology from the traditional approach. Biometric traits such as iris, finger, face, signature, gait etc are among few to be adopted for authentication of an individual. Among these traits, iris and fingerprint are considered to be the most powerful biometric characteristics with no probability of finding the same pattern in any other individual i.e.. uniqueness [2]. But the adversarial models of surreptitious duplication of biometric templates can help imposters to successfully spoof the genuine users. This can lead to lifetime compromise of biometric identity and hence is a threat to the security and privacy of an individual. Thorough analysis suggests that use of same biometric characteristics for different services of different organisations can lead to exchange of information among organisations. This information can lead to cross matching of an individual and hence impedes privacy. To overcome these kinds of security threats, protection of biometric templates was suggested in [7]. In contrast to adoption of cryptography methods in protecting biometric templates, a cancelable scheme of template protection was widely accepted and followed. The alignment-free biometric template protection scheme has been classified in fig. 1 as per various proposed methods. Cancelable templates alike passwords are revocable if compromised. The cancelable templates are inculcated with necessary features like unlinkability and non-invertibility [7]. The cross matching can be avoided if the cancelable template is unlinkable while non-invertibility assures that the adversary could not invert the cancelable template back to original in any case. Unlinkability can be achieved, if for generating a cancelable template, a different one way transformation function is chosen for different services [7]. In addition, the choice of different parameter keys using the same transformation function for different applications can also result in different cancelable templates for the same user. Security for biometric demands that matching should be done in a transformed domain to avoid exposure of original biometric features during authentication. Hence, identification/verification is a challenge in transformed domain, as for a mated instance, the same distorted template generated during enrolment must be reproducible to achieve high recognition performance in a cancelable domain. However, one aspect of increasing the performance accuracy is pre alignment of biometric features. From the acquired biometric image, before feature extraction, pre-alignment is done. But investigation suggests that in recent years, researchers have claimed to achieve high recognition accuracy in transformed domains without pre-aligning the features. Hence, it's worthwhile here to investigate that absolute alignment is really an important requirement for biometric template protection. The need of registration or alignment has been explained very phenomenally in [4] for cancelable methods.Alignment is performed by extracting global features from the image while alignment-free approach considers to only extract local features.For example: in fingerprint feature extraction, to align, the singular points such

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 27,2021 at 10:26:26 UTC from IEEE Xplore. Restrictions apply.

as core and delta are used as a basis for a coordinate system to identify the location of neighbouring minutiae points. In contrast authors preferring alignment-free

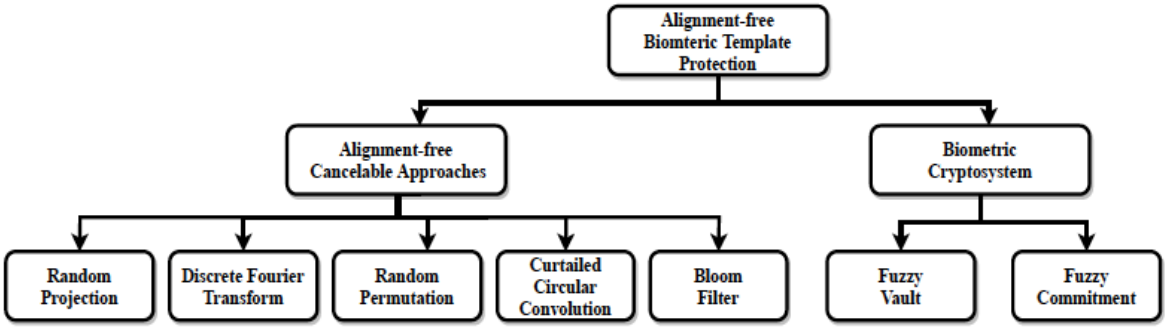


Fig 1: Proposed Classification of Alignment-Free based Biometric Template Protection Approach

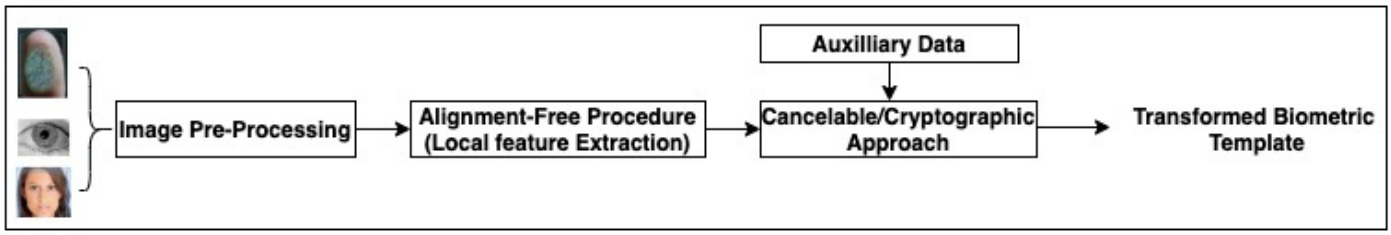


Fig 2: Basic Architecture of Alignment-Free based Biometric Template Protection Approach

approaches, ignore these global features and hence proposed different methods to extract local features. Local feature refers to local structures that consist of attributes which are invariant to global transformation like rotation, translation etc. Local structures, for example in fingerprints, are formed by segmenting the image into regions and then counting the number of minutiae points falling into each region [3]. Another way could be to divide a fingerprint into circular regions,then locating central minutiae in each region to create a reference minutiae point and lastly measuring the distance with neighbouring minutiae point which will be invariant in any case. These calculated distances could then be used as alignment-free features. The local structure reduces the computational complexity and increases the distortion tolerance Similarly, the need of alignment for irides in biometric template protection schemes, is a challenge. Further, this research work will overview the ways to solve the challenge by using the bloom filter. This process of using the local features during enrollment has repudiated the need of alignment in the protected domain during the verification process.To support these alignment free based biometric template protection schemes, a basic architecture has been illustrated in fig. 2. Our goal here is to discuss the proposed methods of local feature extraction along with the various cancelable and biometric applicable cryptographic methods. The article follows the following structure: Section 2 discusses about various alignment-free cancelable approaches for fingerprint followed by description of methods for IRIS in section 3. Section 4 reviews all the influential works on biometric cryptosystems. Following this, section 5 discusses the research gaps and future directions to benefit the future work of alignment free template protection domain.

II. ALIGNMENT-FREE CANCELABLE FINGERPRINT APPROACHES

Central to the entire discussion of biometric template protection is recognition performance. The matcher performance is indirectly associated with alignment of biometric images. This concept and belief has been challenged by many researchers by suggesting alignment-free methods and proving comparable results with alignment dependent cancelable approaches. Lee C et al. [5] pioneered the approach of alignment-free cancelable template generation. In this, the author used local minutiae information of fingerprint to make transformation and verification alignment-free. After pre-processing of image and minutiae extraction, rotation and translation invariant value for each minutia was extracted based on the orientation of its neighbouring regions and a user specific random vector. Further, to generate a cancelable template, minutia extracted was moved by a distance as per the distance changing function and the direction of movement was calculated by adding - original orientation of minutia and the rotational transformation value of orientation changing function. However, the author accepted that extracting invariant value could lead to error in practice. In particular, variations in control parameters of changing function provided a viable solution to influence performance and changeability. Following a similar approach, Chikkerur S et al. [7] proposed to use localized texture based representation. In an alternative approach, during enrollment, author preferred to use information of N*N (N=32) pixel patch around the extracted

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 27,2021 at 10:26:26 UTC from IEEE Xplore. Restrictions apply.

minutiae. These patches were aligned as per the orientation of minutiae. As per the author's assumption, each block of patch provides information about the unique identity of the subject. It was also found that few blocks of patches consist of rare appearances that helps to strongly identify an individual. Each of the gabor extracted patches were further encoded using compact 'signatures'. The signatures were robust and helped to reduce intra user variation and noise. These signatures were then transformed using a transformation key to achieve cancelability. The key used was specific to the user or application. The resultant signatures corresponding to each patch were stored in the database during enrollment and was matched against a set of signatures generated from the query fingerprint during verification. This method provided a viable solution to the privacy and security threats. In a follow-up study, Jin Z et al. [8] proposed Random triangle hashing method based on a cancelable approach. To make approach alignment-free, a reference minutia was selected from the set of minutiae (extracted from the fingerprint image). Based on the coordinates of reference minutia and its orientation , all other minutiae are translated and rotated on a 2D space. In contrast to Chikkerur S et al. [7] instead of patches, the author used a random triangle along with the secret keys to generate a hash string. This secret key is a set of random numbers that decides the vertex of a triangle in 2D space and is a user specific unique number, which makes the approach revocable. The number of minutiae in the random triangle was counted and orientation for each was calculated. These minutiae depending on the orientation range were transformed in a fixed length integer vector. Different keys assigned to the user generate different random triangles and hence different fixed length vectors for the same subject. These fixed length vectors were concatenated to form a hash vector. This key based hash vector is further converted into bit string to reduce the space complexity. Hence, the method claims to ensure high performance recognition accuracy. Zhang N et al. [20] proposed and applied a cancelable scheme on features extracted using Minutiae cylinder code (MCC) which is possibly an alignment-free approach as it deals with the local structure of minutiae. Also the use of MCC assures the removal of non-linear distortion. Here, two procedures - Combo-Plate Transformation and Functional transformation were followed to achieve cancelability.

Wang S et al. [24] proposed a method that ensured overcoming inaccurate singular point detection during recognition. The procedure unfolds as follows: Pair minutiae vectors were formed by calculating the distance, orientation and angle between the minutiae pairs. This vector set is quantized and is encoded into a binary string. The finite binary string and another random chosen finite binary string (act as a secret token) were circularly convolved for the L-point, generating DFT of these binary strings. Further, IDFT of the same helped to eliminate the redundant samples and reduced the size of the binary string, henceforth, cancelability was established. It's noteworthy here that binary string was chosen specifically to reduce inter class and intra class variability. The author claimed that the method renders strong security when both the transformed template and secret key is compromised. Also, that the method is computationally efficient. Although, matching accuracy degrades if the size of the secret key is too large. In the same frame, Sandhya M et al. [30] proposed formation of K-nearest neighbour structure to extract local and global features of minutiae. Further this structure was quantized and mapped onto a 2D array. This 2D array was then encoded into a binary string. The DFT was applied onto the binary string to generate a complex vector. This was further multiplied with a user specific random secret matrix to achieve cancelability. In another similar approach, using KNN clustering method, Nhat tran et al [42] proposed alignment-free approach in conjunction with partial discrete fourier transform. Using a similar approach, Wang S et al. [40] suggested preserving local features of minutiae by allocating all the minutiae in the fingerprint image into multiple circular zones with radius r while considering each and every minutiae as the centre of the zone. Thereafter, minutiae pairs were formed by considering the centre minutia and all other minutia of each zone. To make recognition resilient towards rotational and translational invariances, the features collected were expressed through distance, orientation angle and difference between angles.These collected features were then quantized by choosing a certain step size. The quantized value obtained for zoned minutia pairs were mapped to various locations of cubicles with 1 and 0. All the 1's and 0's stored in cubicles were concatenated to form a binary string. To mitigate the risk of an Attack via record multiplicity attack (ARM), modulo operation was applied to the index of the generated binary string which creates a new shortened binary string. Thereafter, partial DFT, generated using a parameter key, was applied to the new shortened binary string. The variable used for modulo-operation and the parameter key were both user specific. Similarly, Nazmul et al. [44] proposed to use zoning methods to extract local minutiae features to achieve alignment-free matching in cancelable domain. In another alignment-free approach, Wang S et al. [36] uses minutiae-pair vectors to extract local information. These vectors were quantized to be encoded into binary notation of n bits. A binary bin with values 2n was indexed by 1; ˆ else the binary bin was assigned with '0'. This results in a binary string of 1 and 0. Subsequently, a DFT is applied on the binary string which evenly distributes the binary string and was further passed through a FIR filter composed of a user specific parameter key to generate a complex matrix.To ensure non-invertibility, matrix should be necessarily column rank deficient, which depends upon the choice of parameter key.

Lee et al. [11] proposed an approach, where instead of selecting a single reference minutiae, the reference minutia was changed in turn with other minutiae, resulting in formation of m 1-D bit strings followed by permutation to achieve cancelability. However, the method was found to be insecure in the stolen-token scenario. By using polar coordinate space, Ahmad et al. [12] proposed to generate alignment-free templates. If the neighbouring minutiae distance was found greater than the threshold distance, its location was calculated relative to a central minutia based on information such as angle, distance and orientation. Further to perform transformation, polar coordinate space is divided into sectors and sectors were randomly mapped to a different position as determined by random vectors. The angle, distance and orientation of trans-positioned minutiae were used as feature vectors. Yang et al. [21] described the alignment-free

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 27,2021 at 10:26:26 UTC from IEEE Xplore. Restrictions apply.

approach by forming local structure using a delaunay triangulation net followed by a polar transformation approach to achieve non-invertibility.

A process called minutiae vicinity decomposition(MVD) helped Jin et al. [25] to derive a set of random invariant geometric features. The MVD involves formation of vicinity followed by minutiae quadruplet formulation which was disintegrated into

TABLE I: Comparison of Recent Alignment-Free Cancelable Biometric Approaches

Author	Methodology	Modality	Dataset	EER(%Average)	Remarks
Lee C et al [5]	Zoning & Random Mapping	Fingerprint	FVC2002 DB1	3.4	worst case scenario not analysed
Ahmad et al. [12]	Zoning & Random Mapping	Fingerprint	FVC2002 DB1, DB2, DB3	14.3	worst case scenario analysed
Chikkerur et al [7]	Localised Patch & Random Projection	Fingerprint	Self generated	-	worst case scenario not analysed
Lee et al. [11]	Coordinate Mapping	Fingerprint	FVC2004 DB1,FVC2004 DB2,FVC2004 DB3	8.6	worst case scenario analysed but performance is low
Ferrara et al. [18]	MCC(Zoning)	Fingerprint	FVC2002,FVC2006 DB2, FVC2006 DB2b	2.7	worst case scenario not analysed
Das et al. [15]	MDG(Zoning)	Fingerprint	FVC2002-DB1a, FVC2002-DB2a	2.4	worst case scenario analysed
Jin et al [16]	Polar grid based 3 Tuple Quantization Technique	Fingerprint	FVC2002 (DB1 & DB2), FVC2004 (DB1&DB2)	9.54	worst case scenario analysed with good performance
Zhang et al [20]	Minutia Cylinder Code Representation	Fingerprint	FVC2006DB2	-	worst case scenario not analysed
Yang et al [21]	Delaunay Triangle	Fingerprint	FVC2002 DB1 and DB2	4.9	worst case scenario not analysed
Wang et al. [24]	Curtailed Circular Convolution	Fingerprint	FVC2002 DB1, DB2 and DB3	3.6	worst case scenario analysed
Sadhya et al. [39]	Hybrid Approach & DFT	Fingerprint	FVC2002	5.64	worst case scenario analysed
Wang et al. [36]	Quantized Pair Minutiae	Fingerprint	FVC2002 DB1, DB2 and DB3.-	-	worst case scenario analysed
Wang et al. [40]	Zoning & Partial DFT	Fingerprint	FVC2002, FVC2004(DB2)	5.4	worst case scenario analysed
Alam et al. [43]	Polar Grid based 3-tuple Quantization	Fingerprint	FVC2002, FVC2004	12.86	worst case scenario not analysed
Kho et al. [46]	Partial Local Structure (PLS) descriptor and Permutated Randomized Non-Negative Least Square (PR-NNLS)	Fingerprint	FVC2002, FVC2004	4	worst case scenario analysed
Abdullahi et al. [51]	Fourier-Mellin Transform	Fingerprint	FVC2002, FVC2004	2.32	worst case scenario analysed

into four minutiae triplets. Length, angle and orientation, calculated from these four disintegrated mintiate triplets were collected to form a 1-D vector which was then randomly projected to form randomized MVD. To add non-invertibility,

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 27,2021 at 10:26:26 UTC from IEEE Xplore. Restrictions apply.

the obtained MVD was then embedded with graph based hamming technique to seek a binary string of size 128 or 256 bits in correspondence to N minutiae points. These binary strings were converted into integer numbers selecting 8 bit from the binary string. This reduced the size of the cancelable template for storage. The performance of this method is comparable to [12] but shows poor performance in comparison to [18]. In another minutia descriptor based approach, Kho et al. [46] has used partial local structure (PLS) to generate a fingerprint feature vector which is discriminative and compact in size. To add unlinkability and non-invertibility, the feature vector using a user specific permutation key was transformed using randomized non-negative least square method. The method seems to be quite complex, although has maintained high performance in comparison to [18] and [25]. Using the invariance property of Fourier-Mellin Transform, another alignment-free approach was proposed by Abdullahi et al. [51]. The cancelable feature in form hash was generated by using fractal coding. Extending the idea of polar grid based 3-tuple quantisation proposed by [17], [43] added bit toggling strategy to add noise into biometric template and hence incorporates non invertibility. To increase the complexity for adversary, DFT along with random projection was applied on noise added biometric template to generate complex vector.Following a similar approach, Jin et al. [16] method seems to be a combination of [11] and [12] to attain non invertibility as well as cancelability. In another approach, Wang et al. [14] proposed a mathematical approach that realized many to one mapping by representing intersection of higher dimensional hyperplanes, formed using minutia pairs, into algebraic linear transformation. The process was termed as dense infinite to one mapping (DITOM). Walia et al. [52] proposed to use deep learning feature extraction procedure along with random projection to add non-invertibility. To conclude, a hybrid approach was proposed by Walia et al. [50]. This approach proposed an alignment-free method using graph based fusion. The method utilises key based feature extraction of face, iris and fingerprint. A similarity graph and its normalised version is produced for all the modalities. These normalised graphs were then cross diffused and concatenated with weight vectors to generate a cancelable template. In another approach Walia et al. [47] proposed to use score level fusion and a key biometric feature for multiple modalities. These key features were then combined with the subject biometric feature to generate a cancelable template for each modality. Later these transformed templates of all modalities were diffused using random walk cross diffusion. Also, classifier score of each modality was fused using multistage optimal score level fusion method. Using fingervein, Kirchgasser et al. [48] has proposed an alignment free approach. Fingervein features were extracted in binary form. The distance between location of binary feature vector was calculated and stored as vector. Again, using Kronecker delta function, another vector was formed consisting of 0 and 1. Both these vectors were then combined to generate an alignment-free robust transformation.

These different alignment-free feature extraction approaches differ a lot on the basis of computational efficiency,complexity and size of biometric template produced. Few of the above discussed methods use a complex process to extract alignment-free features while few focus mainly on reducing the size of the feature template in addition to alignment-free capability. Investigation reveals that very few authors have tested these cancelable approaches under the worst case scenario for alignment free approaches. In addition, these suggested methods are restricted to only fingerprint and cannot be extended to other modalities. To cater this need, in the next section, an alignment free approach using iris has been discussed.

III. ALIGNMENT-FREE CANCELABLE IRIS APPROACHES

Permanence and uniqueness, are two features that are strictly common between iris and fingerprint. After extensively re-viewing the literature for alignment-free cancelable methods, it is evident that initially alignment-free approach was restricted to a single modality, namely fingerprint. But with the widespread deployment of automated biometric technology using other popular and unique modalities such as face and iris, necessitated an alignment-free approach for other modalities in the transformed domain for achieving better performance and accuracy.

Rathgeb C et al. [23] first adapted the concept of bloom filters- an idea of probabilistic data structure, for protection of biometric templates. Author [19] proposed an alignment free cancelable approach for iris using adaptive bloom filters. Protection of templates, depends upon the set of bloom filters used. Extracted iris code was divided into blocks of fixed length. Each block consists of a fixed amount of columns. These columns were converted into codewords. These code words suggested the indexes within the bloom filter that need to be stored with value 1. Hence, at the time of verification for a mated instance, certain blocks would consist of an equal number of columns, leading to identically mapped indexes within the bloom filter. This particularly eliminated self-propagating errors caused by misalignment of iris codes. Misalignment due to rotational variance was also taken care off due to use of location information of iris codes columns. The proposed concept of bloom filter is used along with an application specific secret key to achieve unlinkability and irreversibility, hence preserving privacy and secrecy. This method mixed the iris codes of irides at feature level before any application of bloom filter to lower the intra-class variability. Hence mixing of multi-biometric characteristics enhances the recognition accuracy.

In another approach [22], to ensure irreversibility, feature level mixing of iris codes is performed. In this process, the extracted iris codes were divided into K blocks of equal size. Each block consists of a sequence of columns. Each column regarded as codeword of the block was converted into a decimal value which was further XOR-ed with the application specific secret key and was subsequently mapped to the bloom filter. Thereby, a protected biometric template is generated by concealing the original positions of codewords and mapping the diverse columns to the same index. Rathgeb C et al. [31] in response to the security vulnerabilities analysis done by Hermans et al. [26][28], proposed a solution to mitigate the risk of cross matching, hence ensuring unlinkability. Post bloom filter based transformation using face and iris

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 27,2021 at 10:26:26 UTC from IEEE Xplore. Restrictions apply.

templates, feature level fusion of both the transformed template is done using OR logic.This OR-ing of features results in concealment of biometric features and hence reduces the risk of privacy invasion. Bringer J et al.[28][38] proposed security and privacy analysis of bloom filter based cancelable templates.Marta G B et al. [34] proposed an improved cancelable face template protection scheme using bloom filters along with a framework for unlinkability and irreversibility. Followed by feature extraction was a structure preserving feature re-arrangement objectively performed for irreversible transforms on biometric feature vectors. To achieve high recognition accuracy, the method preserves certain discriminative words of feature blocks. The feature blocks and groups of feature blocks were then applied with different permutation keys to perform perm-transform. The unlinkability and irreversibility of this method was affected by the size of the permutation key. To overcome the weakness of [8], Sandhya et al. [39] proposed to use a unique key for each bloom filter while considering the size of the key equal to the size of bloom filter. Although this increases the space complexity but does not degrade the performance. In another approach, Lai L et al. [41] mixed index first one hashing approach with bloom filter to achieve alignment-free cancelability. In another different approach, Jacob et al. [49] has put forth a concept of improving biometric recognition accuracy using capsule based networks replacing CNN. The capsule based network is capable of getting trained with less biometric samples in comparison to CNN where a large dataset is required to train the deep learning system.

TABLE II: Comparison of Recent Alignment-Free Cancelable Biometric Approaches

Author	Methodology	Modalities	Dataset	EER(%Average)	Remarks
Rathgeb et. al. [19]	Adaptive bloom filter	Iris	CASIA-v3-Interval Iris	2.5	worst case scenario not analysed
Rathgeb et. al. [22]	Adaptive Bloom Filter	Iris, Face	IITD	5.0	worst case scenario not analysed
Rathgeb et. al. [31]	Adaptive Bloom Filter	Iris, Face	IITD	6.0	worst case scenario not analysed
Gomez et. al. [34]	Adaptive Bloom Filter	Iris Face	Biosecure Multimodal Database	6.0	worst case scenario analysed
Lai et. al. [41]	IFO hashing & Bloom Filter	Iris Face	CASIA V3-Internal Iris	6.9	worst case scenario not analysed

The only alignment-free approach found during investigation for iris is using bloom filters. This approach has been constantly improved in terms of performance since its first inception using a multimodal approach. Table 1 and Table 2 summarises all the alignment-free cancelable approaches for fingerprint and iris, respectively. The security of biometric templates is not only limited to cancelable schemes. Hence, the next section introduces alignment-free biometric cryptosystem of template protection.

IV. ALIGNMENT-FREE BIOMETRIC CRYPTOSYSTEM

The use of cryptographic methods in conjunction with error control codes has led to a key binding technique - the fuzzy commitment scheme, where biometric characteristics bounded with secret key leads to protections of biometric template [37]. The process involves generating a secret key using error control code of the same length as the biometric feature vector. Subsequently, XOR of the key and biometric feature generates a difference feature vector which is stored along with the hash value of key during enrollment. While performing key retrieval, the query feature template is XOR-ed with the stored difference feature vector followed by the decoding process. The key is only retrieved, when a biometric feature matches the stored template and is matched against the hashed key [9]. In another cryptographic key binding approach to protect biometric templates, Fuzzy vault was introduced [1]. This primitive approach involves generation of polynomials using secret data. This binomial feature was projected on this polynomial to generate genuine point pairs. In addition, chaff points were generated that do not lie on a polynomial. Chaff points mixed with genuine point pairs were then stored as fuzzy vaults. To access the secret data, the query biometric template was matched against the stored vault. If the difference between the two was found small, the genuine points get discriminated from chaff points. Finally, the polynomial was reconstructed and secret data was extracted. The fuzzy vault doesn't stress on the sequence but equality in sets of biometric feature vectors. Both these key binding schemes are limited to unimodal systems such as fingerprint [6] and iris [37].

Alignment of biometric features was considered as the necessary condition for successful retrieval of secret data [19]. To perform alignment for query feature template, auxiliary alignment data is needed to be stored in a database which can certainly leak the biometric information if stolen [6]. To overcome this issue, researchers came across an idea of using an alignment-free fuzzy vault and fuzzy commitment scheme. Local structure formation using relative information from biometric images, features into a less discriminative information. In reference to this limitation, Li C et al. [35] proposed a fuzzy vault based cryptosystem using minutiae pair-polar structure. Experimentally, the author claims that feature extraction using pair-polar method provides more

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 27,2021 at 10:26:26 UTC from IEEE Xplore. Restrictions apply.

discriminative features. These discriminative features are finely quantized and then transformed using a pre-set seed and an invariant parameter, before being used in a fuzzy vault. This transformation destroys the correlation between the original feature and the features extracted using pair polar features, hence enhancing privacy. The fuzzy vault scheme reversing the transformed template to original. In addition, the global minutiae matcher has also been transformed into a transformation invariant feature application version. The method ensures accurate recognition performance, revocability, irreversibility and unlinkability to a greater extent through its experimental evaluation. Tams et al. [33] used three different minutiae descriptor- minutia orientation descriptor, minutia frequency descriptor and local minutia structure to enrich fuzzy vault with alignment free approach. In another approach, Rathgeb et al.[37] claimed an alignment-free method for iris. The method proposed fusion of most discriminative features from multiple instances of iris (left and right eye) at feature level. Thereafter, the fused features were locked in a fuzzy vault to generate non-invertible and unlinkable templates. This method was introduced to protect templates from cross-matching attacks without reducing the biometric recognition performance. In another fusion based approach, Li et al. [10] proposed an alignment-free based fuzzy vault based biometric template protection system. The method employed three different fusion strategies such as separated rule, product rule and sum rule to fuse different kinds of local features such as minutia descriptor and minutia local structure.In spite of using huffman coding to compress the storage size required by fingerprint vault, the need for large storage space remains an unsolved issue. Another method used a near equivalent version of dual layer structure check proposed by Xi et al. [13] to introduce alignment-free generation of biometric keys in fuzzy extractor. A concept of alignment free multiple fuzzy vaults was proposed by Nasiri et al. [29]. While performing verification, if two from multiple vaults, generated during enrolment, are decoded successfully by query fingerprint template, the cryptographic key can be recovered. Using modified voronoi local structure, Yang et al. [27] proposed an alignment-free secure sketch method to enhance security and privacy. A hybrid approach of key binding and cancelable scheme was proposed by Chai et. al. [45]. The approach uses fuzzy commitment without error correction code. The cryptographic key was bonded with a cancelable iris template based on a randomly generated pattern of 1 and 0. The iris template is made cancelable using alignment free bloom filter and random permutation. The key generated during enrollment is matched against the key generated by the query template. In another hybrid approach, Leng et al. [32] blended cancelable palmprint template with primitive fuzzy vault scheme to generate alignment free fuzzy vault. Li et al [35] proposed to use polar-pair minutiae structure for producing highly discriminative features to get rid of dependency of registration.

Investigation reveals that the major inclination of the biometric research community is towards cancelable schemes, due to two limitations - Firstly, the security analysis done for fuzzy vault and fuzzy commitment schemes shows the system vulnerability towards more security and privacy attacks. Secondly, high computational and space complexity shown by which in general selects the chaff points that are ‘d’ distance apart, has been replaced by an idea of selecting chaff and genuine minutiae points that are 2d distance apart. This step in particular protects the transformed genuine minutiae features by increasing the complexity for adversary biometric cryptosystems. It's noteworthy that conjunction of non-invertible transformation and fuzzy vault can enhance the secrecy of biometric templates and could avoid cross matching attacks. Hence, the need of the hour is to include hybrid approaches of biometric template protection schemes in conjunction with deep learning algorithms that hopefully makes systems more adaptable to dynamic environments. But the biometric community has still not completely adopted machine and deep learning methods to improve recognition accuracy due to limited availability of biometric datasets.

V. CONCLUSION

Cancelable and Cryptographic approaches suggested for different modalities such as iris, face, fingerprint etc, all demand pre-alignment, especially when matching has to be done in a transformed domain. Investigation shows that global registration only increases computational complexity and has minimum tolerance for distortion. While the local feature extraction method not only gives us comparable recognition performance but also decreases computational complexity and distortion. The use of cancelable or cryptographic approaches on invariant attributes collected from local structure to form feature vectors enhances recognition performance in transformed domains. Review suggests that the different methods proposed for extracting local features seem to be variants of few novel approaches. Hence, more novel methods for local feature extraction must be proposed. The cancelable and cryptographic methods discussed in this review seem to be in common with alignment dependent template protection approaches. One of the main issues found with local feature extraction is low distinctiveness for intra-class variation that highly affects recognition performance. To enhance performance accuracy, the first possible viable solution is hybrid approach, where advantages of global and local registration could be combined to achieve high performance accuracy. The second could be the use of deep learning for feature extraction and pre-processing of images. TheV third is to use fusion of multiple modalities that also eradicates the need of alignment. Further, the future research of alignment-free biometric template protection must not be limited to only fingerprint and iris. There is a need to explore and suggest methods that allow systems to be adaptive in the dynamic environment.

REFERENCES

[1] A. Juels and M. Sudan, “A fuzzy vault scheme,” IEEE Int. Symp. Inf. Theory - Proc., p. 408, 2002.
[2] A. K. Jain, A. Ross, and S. Prabhakar, “An Introduction to Biometric Recognition,” IEEE Trans. Circuits Syst. Video Technol., vol. 14, no. 1, pp. 4–20, 2004
[3] D. Maltoni, D. Maio, A. K. Jain and S. Prabhakar, Handbook of Fingerprint Recognition, New York:Springer-Verlag, 2003.
[4] N. K. Ratha, S. Chikkerur, J. H. Connell, and R. M. Bolle, “Generating cancelable fingerprint templates,” IEEE Trans.Pattern Anal.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 27,2021 at 10:26:26 UTC from IEEE Xplore. Restrictions apply.

Mach. Intell., 2007
[5] C. Lee, J. Y. Choi, K. A. Toh, S. Lee, and J. Kim, "Alignment-free cancelable fingerprint templates based on local minutiae information," IEEE Trans. Syst. Man, Cybern. Part B Cybern., vol. 37, no. 4, pp. 980–992, 2007.
[6] K. Nandakumar, A. K. Jain, and S. Pankanti, "Fingerprint-based fuzzy vault: Implementation and performance," IEEE Trans. Inf. Forensics Secur., vol. 2, no. 4, pp. 744–757, 2007, doi: 10.1109/TIFS.2007.908165.
[7] S. Chikkerur, N. K. Ratha, J. H. Connell, and R. M. Bolle, "Generating registration-free cancelable fingerprint templates," in BTAS 2008 - IEEE 2nd International Conference on Biometrics: Theory, Applications and Systems, 2008 .
[8] Z. Jin, A. B. Jin Teoh, T. S. Ong, and C. Tee, "Secure minutiae-based fingerprint templates using random triangle hashing," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2009.
[9] S. Kanade, D. Petrovska-Delacretaz, and B. Dorizzi, "Cancelable iris ´ biometrics and using error correcting codes to reduce variability in biometric data," 2009 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work. CVPR Work. 2009, vol. 2009 IEEE, pp. 120–127, 2009.
[10] P. Li, X. Yang, K. Cao, X. Tao, R. Wang, and J. Tian, "An alignment free fingerprint cryptosystem based on fuzzy vault scheme," J. Netw. Comput. Appl., vol. 33, no. 3, pp. 207–220, 2010.
[11] C. Lee and J. Kim, "Cancelable fingerprint templates using minutiae based bit-strings," J. Netw. Comput. Appl., vol. 33, no. 3, pp. 236–246, 2010, doi: 10.1016/j.jnca.2009.12.011.
[12] T. Ahmad, J. Hu, and S. Wang, "Pair-polar coordinate-based cancelable fingerprint templates," Pattern Recognit., vol. 44, no. 10–11, pp. 2555–2564, 2011.
[13] K. Xi, J. Hu, and F. Han, "An alignment-free fingerprint fuzzy extractor using near-equivalent Dual Layer Structure Check (NeDLSC) algorithm," Proc. 2011 6th IEEE Conf. Ind. Electron. Appl. ICIEA 2011, pp. 1040–1045, 2011.
[14] S. Wang and J. Hu, "Alignment-free cancellable fingerprint template design: A densely infinite-to-one mapping (DITOM) approach", Pattern Recognit., vol. 45, pp. 4129-4137, Dec. 2012.
[15] P. Das, K. Karthik, and B. Chandra Garai, "A robust alignment-free fingerprint hashing algorithm based on minimum distance graphs," Pattern Recognit., vol. 45, no. 9, pp. 3373–3388, 2012.
[16] Z. Jin, A. B. Jin Teoh, T. S. Ong, and C. Tee, "Fingerprint template protection with minutiae-based bit-string for security and privacy preserving," Expert Syst. Appl., vol. 39, no. 6, pp. 6157–6167, 2012.
[17] Z. Jin, T. S. Ong, C. Tee, and A. B. J. Teoh, "Generating revocable fingerprint template using polar grid based 3-tuple quantization technique," in Midwest Symposium on Circuits and Systems, 2011.
[18] M. Ferrara, D. Maltoni, and R. Cappelli, "Noninvertible minutia cylinder-code representation," IEEE Trans. Inf. Forensics Secur., vol. 7, no. 6, pp. 1727–1737, 2012.
[19] C. Rathgeb, F. Breitinger, and C. Busch, "Alignment-free cancelable iris biometric templates based on adaptive bloom filters," in Proceedings - 2013 International Conference on Biometrics, ICB 2013, 2013.
[20] N. Zhang, X. Yang, Y. Zang, X. Jia, and J. Tian, "Generating registration free cancelable fingerprint templates based on Minutia Cylinder-Code representation," in IEEE 6th International Conference on Biometrics: Theory, Applications and Systems, BTAS 2013, 2013.
[21] W. Yang, J. Hu, S. Wang, and J. Yang, "Cancelable fingerprint templates with Delaunay triangle-based local structures," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 8300 LNCS, pp. 81–91, 2013.
[22] C. Rathgeb and C. Busch, "Cancelable multi-biometrics: Mixing iris codes based on adaptive bloom filters," Comput. Secur., vol. 42, pp. 1–12, 2014.
[23] C. Rathgeb, F. Breitinger, C. Busch, and H. Baier, "On application of bloom filters to iris biometrics," IET Biometrics, 2014.
[24] S. Wang and J. Hu, "Design of alignment-free cancelable fingerprint templates via curtailed circular convolution," Pattern Recognit., vol. 47, no. 3, pp. 1321–1329, 2014.
[25] Z. Jin, M. H. Lim, A. B. J. Teoh, and B. M. Goi, "A non-invertible Randomized Graph-based Hamming Embedding for generating cancelable fingerprint template," Pattern Recognit. Lett., vol. 42, no. 1, pp. 137–147, 2014.
[26] J. Hermans, B. Mennink and R. Peeters, "When a bloom filter is a doom filter: Security assessment of a novel iris biometric template protection system", Proc. BIOSIG, pp. 1-6, 2014.
[27] W. Yang, J. Hu, S. Wang, and M. Stojmenovic, "An alignment-free fingerprint bio-cryptosystem based on modified Voronoi neighbor structures," Pattern Recognit., vol. 47, no. 3, pp. 1309–1320, 2014.
[28] J. Bringer, C. Morel, and C. Rathgeb, "Security analysis of Bloom filter-based iris biometric template protection," in Proceedings of 2015 International Conference on Biometrics, ICB 2015, 2015.
[29] A. A. Nasiri and M. Fathy, "Alignment-free fingerprint cryptosystem based on multiple fuzzy vaults," Proc. Int. Symp. Artif. Intell. Signal Process. AISP 2015, pp. 251–255, 2015.
[30] M. Sandhya and M. V. N. K. Prasad, "K-Nearest Neighborhood Structure (k-NNS) based alignment-free method for fingerprint template protection," in Proceedings of 2015 International Conference on Biometrics, ICB 2015, 2015.
[31] C. Rathgeb, M. Gomez-Barrero, C. Busch, J. Galbally, and J. Fierrez, "Towards cancelable multi-biometrics based on bloom filters: A case study on feature level fusion of face and iris," 3rd Int. Work. Biometrics Forensics, IWBF 2015, 2015.
[32] L. Leng and A. B. J. Teoh, "Alignment-free row-co-occurrence cancelable palmprint Fuzzy Vault," Pattern Recognit., vol. 48, no. 7, pp. 2290–2303, 2015.
[33] B. Tams, J. Merkle, C. Rathgeb, J. Wagner, U. Korte, and C. Busch, "Improved fuzzy vault scheme for alignment-free fingerprint features," Lect. Notes Informatics (LNI), Proc. - Ser. Gesellschaft fur Inform., vol. P-245, 2015.
[34] M. Gomez-Barrero, C. Rathgeb, J. Galbally, C. Busch, and J. Fierrez, "Unlinkable and irreversible biometric template protection based on bloom filters," Inf. Sci. (Ny)., vol. 370–371, pp. 18–32, 2016.
[35] C. Li and J. Hu, "A security-enhanced alignment-free fuzzy vault-based fingerprint cryptosystem using pair-polar minutiae structures," IEEE Trans. Inf. Forensics Secur., 2016.
[36] S. Wang and J. Hu, "A blind system identification approach to cancelable fingerprint templates," Pattern Recognit., vol. 54, pp. 14–22, 2016.
[37] C. Rathgeb, B. Tams, J. Wagner, and C. Busch, "Unlinkable improved multi-biometric iris fuzzy vault," Eurasip J. Inf. Secur., vol. 2016.
[38] J. Bringer, C. Morel, and C. Rathgeb, "Security analysis and improvement of some biometric protected templates based on Bloom filters," Image Vis. Comput., vol. 58, pp. 239–253, 2017.
[39] D. Sadhya and S. K. Singh, "Providing robust security measures to Bloom filter based biometric template protection schemes," Comput. Secur., vol. 67, pp. 59–72, 2017
[40] S. Wang, W. Yang, and J. Hu, "Design of Alignment-Free Cancelable Fingerprint Templates with Zoned Minutia Pairs," Pattern Recognit., vol. 66, no. November 2016, pp. 295–301, 2017.
[41] Y. L. Lai, B. M. Goi, and T. Y. Chai, "Alignment-free indexing-first one hashing with bloom filter integration," 2017 IEEE Int. Conf. Intell. Secur. Informatics Secur. Big Data, ISI 2017, pp. 78–82, 2017
[42] Q. Nhat Tran, J. Hu, and S. Wang, "Alignment-Free Cancellable Template with Clustered-Minutiae Local Structure," 2018 IEEE Glob. Commun. Conf. GLOBECOM 2018 - Proc., pp. 1–6, 2018.
[43] B. Alam, Z. Jin, W.-S. Yap and B.-M. Goi, "An alignment-free cancelable fingerprint template for bio-cryptosystems", J. Netw. Comput. Appl., vol. 115, pp. 20-32, Aug. 2018.
[44] R. Nazmul, M. R. Islam, and A. R. Chowdhury, "Alignment-free fingerprint template protection technique based on minutiae neighbourhood information," Adv. Intell. Syst. Comput., vol. 580, pp. 256–265, 2018
[45] T. Y. Chai, B. M. Goi, Y. H. Tay, and Z. Jin, "A new design for alignment-free chaffed cancelable iris key binding scheme," Symmetry (Basel)., 2019.
[46] J. B. Kho, J. Kim, I. J. Kim, and A. B. J. Teoh, "Cancelable fingerprint template design with randomized non-negative least squares," Pattern Recognit., vol. 91, pp. 245–260, 2019.
[47] G. S. Walia, S. Rishi, R. Asthana, A. Kumar, and A. Gupta, "Secure multimodal biometric system based on diffused graphs and optimal score fusion," IET Biometrics, 2019.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 27,2021 at 10:26:26 UTC from IEEE Xplore. Restrictions apply.

[48] S. Kirchgasser and A. Uhl, "Finger-Vein Template Protection based on Alignment-Free Hashing," 10th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS 2019), 2019.
[49] Jacob, I. J., "Capsule Network Based Biometric Recognition System,". Journal of Artificial Intelligence, vol.01/issue 02, pp. 83-94,2019.
[50] G. S. Walia, G. Jain, N. Bansal, and K. Singh, "Adaptive Weighted Graph Approach to Generate Multimodal Cancelable Biometric Templates," IEEE Trans. Inf. Forensics Secur., vol. 15, pp. 1945–1958, 2020.
[51] S. M. Abdullahi, H. Wang and T. Li, "Fractal Coding-Based Robust and Alignment-Free Fingerprint Image Hashing," in IEEE Transactions on Information Forensics and Security, vol. 15, pp. 2587-2601, 202
[52] G. S. Walia, K. Aggarwal, K. Singh, and K. Singh, "Design and Analysis of Adaptive Graph based Cancelable Multi-Biometrics Approach," IEEE Trans. Dependable Secur. Comput., vol. 5971, no. c, pp. 1–13, 2020.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 27,2021 at 10:26:26 UTC from IEEE Xplore. Restrictions apply.

Photonics and Nanostructures - Fundamentals and Applications 43 (2021) 100881

Contents lists available at ScienceDirect

Photonics and Nanostructures - Fundamentals and Applications

journal homepage: www.elsevier.com/locate/photonics

All dielectric metasurface based tunable optical modulator: Design and analysis

Keshav Samrat Modi [a,b], Satya Pratap Singh [b], Jasleen Kaur [a,b], Umesh Tiwari [a,b], Ravindra Kumar Sinha [c,*]

[a] Academy of Scientific and Innovative Research, Ghaziabad, 201002, India
[b] Computational Photonics Lab, Council of Scientific & Industrial Research (CSIR)- Central Scientific Instruments Organisation (CSIO), Chandigarh, 160030, India
[c] TIFAC-Centre of Relevance and Excellence in Fiber Optics and Optical Communication, Department of Applied Physics, Delhi Technological University (Formerly Delhi College of Engineering), Bawana Road, Delhi, 110042, India

ARTICLE INFO

Keywords:
All-dielectric metasurface
Fano resonance
Liquid crystal
Optical modulation

ABSTRACT

In this article, we propose the design and analysis of all-dielectric metasurface based tunable optical modulator. The metasurface structure consists of split silicon arcs mounted over the indium tin oxide layer and the silica substrate. The nematic liquid crystal is used for tuning the Fano resonance peak of the metasurface and is filled between two indium tin oxide layers. The numerical investigation of the proposed metasurface for electro-optic modulation shows the modulator has a broadband tunability of 25 nm and modulation depth of 90 %. We have also numerically investigated the variation of the Fano resonance peak arising due to the fabrication non-uniformities in the thickness, width and the split gap of the silicon arc.

1. Introduction

The Metamaterials are the specially engineered composite materials having a specific function in the desired wavelength range. Metamaterials are either plasmonic and/or all-dielectric composites. The plasmonics metamaterials, have applications in the development of various devices [1–6], show high thermal and ohmic losses due to negative permittivity and high impedance of the metal components while in case of the all-dielectric metamaterials these losses are quite small in the visible and infrared (IR) wavelength range [7–11]. Hence the Q-Factor and spectral contrast of plasmonic metamaterials are relatively low compared to the all-dielectric metamaterials. The all-dielectric metamaterials have control over the enhancement in the field, inside or outside of the materials. In the case of all-dielectric structure, both electric hotspot, as well as a magnetic hotspot, can be achieved which is not possible with the plasmonics structure in the high-frequency regime [8,9,12,13].

The metamaterial is a resonating structure in the desired wavelength range. The resonance line-shape depends on the design structure of the metamaterial. The fundamental resonance line-shape is symmetric Lorentzian, unlike asymmetric Fano line-shape. The Fano resonance peak is narrower and sharper than the Lorentzian resonance. Fano resonance is destructive interference of the anti-phase lattice collective resonance (ALCR) and the in-phase lattice collective resonance (ILCR) and characterised by the dark and bright electromagnetic modes having narrow and broad resonance [3,14–17]. The sharp Fano-resonance peak has high Q-Factor [18,19] and thus the all-dielectric metasurfaces have found applications [8,20] in biological sensing [21–25], nanoantenna, zero-index material [26,27], solar cell, absorbers, filters [16], modulator [18,28,29].

The modulation of light characteristics can be done by using a number of stimuli like acoustic [30,31], electrical, optical signal. The electro-optic modulation of light is based on the electro-optic effect or Pockel's effect. The electro-optic effect can be achieved by electrically controlling the beta barium borate (BBO) crystals [32], $LiNbO_3$ crystal [33], using piezoelectric buzzer [34] or by electrically controlled the liquid crystal alignment [28,35,36]. The response of the metasurface based modulator can be tuned by the using electrical [18], acoustic, thermal [37], optical [2] signals as per the desired application. In case of electrical tuning, nematic liquid crystals are the material used for tuning of the response of the metasurface. Liquid crystals are the anisotropic substance having rod-like molecule shape [38]. The liquid crystal in optical wavelength range have properties like low absorption, high transmission and high birefringence thus promising material for

* Corresponding author.
E-mail address: dr_rk_sinha@yahoo.com (R.K. Sinha).

https://doi.org/10.1016/j.photonics.2020.100881
Received 12 March 2020; Received in revised form 9 October 2020; Accepted 16 November 2020
Available online 26 November 2020
1569-4410/© 2020 Elsevier B.V. All rights reserved.

fabrication of electro-optic modulator like optical devices [39].

The electro-optic modulators implemented using liquid crystal based metasurface were studied [16,18,28,29,40–44]. The metasurface based electro-optic modulator shows the maximum modulation depth up to 85 % [18,43] while the shift in wavelength due to variation in the refractive index is just 10 nm [18]. In another case where the shift in wavelength is high up to 100 nm [29,40] but the modulation index is as low as 30 % [40]. Thus, a metasurface based electro-optic modulator having both high modulation depth as well as large wavelength shift is not reported.

In this paper, we are presenting an all-dielectric metasurface based modulator having both high modulation depth as well as large wavelength shift. To achieve this device, we have utilized an all-dielectric metasurface structure having split silicon arcs mounted over a silica substrate coated with the thin layer of indium tin oxide. A high birefringence liquid crystal is filled between the metasurface and the other indium tin oxide layer for electro-optical modulation applications.

2. Design of metasurface and numerical study

As shown in Fig. 1 the proposed modulator comprised of the liquid crystal cell which consists of asymmetric silicon arcs mounted over the ITO layer and the silica substrate. The thickness of the arcs is 345 nm and depth are 180 nm. The asymmetric minor and major arcs have the extent which creates the central angle of 150° and 160° respectively. The silicon arcs are mounted over the silica substrate of 180 nm thickness. The Si arcs are bisected by the split gap of 30 nm. The refractive index of the silicon is taken as 3.48 and that of the silica is 1.45. Since dielectric materials are used here so the structure has negligible ohmic losses. It is considered during the simulation that the two-indium tin oxide (ITO) layers are deposited over the silica layer enclosing nematic liquid crystals and asymmetric silicon arcs. The ITO layers act as the terminals to the applied potential biasing. The refractive index of the 180 nm thick indium tin oxide layers is taken to be 1.7552. The Electromagnetic Waves, Frequency Domain under Wave Optics module of COMSOL Multiphysics software tool was utilized for simulation. The above-described structure is periodic in the x- and y-directions. The x-polarized plane wave electromagnetic field (light) is incident normally (along the z-direction) to the metasurface. The perfectly matching layers are defined at the open boundaries in the z-direction. The element size of the tetrahedra mesh of $\lambda/10$ is considered. The optical constants for the materials are considered from the material library of COMSOL Multiphysics.

The $5TOLCl_0$ liquid crystal is taken as the active material due to the high value of birefringence ~0.2828 [45]. The ordinary and the extraordinary principal axis have refractive indices of $n_o = 1.5964$ and $n_e = 1.8792$ respectively. Switching time is greatly reduce by the higher birefringence value as it reduces the gap. In the influence of the electric field, the molecules of the nematic liquid crystal get aligned along the field. The degree of alignment of the liquid crystal molecules depends on the applied electric field, the stronger the field the stronger will be the alignment of the molecule.

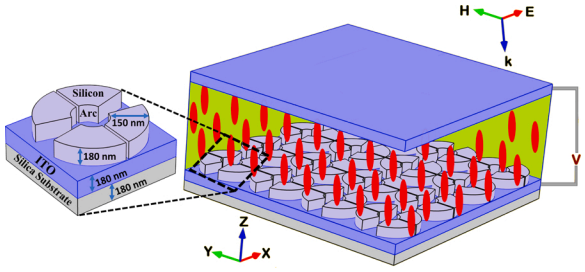


Fig. 1. The schematic representation of the proposed all-dielectric metasurface based electro-optic modulator. Asymmetric silicon arcs (lavender colour) are mounted over the silica substrate (grey colour) having ITO layer (violet colour) spread on it. Nematic liquid crystal molecules (red colour) are enclosed between the two ITO layers. Zoom-in shows the asymmetric silicon arcs over the substrate.

The molecules in the liquid crystal are rod-like which are aligned arbitrarily. In the case when there is no electric field across the ITO layers and the system is considered to be in null condition. As the electric potential applied across the ITO layer from zero volts (low state) to maximum value (high state) the molecules start getting dipolar in nature due to the potential created between the two ITO (positive and negative) layers. Consequently, the dipolar molecules getting oriented along the direction of the applied electric field. Thus, the applied electric field actively control the dielectric environment around the silicon arcs as liquid crystal filling the air gap between the silicon arcs. Due to the change in the dielectric property the resonance characteristic shows the variation. Conversely, with the variation in the applied biased electric field across the liquid crystal the transmission values vary for the particular wavelength. The molecule size of the liquid crystal is very small compared to the applied wavelength of the light so the liquid crystal can be treated as an effective medium with the average refractive index n_a and the effective refractive index n_{eff} of the liquid crystal can be calculated with help of the macroscopic order parameter q [18] as

$$n_{eff} = n_a - \frac{1}{3}\Delta n q \tag{1}$$

where, n_a is given by

$$n_a = \frac{(2n_o + n_e)}{3} \tag{2}$$

and birefringence is given by

$$\Delta n = n_e - n_o \tag{3}$$

The macroscopic order parameter ‘q’ varies from low state ‘0’ to the high state ‘1’. When no biasing across the ITO plates, the parameter q value is 0, i.e., the liquid crystal molecules are arbitrary aligned. When the biasing potential varies from the low state to high state, the parameter q value increases from 0 to 1 and the arbitrarily aligned molecules start to align in a particular fashion. Due to this, the effective refractive index starts to decrease as defined by the Eq. (1). When the biasing potential reaches the highest value, the parameter q value become 1 and the liquid crystal molecules become strongly aligned in a uniform fashion along the direction of the electric field. Consequently, the effective refractive index reaches its lowest value n_o.

3. Results and discussions

When the x-polarized electromagnetic field interacts with the liquid crystal filled the metasurface, profile of the response has a resonance wavelength λ_0 at 1552 nm, which lies in the commercial communication band as shown in Fig. 2. The Q-Factor is the ratio of the resonance wavelength and the linewidth of the resonance peak. The linewidth, $\Delta\lambda$ (Full Width at Half Maxima, FWHM) of the resonance is defined by the difference between the peak and the dip wavelengths [21]. The linewidth of the observed resonance is the 12 nm while the Q-Factor is calculated to be 129.33. The modulation depth is the ratio of the difference between the transmission values at the peak and dip to the transmission value at the peak [46]. The modulation depth for the resonance wavelength at 1552 nm is 99.33 % which is better than many reported metamaterial designs [28,40,41] as per our best knowledge. The numerical study is done by considering parameter q value as ‘0.5’. The response of the metasurface in absence and presence of liquid crystal is compared as shown in Fig. 2. The resonance in absence of liquid crystal is at 1547 nm wavelength.

The effects of the variation of parameters of liquid crystal filled all-dielectric metasurface structure on the resonance are presented in Fig. 3(a)–(d). The 116 nm of redshift in the resonance position is observed for the change in the external radius of Si arcs from 400 nm to

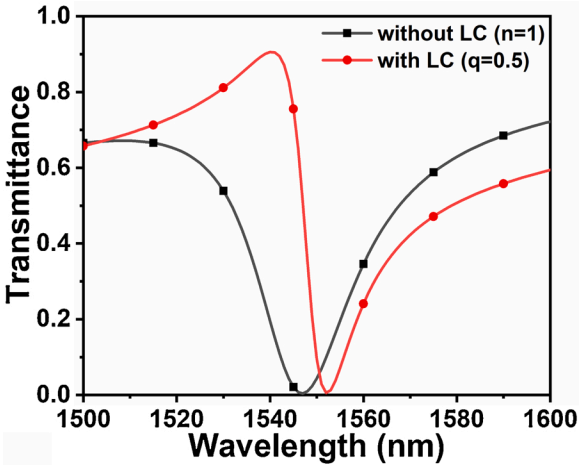


Fig. 2. Shows the response of the metasurface in absence of liquid crystal have resonance at the 1547 nm (black curve) and the response of the metasurface filled with liquid crystal have resonance at the 1552 nm (red curve) calculated using COMSOL Multiphysics.

505 nm and for this change in external radius of Si arcs, the linewidth of the resonances is nearly constant for large radius value but decrease for lower radius values as shown in Fig. 3(a). The blue shift of 21 nm in the resonance position is observed for the change in the internal radius of Si arcs from 125 nm to 185 nm. The linewidth of the resonance is observed to vary from 8 nm to 18 nm due to the change in the internal radius of Si arcs as shown in Fig. 3(b). When the split gap of Si arcs are varied from 1 nm to 50 nm, the blue shift of 33 nm in the resonance position and 2 nm of variation in the linewidth of the resonance is observed as shown in Fig. 3(c). The redshift of 78 nm in the resonance position and variation 3 nm in the linewidth of the resonance is observed due to the change in the thickness of Si arcs as shown in Fig. 3(d).

In the subsequent study, the parameter q value varies from '0' to '1' to observe the response of the increase in the bias potential field on the liquid crystal cell. Fig. 4(a) represents the effect of varying the parameter q value, on the resonance position. When parameter q is '0' the resonance peak is observed at the 1570 nm and the resonance shows the blueshift in peak wavelength when parameter q value is varied from '0' to '1'. The resonance wavelength shifts to 1535 nm when the parameter q value is set to '1'. This blueshift in the resonance is due to the decrease in the effective refractive index of the liquid crystal experienced by the all-dielectric metasurface.

As the value of parameter q varies the resonance peak wavelength shifts in position, this shift is linear with respect to the parameter q as plotted in Fig. 4(b). The total shift of 35 nm is observed. The resonance position shift is larger than the other metamaterial-based modulators [18,44,47].

We observed that the linewidth of the Fano resonance curve, which is 12 nm, remains constant and has no effect of the parameter q variation. The Q-Factor for the resonance peaks at the different parameter q values are nearly constant, and the maximum observed Q-Factor value is of 130.75 at the resonance wavelength of 1570 nm, the trends of linewidth and Q-Factor can be seen in Fig. 4(c).

The calculated values of modulation depth remain fairly constant and close to 100 %. The modulation depth varies from 99.14 % for parameter q value of '1' to 99.30 % for parameter q value of '0' as shown in Fig. 4 (c). As per our knowledge, this is the best modulation depth values among all the reported metasurface based structures [18,28,40,41,43], which are ranging from 30 % to 85 %. At wavelength 1550 nm the calculated modulation depth of 99 % is achieved.

To observe the tunability aspect of the proposed all-dielectric metasurface based optical modulator, we have investigated the dependence of the transmission by varying the q parameter keeping the wavelengths constant as shown in Fig. 5. It is clear from the figure that a broadband tunability more than 25 nm can be achieved by considering the variation of the entire range of q parameter (from 0 to 1). For the wavelength of 1550 nm, the response shows the highest transmission value at the parameter q value of 0.2 and the lowest transmission at the parameter q value of 0.55 as depicted in Fig. 5. The transmission decreases unvaryingly between parameter q values of 0.2 and 0.55 with a steep slope. It is clear from the Fig. 5 that one can achieve the transmission from max to min by changing the parameter q from 0.2 to 0.55 which implies the

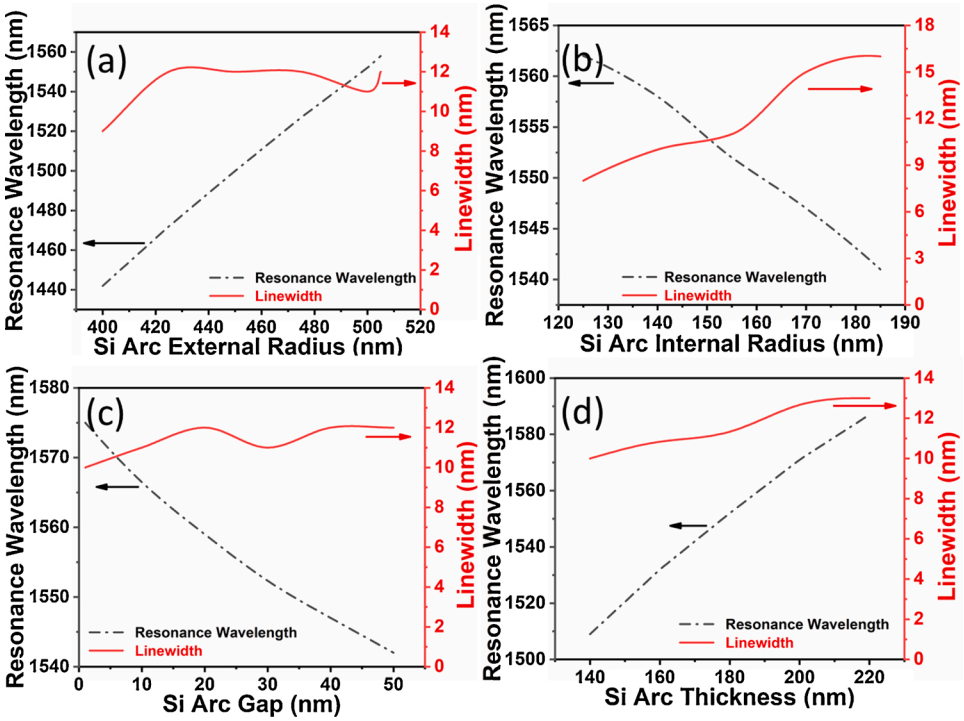


Fig. 3. Shows the effect of parameters variation of the proposed structure on the resonance wavelength (black dash-dot curve) and linewidth (red smooth curve). (a) The redshift in the resonance position and variation in the linewidth of the resonance due to the change in the external radius of Si arcs, (b) The blue shift in the resonance position and variation in the linewidth of the resonance due to the change in the internal radius of Si arcs, (c) The blue shift in the resonance position and variation in the linewidth of the resonance due to the change in the split gap of Si arcs, (d) The redshift in the resonance position and variation in the linewidth of the resonance due to the change in the thickness of Si arcs.

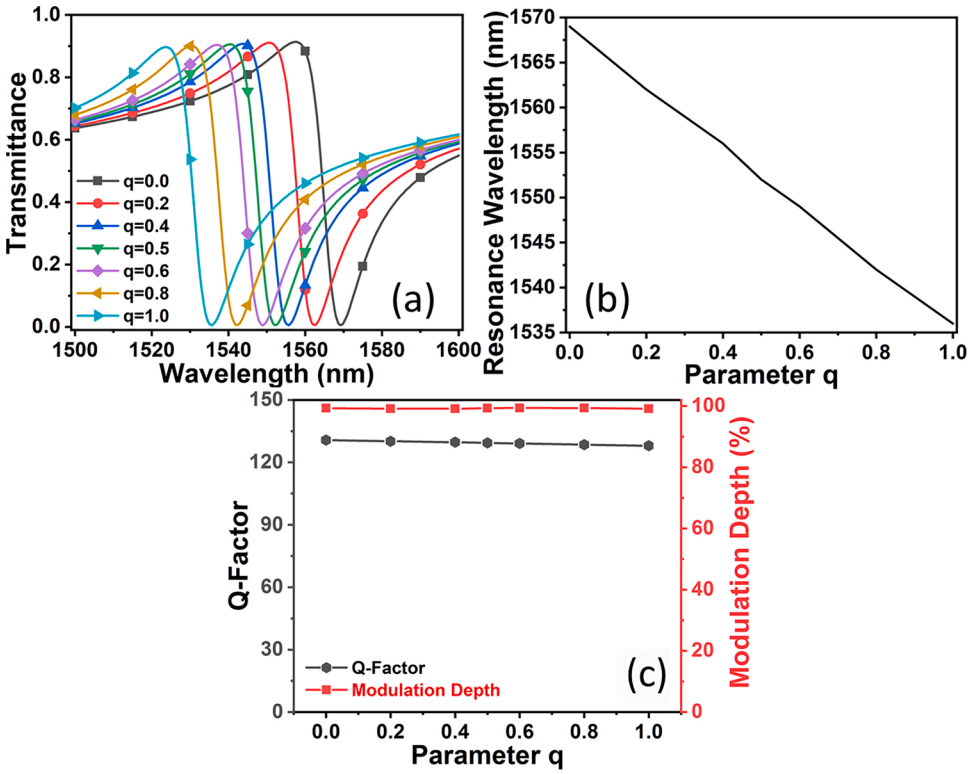


Fig. 4. (a) The blue shift in the resonance position with the change in the parameter q value. The total shift of 33 nm is observed, (b) The blue shift in resonance wavelength position due to change in the parameter q value is a linear function, (c) The characteristic response of the proposed modulator design showing Q-Factor (black) and modulation depth (red).

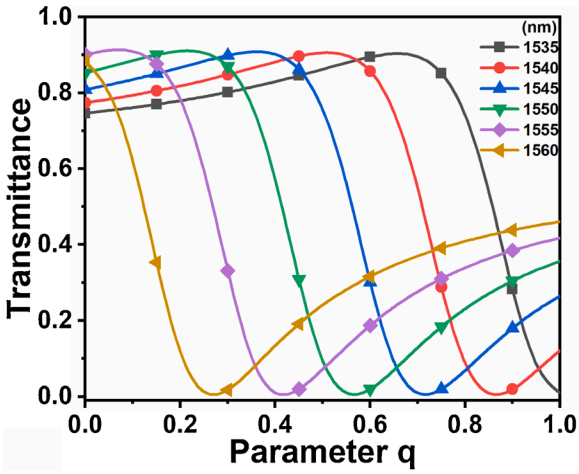


Fig. 5. The transmission amplitude profile at the 1550 nm wavelength for parameter q value of '0' to '1'. The highest transmission amplitude is at 0.2 which decrease steeply to zero at 0.55.

requirement of less voltage difference to achieve the desired modulation. The modulation depth higher than 90 % and more can be observed for the resonance wavelengths in the range between 1535 nm–1560 nm. This wavelength band of 25 nm gives the flexibility of tuning in the desired resonance position.

The transmittance profile for different polarizations of incident field is shown in Fig. 6. When polarization angle of the incident field is varied from 0° (x-polarization) to 90° (y-polarization), the linewidth of the resonance increases from 12 nm to 19 nm, respectively, and a red-shift of 8 nm in resonance position is observed as shown in Fig. 6(a). Due to this increase in the linewidth, the Q-Factor of the resonance decreases from 130 to 82 (x- to y-polarization) as shown in Fig. 6(b). We also observe that when the polarization of incident light changes its effect on the modulator bandwidth is not prominent as bandwidth remains constant that is 25 nm for all angle of polarization.

The metasurface structure of asymmetric silicon arc can be fabricated using the plasma-enhanced chemical vapour deposition (PECVD) method on the silica substrate having electrode terminals layer of indium tin oxide layer as done previously [16,18]. Fabrication of the nanophononics structures through the available techniques may introduce deformities in the structure in terms of the dimensions. We performed a numerical study to analyse the effect of the fabrication error on Fano resonance. For this analysis, we have considered fabrication error embedded in the structural parameter the thickness and the width of the silicon arcs and the split gap in the arcs. In this analysis, the fabrication error in the silicon arcs thickness and width is taken as 180 ± 6 nm and 345 ± 10 nm, respectively, and the fabrication error in the split gap of the silicon arcs is taken to be 30 ± 2 nm. We observe that resonance peak wavelength shows a blue shift of 35 nm for mean parameter values with variation in the parameter q value. When the deformity error is introduced in the arc thickness, the resonance peak wavelength shift ± 6 nm about the mean value for the constant parameter q value as shown in Fig. 7(a). Similarly, when fabrication deformity is introduced in the split gap resonance peak wavelength shift ± 1 nm from the mean position as shown in Fig. 7(c). The equal shift in either side of resonance peak wavelength during the error analysis is observed and this is due to the fact that a varying deformity introduces a linear change in the volume of the metasurface (filling factor) results in a linear change in the effective parameters experienced by the incident field as per the effective medium theory [48]. When the response due to fabrication error in the arc width is introduced, the resonance peak wavelength has a total shift of 14 nm ($+4$ nm and -10 nm) as shown in Fig. 7(b). The uneven shift in the resonance peak is due to the non-linear change in the volume of the metasurface and hence the effective parameters as per the effective

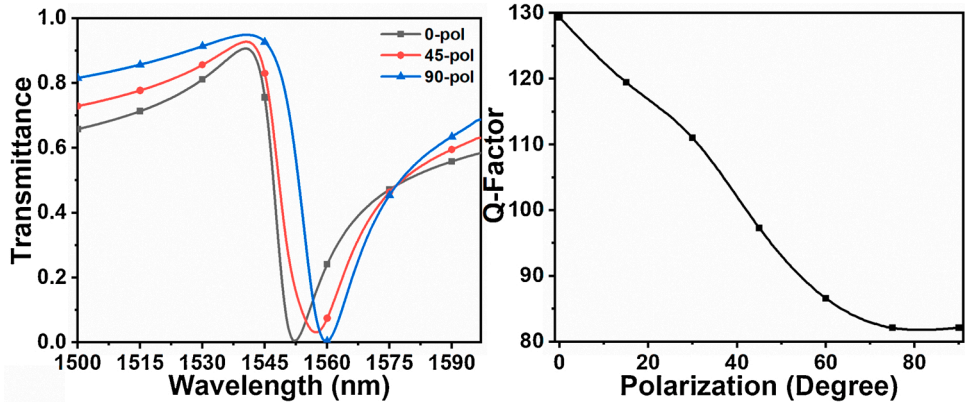


Fig. 6. The transmission profile for the incident light polarization at 0°, 45°and 90° under the increased parameter q from 0 to 1 and effect on Q-Factor of resonance.

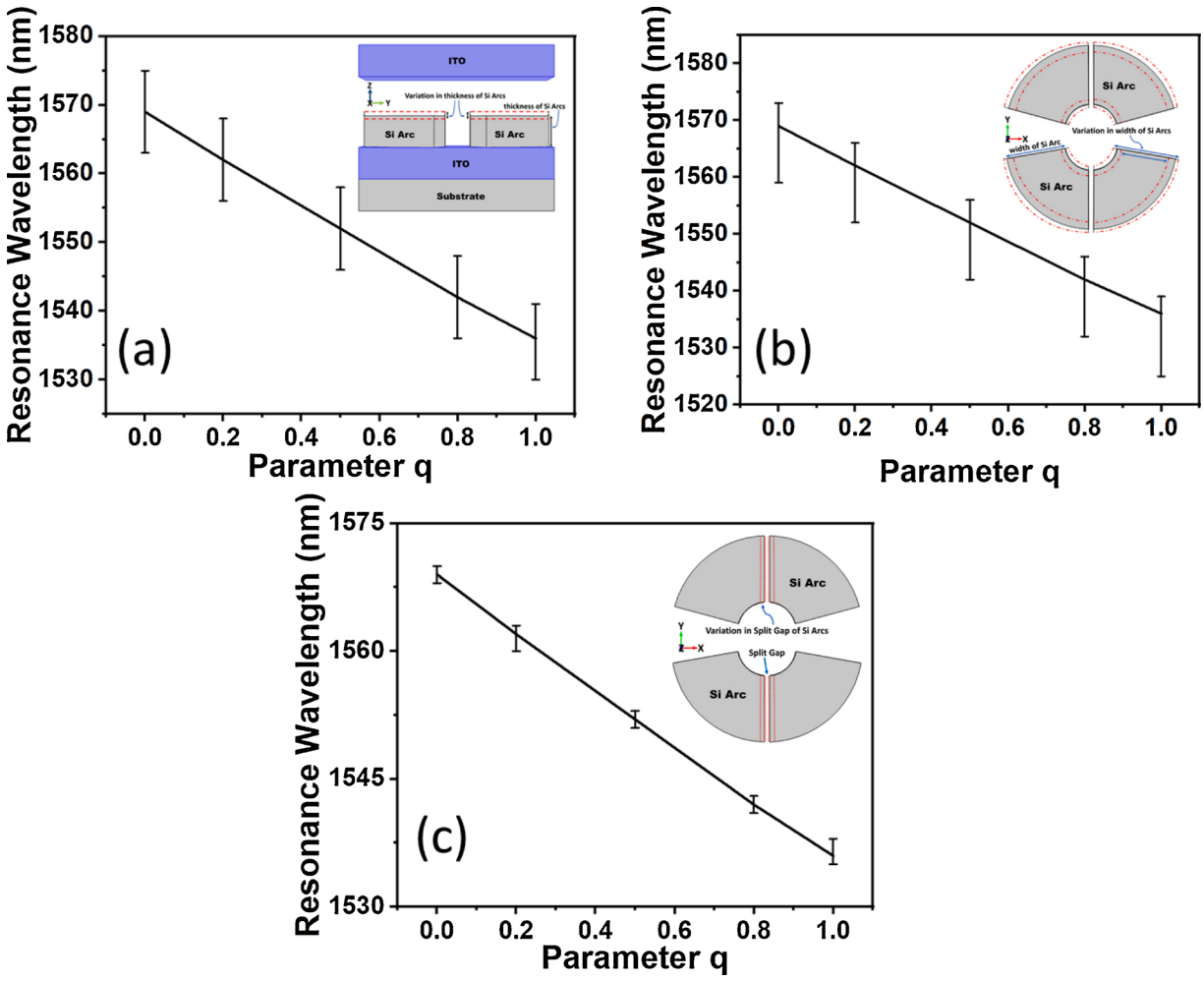


Fig. 7. The blue shift in the resonance position with the change in the parameter q value is a linear function. (a) Error bar shows the shift in resonance due to the fabrication error in the thickness of the silicon arc, (b) Error bar shows the shift in resonance due to the fabrication error in the width of the silicon arc, (c) Error bar shows the shift in resonance due to the fabrication error in the split gap of the silicon arc. The schematic view of the investigated parameter is presented in the inset of each figure.

medium theory.

The modulation speed of the proposed device is in order of a millisecond [49] and can be further increased to microsecond by doping nanoparticles in the liquid crystal or ferroelectric liquid crystal [49,50]. Response time depends on the LC layer, the viscosity of the material, temperature, and applied biasing voltage. LC based electro-optic devices have promising applications in the design and development of tunable LC filters [51], LC display [52] and LC switches [53].

4. Conclusions

We conclude that the proposed design of all-dielectric metasurface with high birefringence nematic liquid crystal may be utilized to design a broadband tunable optical modulator. The variation in the orientation

of the crystal molecules enables the variation in the effective refractive index of the metasurface, which results in the blueshift in the Fano resonance. The numerical simulations show that the proposed modulator has a broadband tunability of 25 nm and modulation depth of 90 %. The Q-Factor of resonance decrease when the polarization of incident field varies from x- to y-polarization. The numerical study that relates the modulator performance to the fabrication errors is performed. The study shows that the non-uniformities in the thickness and width of silicon arc are dominating factor while the effect of the non-uniformities in the split gap of the silicon arc has a relatively small effect on the shift of resonant wavelength.

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.

Acknowledgements

The authors gratefully acknowledge the support from the ICMR for funding from grant number 5/3/8/62/ITR-F/2019-ITR ICMR, Department of Science and Technology, India for a project titled “All-Dielectric, plasmonic and hybrid photonic nanostructure” and CSIR mission project 4M titled “Photonic meta-surfaces for smart applications”. Authors are also thankful to CSIR-CSIO, Chandigarh, India for providing research facilities. Authors also acknowledge Dr Mukesh Sharma, Tel-Aviv University, Israel for the fruitful discussion.

References

[1] R. Singh, W. Cao, I. Al-Naib, L. Cong, W. Withayachumnankul, W. Zhang, Ultrasensitive terahertz sensing with high-Q Fano resonances in metasurfaces, Appl. Phys. Lett. 105 (2014) 171101.
[2] M. Sharma, T. Ellenbogen, An All-Optically Controlled Liquid-Crystal Plasmonic Metasurface Platform, Laser & Photonics, Reviews 14 (2020) 2000253.
[3] B. Luk’yanchuk, N.I. Zheludev, S.A. Maier, N.J. Halas, P. Nordlander, H. Giessen, C. T. Chong, The Fano resonance in plasmonic nanostructures and metamaterials, Nat. Mater. 9 (2010) 707.
[4] Y. Ma, H. Zhang, C. Hu, Tunable omnidirectional band gap and polarization splitting in one-dimensional magnetized plasma photonic crystals with a quasi-periodic topological structure, J. Opt. 22 (2020) 025101.
[5] Y. Ma, H. Zhang, T. Liu, W. Li, Study on the properties of unidirectional absorption and polarization splitting in one-dimensional plasma photonic crystals with ultra-wideband, J. Opt. Soc. Am. B 36 (2019) 2250.
[6] Y.-P. Li, H.-F. Zhang, T. Yang, T.-Y. Sun, L. Zeng, A multifunctional polarization converter base on the solid-state plasma metasurface, IEEE J. Quantum Electron. 56 (2020) 2975019.
[7] B. Slovick, Z.G. Yu, M. Berding, S. Krishnamurthy, Perfect dielectric-metamaterial reflector, Phys. Rev. B 88 (2013) 165116.
[8] S. Jahani, Z. Jacob, All-dielectric metamaterials, Nat. Nanotechnol. 11 (2016) 23.
[9] A. García-Etxarri, R. Gómez-Medina, L.S. Froufe-Pérez, C. López, L. Chantada, F. Scheffold, J. Aizpurua, M. Nieto-Vesperinas, J.J. Sáenz, Strong magnetic response of submicron silicon particles in the infrared, Opt. Express 19 (2011) 4815–4826.
[10] F.J. Bezares, J.P. Long, O.J. Glembocki, J. Guo, R.W. Rendell, R. Kasica, L. Shirey, J.C. Owrutsky, J.D. Caldwell, Mie resonance-enhanced light absorption in periodic silicon nanopillar arrays, Opt. Express 21 (2013) 27587–27601.
[11] W.X. Lim, R. Singh, Universal behaviour of high-Q Fano resonances in metamaterials: terahertz to near-infrared regime, Nano Converg. 5 (2018) 5.
[12] J. Van de Groep, A. Polman, Designing dielectric resonators on substrates: combining magnetic and electric resonances, Opt. Express 21 (2013) 26285–26302.
[13] A.B. Evlyukhin, S.M. Novikov, U. Zywietz, R.L. Eriksen, C. Reinhardt, S. I. Bozhevolnyi, B.N. Chichkov, Demonstration of magnetic dipole resonances of dielectric nanospheres in the visible region, Nano Lett. 12 (2012) 3749–3755.
[14] S. Hayashi, D. Nesterenko, A. Rahmouni, H. Ishitobi, Y. Inouye, S. Kawata, Z. Sekkat, Light-tunable Fano resonance in metal-dielectric multilayer structures, Sci. Rep. 6 (2016) 33144.
[15] M.F. Limonov, M.V. Rybin, A.N. Poddubny, Y.S. Kivshar, Fano resonances in photonics, Nat. Photonics 11 (2017) 543.
[16] W. Zhao, X. Leng, Y. Jiang, Fano resonance in all-dielectric binary nanodisk array realizing optical filter with efficient linewidth tuning, Opt. Express 23 (2015) 6858–6866.
[17] J. Zhang, W. Liu, Z. Zhu, X. Yuan, S. Qin, Strong field enhancement and light-matter interactions with all-dielectric metamaterials based on split bar resonators, Opt. Express 22 (2014) 30889–30898.
[18] W. Zhao, H. Jiang, B. Liu, Y. Jiang, C. Tang, J. Li, Fano resonance based optical modulator reaching 85% modulation depth, Appl. Phys. Lett. 107 (2015) 171109.
[19] Y. Yang, I.I. Kravchenko, D.P. Briggs, J. Valentine, High Quality Factor Fano-resonant All-dielectric Metamaterials, arXiv preprint arXiv:1405.3901, 2014.
[20] A. Krasnok, S. Makarov, M. Petrov, R. Savelev, P. Belov, Y. Kivshar, Towards all-dielectric metamaterials and nanophotonics, Metamaterials X, Int. Soc. Optics Photonics (2015) 950203.
[21] Y. Zhang, W. Liu, Z. Li, Z. Li, H. Cheng, S. Chen, J. Tian, High-quality-factor multiple Fano resonances for refractive index sensing, Opt. Lett. 43 (2018) 1842–1845.
[22] G.-D. Liu, X. Zhai, L.-L. Wang, Q. Lin, S.-X. Xia, X. Luo, C.-J. Zhao, A high-performance refractive index sensor based on fano resonance in si split-ring metasurface, Plasmonics 13 (2018) 15–19.
[23] K.S. Modi, J. Kaur, S.P. Singh, U. Tiwari, R.K. Sinha, Split-arc-based metasurface for refractive index sensing applications, High Contrast Metastructures VIII, Int. Soc. Opt. Photonics 10928 (2019) 109281V.
[24] K.S. Modi, J. Kaur, S.P. Singh, U. Tiwari, R.K. Sinha, Extremely high figure of merit in all-dielectric split asymmetric arc metasurface for refractive index sensing, Opt. Commun. (2020) 125327.
[25] K.S. Modi, J. Kaur, S.P. Singh, U. Tiwari, R.K. Sinha, Fano-resonance Based Tunable All-dielectric Metasurfaces. High Contrast Metastructures IX, International Society for Optics and Photonics, 2020, p. 112901F.
[26] N. Shankhwar, Y. Kalra, R.K. Sinha, All dielectric zero-index metamaterial for TE/TM polarization, J. Opt. 20 (2018) 115101.
[27] N. Shankhwar, Y. Kalra, Q. Li, R.K. Sinha, Zero-index metamaterial based all-dielectric nanoantenna, AIP Adv. 9 (2019) 035115.
[28] A. Komar, Z. Fang, J. Bohn, J. Sautter, M. Decker, A. Miroshnichenko, T. Pertsch, I. Brener, Y.S. Kivshar, I. Staude, Electrically tunable all-dielectric optical metasurfaces based on liquid crystals, Appl. Phys. Lett. 110 (2017) 071109.
[29] J. Sautter, I. Staude, M. Decker, E. Rusak, D.N. Neshev, I. Brener, Y.S. Kivshar, Active tuning of all-dielectric metasurfaces, ACS Nano 9 (2015) 4308–4315.
[30] A. Korpel, Acousto-optics—a review of fundamentals, Proceedings of the IEEE 69 (1981) 48–53.
[31] K. Wijewardena Gamalath, G. Jayawardena, Diffraction of light by acoustic waves in liquids, International Letters of Chemistry, Phys. Astronomy 4 (2012) 39–57.
[32] G. Sinkevicius, A. Baskys, Investigation of piezoelectric ringing frequency response of Beta barium borate crystals, Crystals 9 (2019) 49.
[33] D.-e. Zhao, Y.-h. Chen, Z.-b. Wang, Y.-y. Chen, L.-f. Wang, R. Zhang, Study on signal crystal photo-elastic modulator based on lithium niobate piezoelectric and photo-elastic effect, 2012, Symposium on Piezoelectricity, Acoustic Waves, and Device Applications (SPAWDA), IEEE (2012) 223–228.
[34] K. Lau, K. Wong, T. Chan, S. Yeung, An economical piezoelectric phase modulator for fiber optic sensors, Appl. Opt. 35 (1996) 6836–6838.
[35] M. Muriel, J.A. Martin-Pereda, Liquid-crystal electro-optic modulator based on electrohydrodynamic effects, Opt. Lett. 5 (1980) 494–495.
[36] Y.C. Su, C.C. Chiou, V. Marinova, S. Lin, H. Dikov, P. Vitanov, K. Hsu, Liquid crystal electro-optic modulator based on transparent conductive TiO 2/Ag/TiO 2 multilayers, Opt. Quantum Electron. 50 (2018) 242.
[37] L. Liu, I.V. Shadrivov, D.A. Powell, M.R. Raihan, H.T. Hattori, M. Decker, E. Mironov, D.N. Neshev, Temperature control of terahertz metamaterials with liquid crystals, IEEE Trans. Terahertz Sci. Technol. 3 (2013) 827–831.
[38] C. Care, D. Cleaver, Computer simulation of liquid crystals, Rep. Prog. Phys. 68 (2005) 2665.
[39] D.C. Zografopoulos, R. Asquini, E.E. Kriezis, A. d’Alessandro, R. Beccherelli, Guided-wave liquid-crystal photonics, Lab Chip 12 (2012) 3598–3610.
[40] A. Minovich, J. Farnell, D.N. Neshev, I. McKerracher, F. Karouta, J. Tian, D. A. Powell, I.V. Shadrivov, H. Hoe Tan, C. Jagadish, Liquid crystal based nonlinear fishnet metamaterials, Appl. Phys. Lett. 100 (2012) 121113.
[41] O. Buchnev, J. Ou, M. Kaczmarek, N. Zheludev, V. Fedotov, Electro-optical control in a plasmonic metamaterial hybridised with a liquid-crystal cell, Opt. Express 21 (2013) 1633–1638.
[42] A.A. Komar, D.N. Neshev, A.E. Miroshnichenko, Tunable dielectric metasurfaces based on the variation of the refractive index of the environment, JETP Lett. 106 (2017) 709–715.
[43] H. Su, H. Wang, H. Zhao, T. Xue, J. Zhang, Liquid-crystal-based electrically tuned electromagnetically induced transparency metasurface switch, Sci. Rep. 7 (2017) 17378.
[44] S. Xiao, U.K. Chettiar, A.V. Kildishev, V. Drachev, I. Khoo, V.M. Shalaev, Tunable magnetic response of metamaterials, Appl. Phys. Lett. 95 (2009) 033115.
[45] A. Spadlo, R. Dabrowski, J. Dziaduszek, E. Scibior, S. Urban, S. Gauza, S. Wu, Liquid crystalline materials with high birefringence, J. Opt. Technol. 72 (2005) 659–661.
[46] L. Zhang, Z. Dong, Y.M. Wang, Y.J. Liu, S. Zhang, J.K.W. Yang, C.-W. Qiu, Dynamically configurable hybridization of plasmon modes in nanoring dimer arrays, Nanoscale 7 (2015) 12018–12022.
[47] M. Decker, C. Kremers, A. Minovich, I. Staude, A.E. Miroshnichenko, D. Chigrin, D. N. Neshev, C. Jagadish, Y.S. Kivshar, Electro-optical switching by liquid-crystal controlled metasurfaces, Opt. Express 21 (2013) 8879–8885.
[48] W. Cai, V. Shalaev, D.K. Paul, Optical metamaterials: fundamentals and applications, Phys. Today 63 (2010) 57.
[49] B.E. Saleh, M.C. Teich, Fundamentals of Photonics, John Wiley & Sons, 2019.
[50] M. Sharma, A. Sinha, M. Shenoy, Effect of TiO2 nanoparticle doping on the performance of electrically-controlled nematic liquid crystal core waveguide switch, Opt. Mater. 49 (2015) 292–296.

[51] M. Desouky, M. Abdelsalam, M.A. Swillam, Tunable Dielectric Metasurface upon Using Nematic Liquid Crystals for Color Filter Application, 2019, p. 109281T.
[52] Y.-S. Lin, J. Dai, Z. Zeng, B.-R. Yang, Metasurface color filters using aluminum and Lithium niobate configurations, Nanoscale Res. Lett. 15 (2020) 77.
[53] H. Su, H. Wang, H. Zhao, T. Xue, J. Zhang, Liquid-crystal-based electrically tuned electromagnetically induced transparency metasurface switch, Sci. Rep. 7 (2017) 17378.

Silicon
https://doi.org/10.1007/s12633-020-00910-7

ORIGINAL PAPER

Analog and RF Performance Evaluation of Junctionless Accumulation Mode (JAM) Gate Stack Gate All Around (GS-GAA) FinFET

Bhavya Kumar[1] · Rishu Chaujar[1]

Received: 28 June 2020 / Accepted: 16 December 2020
© Springer Nature B.V. 2021

Abstract
This work presents the analog and RF performance evaluation of Junctionless Accumulation Mode (JAM) Gate Stack Gate All Around (GS-GAA) FinFET, and the results acquired have been compared with conventional FinFET and GAA FinFET. It has been observed that in comparison to conventional FinFET, leakage current (I_{off}) reduces by almost thirty times for the GS-GAA FinFET configuration. Thus, revamping the threshold voltage (V_{th}), switching ratio (I_{on}/I_{off}), and subthreshold slope (SS) of the proposed device. Also, major analog parameters like transconductance (g_m), transconductance generation factor (TGF) enhances considerably with early voltage (V_{EA}) and intrinsic gain (A_v) increased by over two times in magnitude for the GS-GAA FinFET configuration. Furthermore, several important RF parameters have been explored, and the outcome of the study is that the GS-GAA FinFET configuration shows superior RF performance. In GS-GAA FinFET configuration, the gain frequency product (GFP) and gain transconductance frequency product (GTFP) amplified by over two times in magnitude with minimal decrease in the cut-off frequency (f_T) and maximum oscillation frequency (f_{max}). Thus, the proposed GS-GAA FinFET device can be looked upon as an appealing option for high-frequency analog/RF applications.

Keywords Analog/RF parameters · Gate Stack (GS) · Gate All Around (GAA) FinFET · Junctionless Accumulation Mode (JAM) · Short Channel Effects (SCEs) · TiN metal gate

1 Introduction

According to Gordon E. Moore, the number of transistors in a given unit of space will double up almost every eighteen months [1]. Thus, to follow up with Moore's Law, CMOS technology is being aggressively scaled-down and is reaching its fundamental limits. As the channel length of the device decreases gradually, the device starts suffering from issues like pin-point ultra-sharp S/D junction's formation, and short-channel effects (SCEs) such as drain induced barrier lowering, mobility degradation [2–4]. Colinge et al. resolved the problem in the form of junctionless field-effect transistors (JLFETs) [5]. In JL transistors, doping species (either n-type or p-type) is the same in all the regions. The junctionless device, in comparison to conventional CMOS devices, has better immunity against the SCEs with effectively increased channel length. However, the difficulty arises in the fabrication of a narrow and thin channel layer, which is essential to completely deplete the channel during the OFF state to attain excellent turn-off characteristics [6]. If the doping concentration of the channel region is relatively low, then carriers will quickly deplete in the OFF state; however, it leads to increased S/D series resistance accompanied by a decrease in drain current. Therefore, to get over these problems, junctionless accumulation mode (JAM) FETs with increased S/D doping is used [7, 8]. In 1999, professor Chenming Hu proposed a "3D" transistor named FinFET to alleviate the SCEs [9]. Several research papers have already been published on FinFET and taking into consideration that as the number of gates increases, electrostatic control over the channel increases, GAA structure has been explored for better device performance and improved SCEs [10]. As per the International Technology Roadmap for Semiconductors (ITRS) guideline, GAA is also known as "ultimate structure" [11].

✉ Rishu Chaujar
chaujar.rishu@dtu.ac.in

Bhavya Kumar
bhavyakumar_phd2k18@dtu.ac.in

[1] Department of Applied Physics, Delhi Technological University, Delhi, India

Published online: 03 January 2021

Another primary concern is the increment in the gate direct tunneling current with the decreasing gate oxide thickness, resulting in the enhanced off-state leakage current [12, 13]. The use of high-κ dielectric materials, like HfO_2 (κ = 25) [14], proved to be the most technological solution. However, the direct deposition of high-κ dielectrics on Si substrate remains a challenging field. The challenges that need to be overcome are mobility degradation accompanied by below-par subthreshold swings, threshold voltage instability, and maintaining the interface quality between the high-κ dielectric and Si substrate [15, 16]. An alternative is to use Gate Stack (GS) architecture comprising of a thin interfacial SiO_2 layer as passivation between the Si substrate and high-κ dielectrics by maintaining the effective oxide thickness (EOT) constant [17]. Thus, keeping in mind all the considerations, we are proposing a JAM GS-GAA FinFET. Nowadays, metal gates are preferred over polysilicon gates because polysilicon shows a poly-depletion effect, thereby affecting the EOT of the device. Also, polysilicon gates, when brought in contact with high-κ dielectrics, are chemically unstable [18]. TiN is picked as the metal gate due to its thermal stability, high purity, low resistivity, and compatibility with CMOS processing [19, 20].

In the last few years, demands for portable and higher battery backup electronic devices have sufficiently increased. To maintain the increased demand, speed performance, and transistors density in the IC needs to go up, and that can be obtained by System-on-Chip (SoC) concept. In SoC technology, the majority of the elements of a system are integrated as an independent system on a semiconductor chip. The device optimization is very challenging with SoC technology [21]. Thus, the purpose of this study is to analyze the device structure and improve the device performance for high-frequency analog/RF circuit applications. Analog and RF performance of JLAM-CGAA MOSFET has been investigated extensively [22]. In another study, Raskin et al. reported the analog/RF performance of multiple gate SOI devices [23]. However, to the best knowledge of the authors, no report has been published on the analog and RF performance of JAM GS-GAA FinFET. In this work, extensive simulations have been performed to investigate the several important electrostatic, analog, and RF parameters of the three different devices, namely, conventional FinFET, GAA FinFET, GS-GAA FinFET. Besides, we have also evaluated various other RF metrics significant for RF applications such as f_{max}, GFP, TFP, and GTFP for RF performance.

It is observed from the simulated results that JAM GS-GAA FinFETs provide several advantages like higher switching ratio, improved subthreshold slope, etc. with minimal decrease in the f_T and f_{max} in comparison to the other two devices.

Starting with the introduction, section 2 narrates the device structure. Section 3 deals with the simulation methodology used and calibration between the simulation data and experimental data. Section 4 analyses the analog and RF performance parameters of the device with concluding remarks in section 5, authenticating the originality of the paper.

2 Device Structure

Figure 1 displays the proposed 3D structure of JAM GS-GAA FinFET, whereas horizontal and vertical cross-sectional view cut along the silicon fin of the proposed device is portrayed in Fig. 2 (a) and (b), respectively. The gate length (L_g) of the device is 7 nm, and the length of S/D regions ($L_{S/D}$) is fixed at 10 nm. The oxide thickness (t_{ox}) is kept constant all around the fin at 1 nm, as depicted in Table 1. In GS-GAA FinFET, SiO_2 and HfO_2 are amalgamated in equal proportion, whereas in the other two structures, only SiO_2 is used as a gate oxide. Silicon material is used in the fin structure with height (H_{Fin}) and width (W_{Fin}) of the fin fixed at 10 nm and 5 nm, respectively, thereby following the property of width quantization, which says that W_{Fin} must be a multiple of H_{Fin} [24, 25]. All the regions are uniformly doped with n-type doping species. The doping concentration in the source/drain regions ($N_{S/D}$) is 10^{19} cm^{-3}, and that of the channel region (N_{Ch}) is 10^{16} cm^{-3}. The work function (φ_m) is kept at 4.65 eV for the TiN metal gate [20].

3 Simulation Methodology and Calibration

The SILVACO ATLAS 3D simulator has simulated all the three different configurations. In the entire device simulation, V_{ds} is fixed at 0.5 V, V_{gs} is changed from 0 V to 1 V, and the T is at 300 K. In amplifier circuits, the region of operation is decided by gate overdrive voltage (V_{gt}). As a result, all the analog and RF analysis has been done against V_{gt}. It is the voltage difference between gate to source voltage and threshold voltage, i.e., $V_{gt} = V_{gs} - V_{th}$. Several physical models are included in device simulation for different purposes like the Arora model is incorporated to grab concentration-dependent

Table 1 Different device parameters used for simulation

Parameter	FinFET	GAA FinFET	GS-GAA FinFET
L_g (nm)	7	7	7
$L_{S/D}$ (nm)	10	10	10
t_{ox} (nm)	1 (SiO_2)	1 (SiO_2)	1 (0.5 SiO_2 + 0.5 HfO_2)
H_{Fin} (nm)	10	10	10
W_{Fin} (nm)	5	5	5
$N_{S/D}$ (cm^{-3})	1×10^{19}	1×10^{19}	1×10^{19}
N_{Ch} (cm^{-3})	1×10^{16}	1×10^{16}	1×10^{16}
φ_m (eV)	4.65 (TiN)	4.65 (TiN)	4.65 (TiN)

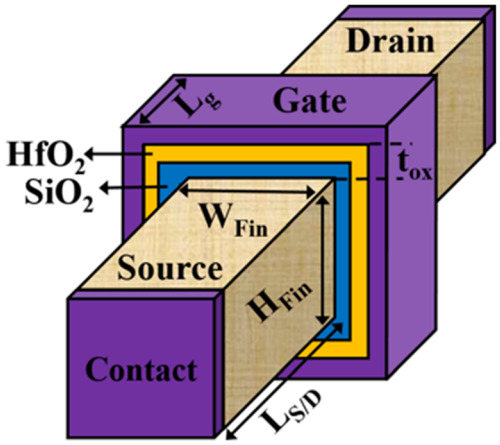


Fig. 1 Systematic 3-D structure of the proposed JAM GS-GAA FinFET

mobility. For recombination effects, Shockley-Read-Hall (SRH) model is employed to imitate the existence of trap charges at the interface. The Fermi-Dirac statistics model is used to increase the precision in the results. Crowell impact ionization and Klaassen band to band tunneling models are also included to incorporate the impact ionization and tunneling effects. Further, Gummel and Newton's methods are used to attain a solution [26]. The physical models discussed above have been validated with the published results of Hyunjin Lee et al. [27]. Figure 3(a) shows the simulated and experimental I_d - V_{gs} characteristics of a 5 nm All-Around-Gate (AAG) FinFET at $V_{ds} = 1.0$ V. Figure 3(b) reflects the I_d - V_{ds} characteristics at $V_{gs} = 0.2$ V and $V_{gs} = 0.4$ V. It is visible from both Fig. 3(a) and (b) that the results are well-calibrated.

4 Results and Discussion

4.1 Device Scalability

In Fig. 4, the variation of drain current (I_d) with gate to source voltage (V_{gs}) at $V_{ds} = 0.5$ V is displayed in both linear and log scale for different configurations. It is observed that I_d increases with the increase in the V_{gs} and is maximum for GS-GAA FinFET structure. Also, it is found that leakage current (I_{off}) reduces significantly in GS configuration due to the enhanced gate coupling capacitances and reduced tunneling current. Thus, as shown in Fig. 5 due to increased drain current and decreased leakage current, higher switching ratio (I_{on}/I_{off}), and consequently, superior switching speed is obtained for GS-GAA FinFET configuration. Figure 6 represents the threshold voltage (V_{th}) for conventional FinFET, GAA FinFET, GS-GAA FinFET. The threshold voltage increases sequentially and recorded maximum for GS-GAA FinFET due to the enhanced control over the channel and better shielding of drain-side potential by the GAA structure, thereby improving the subthreshold characteristics [28].

In designing a microscopic device, the subthreshold swing (SS) is an essential short channel parameter. It is visible from Fig. 7 that SS is lowest and nearest to its ideal value (60 mV/dec) for GS-GAA FinFET. It is because the gate coupling capacitance increases as the permittivity (κ) increases, which increases the gate control on the channel, leading to an increased SS.

4.2 Analog Performance

In this subsection, from the perspective of the analog application, several important analog parameters such as transconductance (g_m), transconductance generation factor (TGF), output conductance (g_d), early voltage (V_{EA}), and intrinsic gain (A_v) are evaluated. Equations (1) and (2) estimates g_m and TGF, respectively, and both parameters should be high for better analog performance [29]. Figure 8 outlines a collaborated plot of g_m and TGF as a function of V_{gt} for all the three different configurations. Transconductance is a parameter that computes the change in drain current to the shift in V_{gs} at constant V_{ds}. Thus, g_m is obtained from the derivative of the I_d-V_{gs} curve. Comparing to their counterparts, GS-GAA FinFET has the maximum value of transconductance because

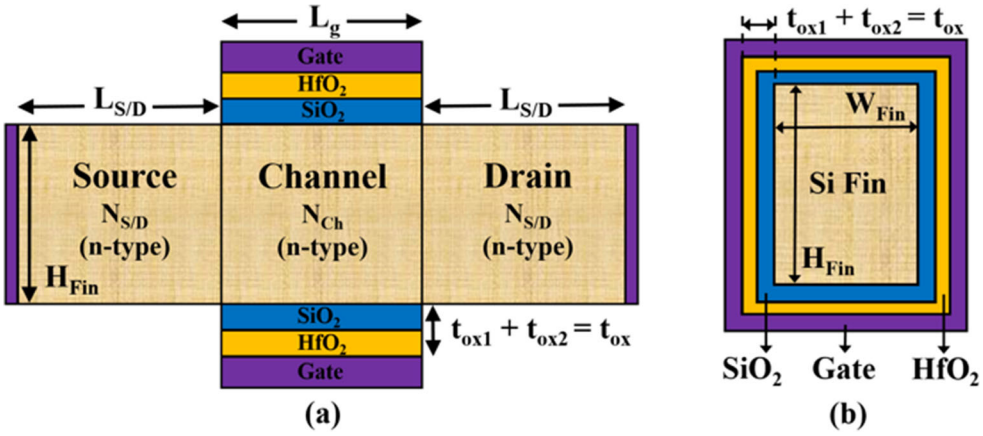


Fig. 2 (**a**) Horizontal (**b**) Vertical cross-sectional view cut along the silicon Fin of the proposed JAM GS-GAA FinFET

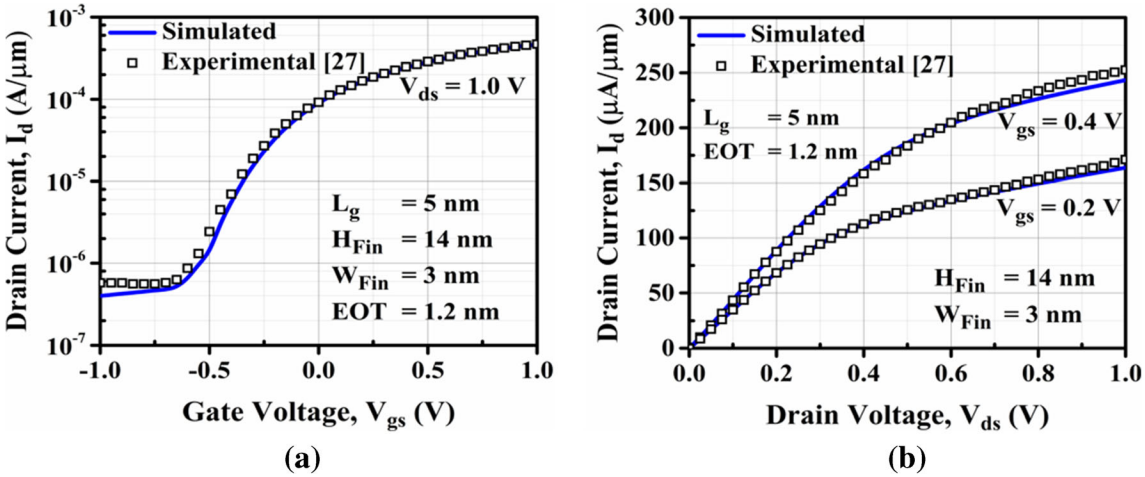


Fig. 3 (**a**) Calibrated I_d-V_{gs} characteristics of 5 nm All-Around-Gate FinFET. (**b**) Calibrated I_d-V_{ds} characteristics of 5 nm All-Around-Gate FinFET

of the enhanced gate control on the channel and reduced short channel effects. Also, gate-stack architecture enhances the average carrier velocity, resulting in improved electron mobility, and consequently, g_m increases [30]. The transconductance generation factor is the proportion of gain generated per unit power loss. The device operating at low supply voltages performs more efficiently for higher values of TGF. It is observed from Fig. 8 that the TGF changes only in the subthreshold region, with all most having the same value in the strong inversion region. GS-GAA FinFET configuration achieves the maximum value of TGF and is closest to its ideal value $\approx 40\ V^{-1}$ at the device's minimum SS (60 mV/dec) [29]. The reason for this is that the higher value of I_d corresponds to higher g_m and, consequently, higher TGF.

$$g_m = \partial I_d / \partial V_{gs} \tag{1}$$

$$TGF = g_m / I_d \tag{2}$$

$$g_d = \partial I_d / \partial V_{ds} \tag{3}$$

$$V_{EA} = I_d / g_d \tag{4}$$

$$A_v = g_m / g_d = (g_m / I_d) \times V_{EA} \tag{5}$$

Figure 9 represents a combined plot of drain current and output conductance as a function of a drain to source voltage at constant $V_{gs} = 0.5$ V. Output conductance determines the driving ability of a device, as defined in Eq. (3) [29]. The region of device operation determines the value of output conductance. Initially, g_d is large in the linear region but keeps on decaying as V_{ds} increases beyond pinch-off voltage due to drain-induced-barrier-lowering (DIBL) and channel length modulation (CLM) [31]. In the saturation region, g_d maintains a constant value. Thus, due to increased gate controllability and suppressed short channel effects, g_d is minimum for GS-

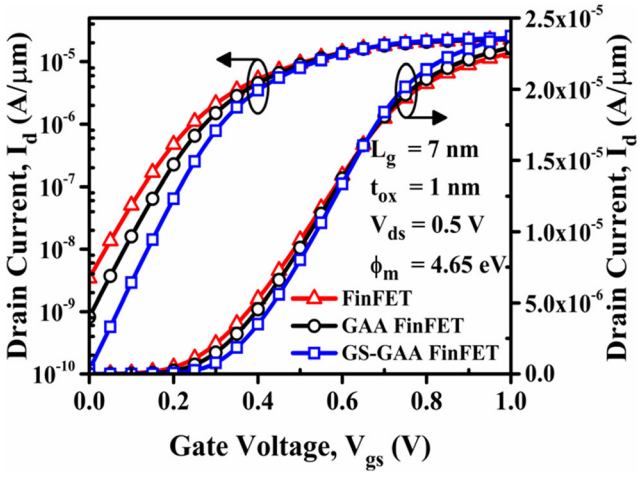


Fig. 4 Variation of I_d-V_{gs} in linear and log scale for different configurations

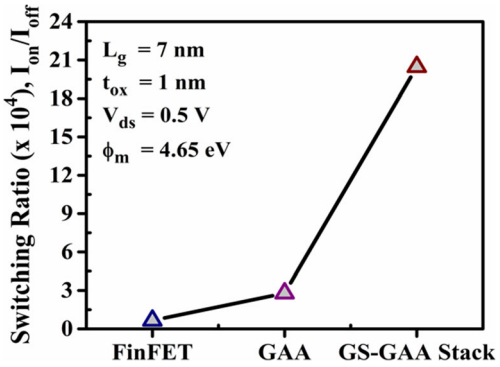


Fig. 5 Switching (I_{on}/I_{off}) ratio comparison for different structures

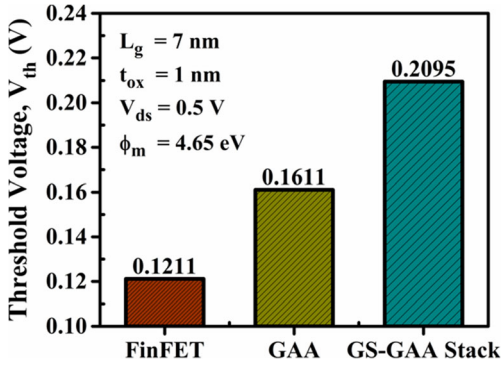


Fig. 6 Threshold Voltage (V_{th}) comparison for different structures

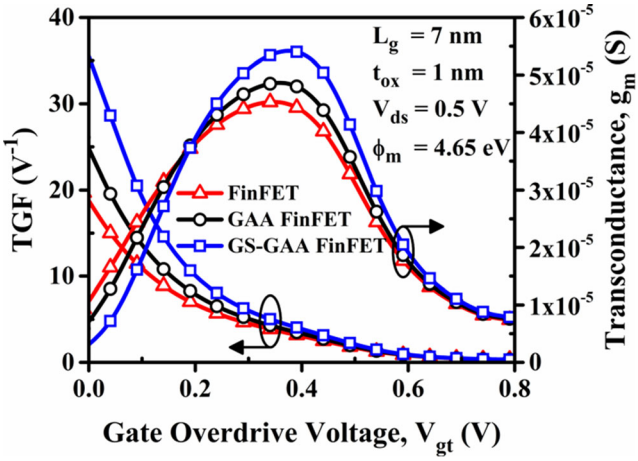


Fig. 8 TGF and g_m vs gate overdrive voltage (V_{gt}) for different structures

GAA FinFET when compared to their counterparts. Moreover, as depicted in Fig. 9, lower g_d means higher output resistance, which means less increase in drain current with V_{ds} in the saturation region.

Early voltage and intrinsic gain are evaluated using Eqs. (4) and (5), respectively, and they must be as high as possible for the enhanced analog performance of the device [29]. Figure 10 shows the variation of early voltage and intrinsic gain against V_{gt} for all three configurations. The maximum value of early voltage is obtained for GS-GAA FinFET structure due to the reduction in short channel effects, as demonstrated in Table 2. Also, the highest peak of intrinsic gain is observed for GS-GAA FinFET because of higher g_m and lower g_d.

4.3 RF Performance

From the RF application's point of view, RF parameters of vital interest are cut-off frequency (f_T), maximum oscillation frequency (f_{max}), gain frequency product (GFP), transconductance frequency product (TFP), and gain transconductance frequency product (GTFP). Figure 11 shows the variation of the gate to source capacitance (C_{gs}) and gate to drain capacitance (C_{gd}) as a function of V_{gt}. C_{gs} is plotted in log scale, and C_{gd} is on a linear scale to differentiate between the two curves. AC small-signal analysis has been performed to extract the values of these intrinsic gate capacitances (C_{gs} and C_{gd}) at an operating frequency of 1 MHz with DC voltage ramped from 0 V to 1 V with a step size of 0.05 V. It is observed from Fig. 11 that in the subthreshold region, both intrinsic gate capacitances increase very slowly with V_{gt}, but with further increase in V_{gt}, C_{gs} and C_{gd} increase swiftly due to the enhanced lateral field, which increases the movement of charge carriers from source side to drain side. In the superthreshold region, as expected, C_{gs} and C_{gd} become constant due to the non-contribution of V_{ds}. The GS-GAA FinFET structure exhibits higher C_{gs} and C_{gd} in comparison to other devices because gate capacitance increases with the increase in the dielectric permittivity [32].

In Fig. 12, total gate capacitance (C_{gg}) and cut-off frequency (f_T) are displayed as a function of V_{gt} for all three devices. The variation of C_{gg} with V_{gt} is the same as that of C_{gs} and C_{gd} because C_{gg} is the sum of intrinsic gate capacitances.

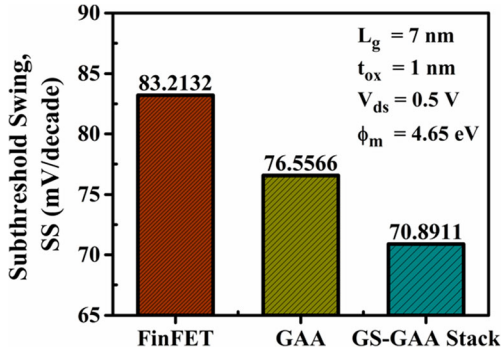


Fig. 7 Subthreshold Swing (SS) comparison for different structures

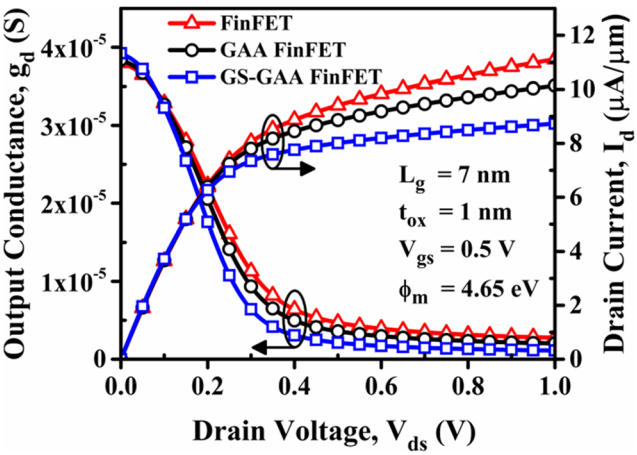


Fig. 9 g_d and I_d vs drain to source voltage (V_{ds}) for different structures

Table 2 Summary of electrostatic and analog parameters for different structures, $V_{ds} = 0.5$ V

Parameter	FinFET	GAA FinFET	GS-GAA FinFET
I_{on} (A/μm)	2.25×10^{-5}	2.29×10^{-5}	2.37×10^{-5}
I_{off} (A/μm)	34.04×10^{-10}	8.22×10^{-10}	1.15×10^{-10}
V_{th} (V)	0.12	0.16	0.20
SS (mV/dec)	83.21	76.55	70.89
g_m (S)	4.52×10^{-5}	4.85×10^{-5}	5.40×10^{-5}
TGF (V^{-1})	19.70	26.09	37.09
g_d (S)	11.29×10^{-6}	9.35×10^{-6}	6.41×10^{-6}
V_{EA} (V)	8.11	11.20	21.10
A_v (dB)	11.30	16.20	32.01

Generally, the frequency at which current gain becomes unity (0 dB) is known as cut-off frequency and is calculated using Eq. (6) [33]. As shown in Fig. 12, f_T decreases slightly for the GS-GAA FinFET structure in comparison to the other two designs. It happens mainly due to the enhanced value of C_{gs} and C_{gd}, although a higher value of g_m significantly compensates the deteriorating f_T.

$$f_T \approx g_m / 2\pi (C_{gs} + C_{gd}) \tag{6}$$

$$f_{max} = f_T / \sqrt{4R_g (g_{ds} + 2\pi f_T C_{gd})} \tag{7}$$

$$GFP = (g_m / g_d) \times f_T \tag{8}$$

Figure 13 represents the maximum oscillation frequency (f_{max}) and gain frequency product (GFP) with respect to V_{gt} for each device structure. The frequency at which maximum unilateral power gain becomes unity (0 dB) is known as maximum oscillation frequency (f_{max}). It is evaluated using Eq. (7) in which f_T, R_g, g_{ds}, C_{gd} denotes the cut-off frequency, gate resistance, drain to source output conductance, and gate to drain capacitance, respectively [34]. The values of R_g, g_{ds}, and C_{gd} are extracted using AC small-signal analysis at an operating frequency of 1 MHz with DC voltage ramped from 0 V to 1 V with a step size of 0.05 V [35, 36]. In Fig. 13, f_{max} is slightly lower for GS-GAA FinFET in the subthreshold and superthreshold region compared to their counterparts. It is due to the enhanced parasitic capacitances in the GS-GAA FinFET structure. Gain frequency Product is an essential parameter used in high-frequency applications, as specified in Eq. (8) [33]. It is noticed from Fig. 13, GFP increases as the gate overdrive voltage increases and then attains a maximum peak before falling to a constant value in the saturation region. GS-GAA FinFET configuration records the highest value of

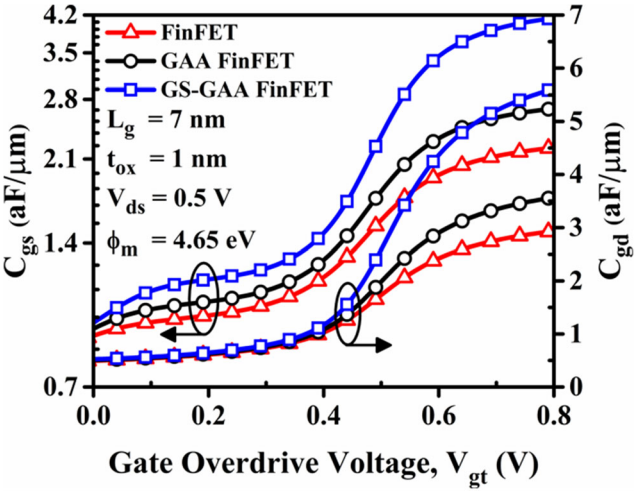


Fig. 11 Variation of C_{gs} and C_{gd} with respect to V_{gt} for each configuration

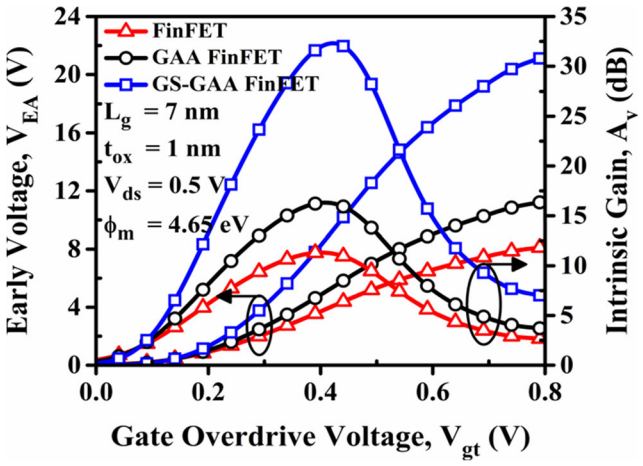


Fig. 10 V_{EA} and A_v vs gate overdrive voltage (V_{gt}) for different structures

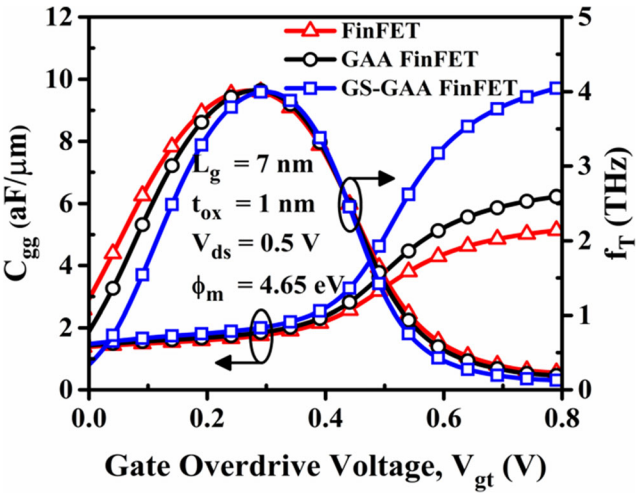


Fig. 12 Variation of C_{gg} and f_T with respect to V_{gt} for each configuration

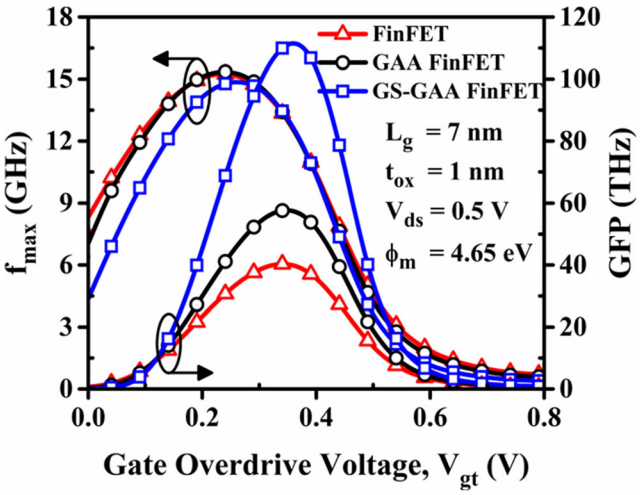


Fig. 13 Variation of f_{max} and GFP with respect to V_{gt} for each configuration

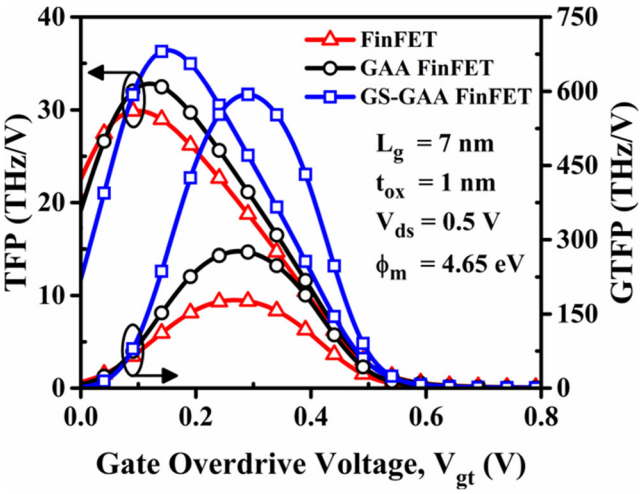


Fig. 14 Variation of TFP and GTFP with respect to V_{gt} for each configuration

GFP due to the improved value of transconductance and output transconductance with a minimal decrease in cut-off frequency.

The other important RF parameters of discussion are transconductance frequency product (TFP) and the gain transconductance frequency product (GTFP), as expressed in Eqs. (9) and (10), respectively [33]. TFP is mainly utilized in high-speed designs as it exhibits an agreement between bandwidth and power [37]. Figure 14 depicts the TFP and GTFP against V_{gt} for each configuration. The graph plotted for GTFP is calibrated by both the intrinsic gain and switching speed. The curve reflects that both TFP and GTFP increases as V_{gt} increases then reaches a maximum value at a particular value of V_{gs}. With further increase in V_{gt}, both TFP and GTFP decreases owing to increment in C_{gg} and maintains a minimum constant value as V_{gt} reaches in a saturation region. Again, for both TFP and GTFP highest value is obtained for GS-GAA FinFET structure due to the enhanced value of g_m, TGF, and reduced value of g_d with almost equivalent f_T. Thus, GS-GAA FinFET is the best suitable structure in terms of speed, transconductance, and gain. In Table 3, the values of different RF parameters are tabulated.

$$TFP = (g_m/I_d) \times f_T \tag{9}$$

$$GTFP = (g_m/g_d) \times (g_m/I_d) \times f_T = A_v \times TFP \tag{10}$$

Table 3 Summary of RF parameters for different structures, V_{ds} = 0.5 V

Parameter	FinFET	GAA FinFET	GS-GAA FinFET
C_{gs} (aF/μm)	2.21	2.67	4.12
C_{gd} (aF/μm)	2.92	3.55	5.58
C_{gg} (aF/μm)	5.13	6.22	9.71
f_T (THz)	3.96	3.92	3.79
f_{max} (GHz)	15.2	15.3	14.7
GFP (THz)	40.3	57.5	109.9
TFP (THz)	29.8	32.4	36.2
GTFP (THz)	176	275	593

5 Conclusion

In this work, we demonstrated the effect of high-κ SiO_2-HfO_2-TiN gate stack on device performance in terms of analog and RF parameters using the SILVACO ATLAS 3D simulator. The proposed JAM GS-GAA FinFET device shows the most improved results with the switching ratio increased by almost thirty-one times in comparison to conventional FinFET. Also, the leakage current and subthreshold swing get reduced by 96.62% and 14.80%, respectively, indicating that JAM GS-GAA FinFET configuration significantly reduces the SCEs. The JAM GS-GAA FinFET structure exhibits a remarkable improvement in analog parameters with A_V, V_{EA}, TGF increased by 183.27%, 160.17%, and 88.27%, respectively. Furthermore, RF parameters like GTFP gets enhanced by 236.93%, GFP by 172.70%, and TFP by 21.47% with a minimal decrease in f_T and f_{max} for GS-GAA FinFET configuration. Thus, from the results obtained in this study, the proposed JAM GS-GAA FinFET device can be considered as an attractive solution for the designing of analog and RF circuits.

Acknowledgments The authors are grateful to the Microelectronics Research Laboratory, Delhi Technological University, for supporting the work.

Authors' Contribution All authors contributed to the study conception and design.

Data Availability The above-mentioned authors have all the relevant data associated with this research work and will be dedicated to share that, if they will be asked to do so in future.

Compliance with Ethical Standard All the Ethical Standards have been seen by the authors and will supposed to follow them in future as well.

Conflict of Interests The authors declare that they have 'no known conflict of interests or personal relationships' that could have appeared to influence the work reported in this paper.

Consent to Participate & for Publication Since, the concerned research paper is based for the 'non-life science journal'. So, 'Not Applicable' here.

However, the authors have gone through all policies of journal and consent the authorities for further processing.

References

1. Moore GE (1998) Cramming more components onto integrated circuits. Proc IEEE 86(1):82–85. https://doi.org/10.1109/JPROC.1998.658762
2. Jeon DY, Park SJ, Mouis M, Barraud S, Kim GT, Ghibaudo G (2013) Low-temperature electrical characterization of junctionless transistors. Solid State Electron 80:135–141. https://doi.org/10.1016/j.sse.2012.10.018
3. Kumar A, Gupta N, Chaujar R (2016) TCAD RF performance investigation of transparent gate recessed channel MOSFET. Microelectron J 49:36–42. https://doi.org/10.1016/j.mejo.2015.12.007
4. Doria RT, Pavanello MA, Lee CW, Ferain I, Dehdashti-Akhavan N, Yan R, Razavi P, Yu R, Kranti A, Colinge JP (2010) Analog operation and harmonic distortion temperature dependence of nMOS Junctionless transistors. ECS Trans 31(1):13–20. https://doi.org/10.1149/1.3474137
5. Colinge JP, Lee CW, Afzalian A, Akhavan ND, Yan R, Ferain I, Razavi P, O'Neill B, Blake A, White M, Kelleher AM, McCarthy B, Murphy R (2010) Nanowire transistors without junctions. Nat Nanotechnol 5(3):225–229. https://doi.org/10.1038/nnano.2010.15
6. Biswas K, Sarkar A, Sarkar CK (2018) Fin shape influence on analog and RF performance of junctionless accumulation-mode bulk FinFETs. Microsyst Technol 24(5):2317–2324
7. Kim TK, Kim DH, Yoon YG, Moon JM, Hwang BW, Moon D, Lee GS, Lee DW, Yoo DE, Hwang HC, Kim JS, Choi YK, Cho BJ, Lee SH (2013) First demonstration of junctionless accumulation-mode bulk FinFETs with robust junction isolation. IEEE Electron Device Lett 34(12):1479–1481. https://doi.org/10.1109/LED.2013.2283291
8. Lee CW, Ferain I, Afzalian A, Yan R, Akhavan ND, Razavi P, Colinge JP (2010) Performance estimation of junctionless multigate transistors. Solid. State. Electron. 54(2):97–103. https://doi.org/10.1016/j.sse.2009.12.003
9. Hisamoto D, Lee WC, Kedzierski J, Takeuchi H, Asano K, Kuo C, Anderson E, King TJ, Bokor J, Hu C (2000) FinFET-A self-aligned double-gate MOSFET scalable to 20 nm. IEEE Trans Electron Devices 47(12):2320–2325. https://doi.org/10.1109/16.887014
10. Huang YC, Chiang MH, Wang SJ, Fossum JG (2017) GAAFET versus pragmatic FinFET at the 5nm Si-based CMOS technology node. IEEE J Electron Devices Soc 5(3):164–169. https://doi.org/10.1109/JEDS.2017.2689738
11. The International Technology Roadmap for Semiconductors 2.0: 2015, Itrpv, 2015 [Online]. Available: http://www.itrs2.net/
12. Lo SH, Buchanan DA, Taur Y, Wang W (1997) Quantum-mechanical modeling of electron tunneling current from the inversion layer of ultra-thin-oxide nMOSFET's. IEEE Electron Device Lett. 18(5):209–211. https://doi.org/10.1109/55.568766
13. Ribes G, Mitard J, Denais M, Bruyere S, Monsieur F, Parthasarathy C, Vincent E, Ghibaudo G (2005) Review on high-k dielectrics reliability issues. IEEE Trans Device Mater Reliab 5(1):5–19. https://doi.org/10.1109/TDMR.2005.845236
14. Robertson J (2004) High dielectric constant oxides. Eur Phys Journal Applied Phys 28:265–291. https://doi.org/10.1051/epjap:2004206
15. Onishi K, Kang CS, Choi R, Cho HJ, Gopalan S, Nieh RE, Krishnan SA, Lee JC (2003) Improvement of surface carrier mobility of HfO2 MOSFETs by high-temperature forming gas annealing. IEEE Trans. Electron Devices 50(2):384–390. https://doi.org/10.1109/TED.2002.807447
16. Kerber A, Cartier E, Pantisano L, Degraeve R, Kauerauf T, Kim Y, Hou A, Groeseneken G, Maes HE (2003) Origin of the threshold voltage instability in SiO2/HfO2 dual layer gate dielectrics. IEEE Electron Device Lett. 24(2):87–89. https://doi.org/10.1109/LED.2003.808844
17. Gupta N, Chaujar R (2016) Optimization of high-k and gate metal work function for improved analog and intermodulation performance of gate stack (GS)-GEWE-SiNW MOSFET. Superlattice Microst 97:630–641. https://doi.org/10.1016/j.spmi.2016.07.021
18. Sjöblom G (2006) Metal gate technology for advanced CMOS devices, Ph.D. dissertation Dept. Engg. Sci., Uppsala Univ., Sweden
19. Liu Y, Kijima S, Sugimata E, Masahara M, Endo K, Matasukawa T, Ishii K, Sakamoto K, Sekigawa T, Yamauchi H, Takanashi Y, Suzuki E (2006) Investigation of the TiN gate electrode with tunable work function and its application for FinFET fabrication. IEEE Trans Nanotechnol 5(6):723–728. https://doi.org/10.1109/TNANO.2006.885035
20. Vitale SA, Kedzierski J, Healey P, Wyatt PW, Keast CL (2011) Work-function-tuned TiN metal gate FDSOI transistors for subthreshold operation. IEEE Trans. Electron Devices 58(2):419–426. https://doi.org/10.1109/TED.2010.2092779
21. Mohapatra SK, Pradhan KP, Singh D, Sahu PK (2015) The role of geometry parameters and fin aspect ratio of sub-20nm SOI-FinFET: an analysis towards analog and RF circuit design. IEEE Trans Nanotechnol 14(3):546–554. https://doi.org/10.1109/TNANO.2015.2415555
22. Trivedi N, Kumar M, Gupta M, Haldar S, Deswal SS, Gupta RS (2016) Investigation of analog / RF performance of High-k spacer junctionless accumulation-mode cylindrical gate all around (JLAM-CGAA) MOSFET, in 2016 IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics Engineering (UPCON), pp. 201–205, https://doi.org/10.1109/UPCON.2016.7894652
23. Raskin JP, Chung TM, Kilchytska V, Lederer D, Flandre D (2006) Analog/RF performance of multiple gate SOI devices: wideband simulations and characterization. IEEE Trans. Electron Devices 53(5):1088–1095. https://doi.org/10.1109/TED.2006.871876
24. Bhattacharya D, Jha NK (2014) FinFETs: from devices to architectures. Adv Electron 2014:21–55. https://doi.org/10.1155/2014/365689
25. Kumar B, Kumar A, Chaujar R (2020) The effect of gate stack and high-K Spacer on device performance of a junctionless GAA FinFET, in IEEE VLSI Device, Circuit and System Conference (VLSI-DCS), pp. 159–163, https://doi.org/10.1109/VLSIDCS47293.2020.9179855

26. ATLAS (2016) User's manual. SILVACO International, CA, Santa Clara, USA
27. Lee H, Yu LE, Ryu SW, Han JW, Jeon K, Jang DY, Kim KH, Lee J, Kim JH, Jeon SC, Oh JS, Park YC, Bae WH, Lee HM, Yang JM, Yoo JJ, Kim SI, Choi YK (2006) Sub-5nm all-around gate FinFET for ultimate scaling. in Digest of Technical Papers - Symposium on VLSI Technology 25(9):58–59. https://doi.org/10.1109/VLSIT.2006.1705215
28. Kumar A, Gupta N, Tripathi SK, Tripathi MM, Chaujar R (2020) Performance evaluation of linearity and intermodulation distortion of nanoscale GaN-SOI FinFET for RFIC design. AEU - Int J Electron Commun 115:153052. https://doi.org/10.1016/j.aeue.2019.153052
29. Pradhan KP, Mohapatra SK, Sahu PK, Behera DK (2014) Impact of high-k gate dielectric on analog and RF performance of nanoscale DG-MOSFET. Microelectron J 45(2):144–151. https://doi.org/10.1016/j.mejo.2013.11.016
30. Narendar V, Girdhardas KA (2018) Surface potential modeling of Graded-Channel gate-stack (GCGS) high-K dielectric dual-material double-gate (DMDG) MOSFET and analog/RF performance study. Silicon 10(6):2865–2875. https://doi.org/10.1007/s12633-018-9826-z
31. Amin SI, Sarin RK (2015) Charge-plasma based dual-material and gate-stacked architecture of junctionless transistor for enhanced analog performance. Superlattice Microst 88:582–590. https://doi.org/10.1016/j.spmi.2015.10.017
32. Malik P, Gupta RS, Chaujar R, Gupta M (2012) AC analysis of nanoscale GME-TRC MOSFET for microwave and RF applications. Microelectron Reliab 52(1):151–158. https://doi.org/10.1016/j.microrel.2011.07.070
33. Mohapatra SK, Pradhan KP, Artola L, Sahu PK (2015) Estimation of analog/RF figures-of-merit using device design engineering in gate stack double gate MOSFET. Mater Sci Semicond Process 31:455–462. https://doi.org/10.1016/j.mssp.2014.12.026
34. Gupta N, Kumar A, Chaujar R (2015) Impact of device parameter variation on RF performance of gate electrode workfunction engineered (GEWE)-silicon nanowire (SiNW) MOSFET. J Comput Electron 14(3):798–810. https://doi.org/10.1007/s10825-015-0715-z
35. Shimizu Y, Kim GC, Murakami B, Ueda K, Utsurogi Y, Cha S, Matsuoka T, Taniguchi K (2004) Drain current response delay of FD-SOI MOSFETs in RF operation. IEICE Electron Express 1(16):518–522. https://doi.org/10.1587/elex.1.518
36. Shin S, Kang IM, Kim KR (2012) Extraction method for substrate-related components of vertical junctionless silicon nanowire field-effect transistors and its verification on radio frequency characteristics. Jpn J Appl Phys 51, https://doi.org/10.1143/JJAP.51.06FE20
37. Kumar A, Tripathi MM, Chaujar R (2017) Investigation of parasitic capacitances of In2O5Sn gate electrode recessed channel MOSFET for ULSI switching applications. Microsyst Technol 23(12):5867–5874. https://doi.org/10.1007/s00542-017-3348-2

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

Annals of Operations Research
https://doi.org/10.1007/s10479-020-03898-w

S.I. : DESIGN AND MANAGEMENT OF HUMANITARIAN SUPPLY CHAINS

Application of Industry 4.0 technologies for effective coordination in humanitarian supply chains: a strategic approach

Pravin Kumar[1] · **Rajesh Kr Singh**[2]

Accepted: 8 December 2020
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC part of Springer Nature 2021

Abstract

The coordination in post-disaster functions is a major challenge for the humanitarian supply chains. Due to uncertainty and temporary nature of the disasters, it becomes very difficult to ensure coordination among the aid organizations engaged in the humanitarian supply chain. The increasing incidents of the natural, technological and man-made disasters have attracted the attention of the researchers, governmental and non-governmental actors. Most of the disasters are not under the control of the human being. Some of the man-made disasters also create a huge loss of the human lives with damages to the public and private properties due to some social and political causes. From literature, it is observed that the technologies of Industry 4.0 may help in improving the coordination among stakeholders of the commercial supply chains but the application of industry 4.0 technologies in humanitarian relief chains is not well explored. Therefore, the present study has bridged the existing research gaps by exploring the application of Industry 4.0 technologies for improving the coordination in humanitarian supply chains. A hybrid approach comprising of Analytic Hierarchy Process and ELECTRE has been used to propose a strategic approach for mapping the measures of coordination with critical success factors. Findings of the study will help stakeholders of humanitarian supply chains in implementing Industry 4.0 technologies across different functions for prompt service and accurate decision making.

Keywords Coordination · Industry 4.0 · Humanitarian supply chain · Analytic hierarchy process · ELECTRE

✉ Rajesh Kr Singh
rksdce@yahoo.com

[1] Mechanical Engineering Department, Delhi Technological University, New Delhi, India

[2] Management Development Institute Gurgaon, Mehrauli Road, Gurugram, India

Published online: 03 January 2021

1 Introduction

In the recent years, humanitarian supply chain (HSC) has attracted the attention of academicians, researchers, and industry professionals due to the frequent outbreak of the epidemic/pandemic diseases., Occurrences of the natural disasters like earthquake, hurricane, cyclone etc., man-made disaster like a terrorist attack, communal riots, etc. are also very frequent. A disaster not only causes a huge loss to human beings, but it also affects all sectors such as tourism and hospitality, events and entertainment, cab-hailing, aviation, pharma sector, manufacturing. In the case of the humanitarian supply chain (HSC), there are high stakes but few established measurements of performances as compared to commercial supply chains (Beamon 2004). The activities required for relief operations in HSC are managed by several actors (Van Wassenhove 2017). According to Maghsoudi et al. (2018), the humanitarian operations include disaster relief related activities but not limited to activities like needs assessment, planning, procurement, warehousing, and distribution of the supplies to beneficiaries. The exact needs assessment is an important component of disaster management. Many countries face social and economic crisis due to the constraints of different resources, infrastructure and facilities. This also means that the relief chain is formed quickly and maintained for shorter periods which make it unpredictable (Oloruntoba and Gray 2006; Maon et al. 2009).

In some cases, the assessment of the disaster is very difficult (Blecken 2010; Whybark et al. 2010). The random occurrence of the disasters demands immediate action from the relief organizations. It has been observed that proper coordination with logistics service providers and other participating agencies is required to serve the people in need (Singh et al. 2018). According to Akhtar et al. (2012), proper coordination is very essential for the quick and effective response to man-made or natural disasters. Sharing information with organizations is very crucial to maximize resource allocation to alleviate people's suffering. The quality information is the basic requirement of disaster management. It is important to have the accurate information of the disaster locations, intensity, level of damages and estimates of causalities or affected people and their needs (Moshtari and Gonçalves 2016; Da Costa et al. 2014). In HSCs, the system is formed hastily, therefore lack of trust prevails among the aid organizations (Tatham and Kovács 2010; Fawcett and Fawcett 2013). Many researchers have worked on the coordination issue in context to the commercial supply chains (Arshinder et al. 2008, Singh et al. 2019), but in the context of the humanitarian supply chain, it has not been well explored (Akhtar et al. 2012).

Due to some major crises such as the escalation of conflicts in several countries, climate change, and a series of natural disasters, humanitarian operations management has gained the special attention within the operations management field (Gunasekaran et al. 2018; Dubey and Gunasekaran 2016; Dubey et al. 2016; Sodhi 2016; Ivanov et al. 2017; Altay et al. 2018). The need for better understanding of the humanitarian supply chain has been increased due to the complexity of operations and increasing cases of natural and man-made disasters (Altay and Ramirez 2010). The major challenge in managing humanitarian operations during disasters is to coordinate diversified activities so that services could be provided on time to needy people (Singh et al. 2018). Ergun et al. (2014) also noted the complexity of coordination in the HSC due to difficult operational situations (Dubey et al. 2019).

According to Dubey and Gunasekaran (2016), the HSC is a complex network of supply chain actors with non-profit motives. Bhattacharya et al. (2014) distinguished the HSCs from commercial supply chains in terms of reactive logistics and rapid response

time. Different evidence of calamities and disasters have shown that effective coordination among different agencies involved in humanitarian operations plays a crucial role in resilient HSC (Singh et al. 2018). This coordination process can be made more effective by exploring the application of Industry 4.0 technologies across different functions of HSCs. Researchers have analyzed coordination issues in supply chains of the manufacturing sector (Singh et al. 2019) but have not analyzed how Industry 4.0 based technologies can accelerate coordination among stakeholders of HSCs. It has been observed that the coordination in the humanitarian supply chain reduces the losses and fulfils the need of the sufferers. Coordination among the aid organizations can be improved by quality information flow and visibility of items movement or distribution in the affected areas. Limited research has been done on the successful application of emerging technologies for humanitarian supply chains in Industry 4.0 environment (Singh et al. 2018). Therefore, there is an urgent need to explore possible applications of recent technologies such as big data analytics, internet of things, cyber-physical systems, cloud computing, etc. for better management of the relief chain.

The major research questions which are addressed in this study are given below as:

(a) What are the important measures of coordination in humanitarian supply chain and how to implement them?
(b) What are the critical factors of implementation of Industry 4.0 technologies in Humanitarian supply chains for improving coordination?
(c) How to prioritize these critical success factors for successful application of Industry 4.0 technologies in HSCs.

Organization of this paper is as follows: Sect. 2 deals with the literature review. Section 3 describes the research methodology. Section 4 discusses the results. Section 5 concludes the research and presents the limitations, future scope and managerial implications.

2 Literature review

A disaster not only affects the daily activities and the lives of the people suffered but also the other people indirectly linked to the disaster (Alexander 1997). The pandemic outbreaks in recent years have badly affected the global value chains, economy, transportation, tourism, industry, society and every individual. Most of the countries in the world are suffering from some infectious disease. It is very difficult to predict the occurrence of a disaster, but the effect of the disaster and the losses can be minimized by proper coordination among all stakeholders of HSCs (Kovacs and Spens 2007).

Moshtari (2016) have observed that coordination and collaboration terms are used interchangeably, whereas both the terms serve a different purpose in context to humanitarian supply chains. According to Ergun et al. (2014), coordination, as well as collaboration in the humanitarian supply chains, are more complex due to difficult operational circumstances. Coordination is more related with the sharing of information and resources, whereas collaboration means working together to create something new (Balakrishnan and Geunes 2004; Ergun et al. 2014; de Camargo et al. 2019).

The coordination can be vertical or horizontal. Vertical coordination is used to coordinate the activities with upstream or downstream directions. Horizontal coordination is used to coordinate the activities at the same level within the chain (Balcik et al. 2010). The

lateral relationships among actors involved in the disaster relief operations can increase the availability of current and valuable information (Altay and Pal 2014; Moshtari 2016; Srinivasan and Swink 2018).

Wentz (2006) highlighted the importance of ICT capability for both military and civil organizations in planning and designing the relief and reconstruction efforts for better coordination of the relief activities (Balcik et al. 2010; Tatham and Kovács 2010; Ergun et al. 2014; Altay and Pal 2014). Many researchers have focused on the application of big data analytics for managing humanitarian supply chains (Fawcett and Fawcett 2013; Goswami et al. 2018; Papadopoulos et al. 2017; Dubey et al. 2018). Dubey et al. (2018) empirically tested the use of big data and predictive analytics (BDPA) capability in improving visibility and coordination.

A vast range of theories can be found in HSCM (Humanitarian Supply Chain management. Tabaklar et al. (2015) reviewed the theories in HSCM from 1995 to 2014. They reviewed approx. 60 articles based on the theories in HSC and observed that 57 articles used middle-range theory. All the middle-range theories are operations research including inventory control theory, ant colony optimization, queuing theory, graph theory, fuzzy set theory, Bayes' theorem, game theory, oscillation physics theory, and auction theory. They found only one article that was concerned with grand theory (Complexity theory). Another theory is the grounded theory which is concerned with the methodological approach. They also observed that very few articles are based on the cultural bias theory or structural theory which is not in common use. Most of the articles are based on inventory management theory and systems theory. According to mixed form of systems theory and resource-based view for better vertical and horizontal coordination in HSC, proper utilization of the resources can be done by using Industry 4.0 technologies. In this study, for effective resource utilization, some critical success factors for the implementation of the Industry 4.0 technologies have been identified and mapped with the measures of coordination in Humanitarian supply chains.

2.1 Need for coordination in HSC

Coordination among different stakeholders involved in relief operations make humanitarian supply chains resilient and efficient (Singh et al. 2018). Coordination among the aid organizations and other stakeholders helps in the assessment of the impact and in formulating effective action plan to minimize the post-disaster effect (Cozzolino 2012; Datta and Christopher 2011). It is also required for effective utilization of the limited available resources and procurement of the additional resources required at the minimum cost to provide better relief (Balcik et al. 2010). Samii and Van Wassenhove (2003). Proper coordination provides a common strategy under the leadership of the umbrella organization and also strengthens the financial and human resources (Akhtar et al. 2012; Dubey et al. 2019). Coordination reduces the cost of humanitarian operations and improves the response time for providing the best services to the needy (Akhtar et al. 2012). Lack of timely action plan and financial resources lead to various challenges in ensuring coordination (Tomasini and Van Wassenhove 2009; McLachlin and Larson 2011). Balcik et al. (2010) explored some factors that affect the coordination, such as diversity, donors' expectations, funding structures, competition for funding, uncertainty, resource scarcity, and oversupply, etc. Lack of coordination among humanitarian actors may cause the prolongation of the suffering of the affected people (Jin et al. 2015). There is the chance of duplication of efforts, waste of resources, and inequitable distribution of the resources due to poor management and

lack of coordination (Balcik et al. 2010; Moshtari 2016). Kabra and Ramesh (2015) have also shown their concern for duplication of resources and services due to poor coordination, and they emphasized the requirement of synchronization to work together to achieve a common goal and enable a coherent, effective, and efficient response (Gillmann 2009; Dubey and Altay 2018).

Fritz Institute (2005) observed some challenges in coordination of relief operations such as the use of manual supply chain, improper assessments and poor planning, lack of collaboration and coordination, unavailability of expert logisticians. Balcik et al. (2010) and Whiting and Ayala-Öström (2009) have shown their concern for the shortage of expert logisticians as it requires good professionalism. It has been observed that Information Technology may help in the effective coordination of HSC (Thomas 2003; Gustavsson 2003).

As per the resource-based view theory, organizations may achieve the desired competitive advantage through strategic resources (Barney 1991). Although contingency theory suggests that organisations should select the resources carefully depending upon prevailing conditions (Eckstein et al. 2015). Dubey and Altay (2018) identified the drivers of coordination in the humanitarian supply chains on basis of two theories. These are the resource view and relationship theory. They have used ISM (Interpretive Structural Modelling) for linking these coordination drivers. These drivers are information and communication technologies (ICTs), information sharing, visibility, training, mutual learning, contingency leadership, performance measurement systems (PMS), swift trust, commitment, cultural cohesion, and regular meetings. These drivers are very helpful in deciding the measures for the post-disaster humanitarian supply chains.

2.2 Post-disaster measures of humanitarian supply chains

The annual growth rate of humanitarian funding has been recorded as 8.9%. It has been increased from 16 billion US$ in 2009 to 24.9 billion US$ in 2015 (Global Assistance Report 2015). This surge in funding should be used to enhance the training of the workforce and for exploring the application of emerging technologies, across different operations of HSCs. It will also help in developing a robust mechanism for coordination among different agencies involved across the world. Various studies have been done to measure the coordination in commercial supply chains but coordination in context to HSCs is not well explored. Just after the occurrence of disasters, the chaotic environment and lack of sufficient resources at that time of need is a big obstacle for the effective coordination among the aid organizations involved in the relief operations (McLachlan and Larson 2011). John et al. (2019) also observed that the lack of coordination and communication among different actors during a relief operation was a common phenomenon which leads to delays in emergency relief and hence the sufferings of victims (Dasaklis and Pappis 2018). Dubey et al. (2019) investigated the role of swift trust in achieving supply chain coordination. They explored some enablers of swift trust which can be translated into the coordination. Jeong et al. (2014) proposed a framework for the design of logistics network based on an integrated model of efficiency, risk and robustness metrics. Based on the literature review, some of the important measures of successful coordination and collaboration are discussed in the following subsections:

The flow of quality information (C1) Here quality information means reliable and real-time information. During the disasters, many sources of information start to work which leads to a chaotic situation. Ergun et al. (2014) observed that the information sharing in the HSC has always been undervalued aspect. They found two major reasons for it. These

are (a) selective collection of the data catering to own needs of the organizations and (b) inconsistency in the data collection. It causes difficulty in inter-organization data processing (Seybolt 2009). Information sharing is a very important and urgent need of humanitarian organizations., Collection and processing of the real-time data must be reliable and consistent. The information-sharing also develops trust among the organizations and helps in post-disaster planning. A coordinator of HSC has to conduct activities like appropriate information sharing and communication, travelling, documentation, etc. (Akhtar et al. 2012). During the crisis, information sharing is declared as the determinant of success (Long and Wood 1995; Gunasekaran and Ngai 2003; Soin 2004; UNDP 1993; Long 1997).

Flexibility in humanitarian operations (C2) In the humanitarian supply chains, the ability of an organization to deliver the items on time and its ability to change the volume of supply is an important factor in determining the level of coordination (Abidi et al. 2013). The flexibility can be further divided into different performance indicators. These are volume flexibility, mix flexibility and percentage of non-prepositioned goods (Santarelli et al. 2015). To respond to the changing scenario in the relief operation, flexibility in providing the product and processes is considered as pre-requisites. The inventory carrying costs are nothing compared to the time values of commodities in humanitarian aid (Long and Wood 1995). The humanitarian supply chains must be agile consisting of the flexibility and speed to respond to the need of the people in the disaster-affected area with the necessary aid. (Dubey and Gunasekaran, 2016; Dubey, 2019).

Service reliability (C3) The supply chain network in HSC is composed of suppliers, distributors, fixed/movable warehouses, victims, etc. The actual problems may be different from our assumptions for the conventional supply chain and the facilities might be inaccessible during and after the disasters (Diabat et al. 2019). In this situation, effective planning and fast response are the most important management tools to minimize the consequences of disasters. The reliable network design for the relief operation is very crucial for fulfilling the demand of the disaster affected people (Yahyaei and Amiri 2019). Reliability of the relief network is related to ensuring the recovery of the disrupted supply chain and supplying the necessary aid to the victims.

Response Time (C4) Timely delivery of relief to needy people is very important for effective HSC (Abidi et al. 2013). Disasters can be classified into two categories: Slow-onset disasters and sudden-onset disasters (Jabbour et al. 2019). In the slow-onset disasters, actors get some time to organize the relief activities. But, when sudden-onset disasters occur, the response time is very short. Santarelli et al. (2015) proposed five performance parameters. These are project duration, response time, on-time delivery and warehouse location. The short duration of time for relief operation is very important. The main requirement during the crisis is to supply the right items at the right time in the right location to the sufferers. The initial three to five days just after a disaster is very crucial for a country to meet the challenge of providing emergency relief (Tatham and Kovacs 2007). The donors and logisticians both have to work in proper coordination to meet the needs of every disaster (Fritz Institute 2005).

Reducing the impact of the diversity actors (C5) One of the important characteristics of the relief operation is having the diversity actors. The diversity actors stem from two sources. Some actors, who reach the site quickly may not have proper training and experience to handle the relief operations. Since they are the first ones to act when a disaster occurs, they also act as primary sources of information. Timely application of ICT tools may also ensure the flow of right information to other stakeholders. In addition to it, many NGOs and government agencies are also not well equipped with the use of different technologies. Hence, it is essential to impart the training to these actors for improving

the coordination (Falasca and Zobel 2012). Differences among the aid organizations may be also a hurdle in the coordination of the relief operations (Van Wassenhove 2006). For example, the language and culture differences between local and international organizations may create obstacles in the coordination (Moore et al. 2003; Balcik et al. 2010).

In general, the government of the affected country conducts the relief operations but other actors and agencies may also help in relief operations as per the laws of the country. It creates diversity among the actors. Sometimes, the government may be dysfunctional due to lack of experience, lack of resources and lack of knowledge regarding the relief operations (Seaman 1999). The HSC operates in the chaotic environment involving the host government, military, local and international NGOs, private companies etc. Converging at the same location, the coordination between the actors demands rapid trust-building (Tatham and Kovács 2010; Heaslip et al. 2012; John et al. 2019). Trust is usually developed through long-term association and personal relationships. Thus, reducing the impact of diverse actors is very important for the efficient humanitarian supply chain.

Effective utilization of resources (C6) Due to the uncertainties associated with disaster locations, timing, and intensity, the matching of demand and supply are very difficult in HSC. Lack of supporting resources creates difficulties in coordination. The volunteers/temporary workers in the relief operations are also working on the short term basis, which leads to lack of experience in coordination activities during a disaster relief operation (Balcik et al. 2010).

The humanitarian supply chain faces a problem of scarcity as well as the abundance of resources at different stages. Due to poor resource planning, the resources are scarce during the earlier stages of a disaster; and are in abundance in the later stages of a disaster due to the number of donors coming forward at different points of time. The organizations lose the track of resources and some donations end up unsolicited and end up taking a bulk of inventory storage space (Balcik et al. 2010; Reitjens et al. 2007). There may be the problem of visibility of stocks due to poor inventory management and wastage of aid (Bealt et al. 2016). The chances of duplication and overlapping of efforts are also observed in HSC (Heaslip et al. 2012; Jahre and Jensen 2010). The management has to address the issues of existing inventory and sources of supplies (Long and Wood 1995), lead times (UNDP 1993), and forecast of the demand (Gunasekaran and Ngai 2003).

In the humanitarian supply chain, the cost is not a predominant factor as in the case of the commercial supply chain (Santarelli et al. 2015). But, it is still considered a relatively important factor. Cost evaluation is very difficult under demand uncertainty. The cost can be evaluated only when normalcy is restored (Santarelli et al. 2015). Santarelli et al. (2015) proposed the five performance indicators in this category. These are the cost of the goods, transportation costs, warehousing costs, percentage of claims and the goods not distributed.

Effective Logistics (C7) As per the report of Trunick (2005), approx 80 per cent of the activities of relief operations are only related to logistics. Logistics operations include acquiring, inventory management, opening and controlling the emergency centres, evacuating the disaster area, and opening of shelters and providing proper care (Altay and Green 2006). The main challenge arises from the damage of infrastructure post-disaster and limited transportation (Balcik and Beamon 2008). Generally, relief agencies have to hire local vehicles and drivers. During the emergency, hiring of the vehicles may be costly and scarce. The topography of a region may also be a big hurdle in assessing the affected populations (Balcik et al. 2010).

Wagner et al. (2020) presented two major examples of the World Food Program (WFP) and the International Committee of Red Cross for the year 2018 to show the complexity of humanitarian logistics. The world's largest humanitarian organization (WFP) purchased

3.6 million metric tons of foods and distributed among 86.7 million people in 83 countries 92 aircraft, 20 ships and 5600 trucks daily in 2018. Similarly, the Logistics Division of the International Committee of the Red Cross (ICRC) purchased goods and services worth CHF 565 million, of which items worth CHF 315 million were delivered to people in need. ICRC's supply chain manages 210,000 square meters of warehouse space, 22 aircraft and 5624 units of vehicles and equipment. Thus, logistics management is an important part of the humanitarian supply chain and it must be effective. Deprivation cost is more than the logistics cost. Shao et al. (2020) worked on the importance of deprivation cost in humanitarian logistics.

2.3 Application of Industry 4.0 technologies for coordination in HSC

The application of Industry 4.0 technologies for the digitization of manufacturing processes and supply chain management has been addressed by many researchers and practitioners (Singh et al. 2019). But, the application of Industry 4.0 in the humanitarian supply chain is at the nascent stage. However, the learnings from the manufacturing sector can lead to exploring possible applications of Industry 4.0 technologies in HSCs. It may be used to integrate the systems of stakeholders involved in HSC, data processing and information sharing and for improving traceability of items. Dai et al. (1994) and Yang et al. 2013) examined the use of IoT in disaster management. Although some studies are available in the literature (Dai et al. 1994; Yang et al. 2013) that examines the usefulness of IoT in disaster management. This is not in the context of developing countries like India. The paradigm of IoT may be used for effective data collection and to enhance the ability to share the data. This may be applied for disaster risk management and prevention, Emergency response, and Disaster recovery (Sinha et al. 2019). Telephone lines in the disaster area become overloaded due to a large number of parallel calls. The cloud computing improves the computer database providing detailed and real-time information. Recovery of the data from the affected area becomes very difficult due to the need for replacement of the hardware and software of the local damaged devices. Although the internet can be disrupted the user data and cloud server can be protected. Cloud computing can be very useful here. The major advantage of cloud computing is to store the data on many systems at different locations (Velev and Zlateva 2011).

Cyber-Physical System (CPS) ensures seamless integration of computational algorithms and physical components (Zander et al. 2015). CPS may be applied for achieving the goal of smart emergency response. However, its application may vary as per the existing scenario. The human–machine interactions must be also considered while designing smart systems of tomorrow. The environmental emergency can occur in different forms, such as an earthquake, tornado, fire, hurricane, cyclone, and so forth. All of these different scenarios require different treatments. The CPS may help in getting the exact location of resources and better service to needy during the disaster.

Industry 4.0 comprises technologies such as autonomous robots, simulation, horizontal and vertical systems integration, cybersecurity, augmented reality, the cloud, additive manufacturing, industrial internet of things. Based on learnings from the manufacturing sector, the application of Industry 4.0 technologies may be explored for humanitarian operations as shown in Table 1.

Table 1 Application of Industry 4.0 technologies in HSCs

SI. No.	Technology	Application in humanitarian supply chain	References
1.	Big data analytics	Big data and predictive analytics (BDPA) capability can improve decision making, data management and coordination in humanitarian supply chains	Dubey et al. (2018), Fawcett and Fawcett (2013), Goswami et al. (2018), Papadopoulos et al. (2017), Jabbour et al. (2017), Witkowski (2017), Akoka et al. (2017)
2.	Autonomous robots	They can work in places where human workers are restricted to work. They can work efficiently within the given time limit	Lorenz et al. (2015), Bauer (2016)
3.	The industrial Internet of Things	It may help in interaction and response between the objects, intelligent shelving in warehouse storage, tracking and tracing of goods	Dai et al. (1994), Yang et al. (2013), Neugebauer et al. (2016), Zhang et al. (2017),
4.	Simulation	Simulation of real-time data and producing the virtual world. This may help in creating the virtual model of the incidents	Lorenz et al. (2015), Simons et al. (2017), Vaidya et al. (2018).
5.	Augmented reality	Presentation of composite view using a computer-generated image of the real scenario of the disaster	Van Lopic et al. (2020), Gattullo et al. (2019), Masoni et al. (2017).
6.	Cloud computing	It may help in preserving the data far from the area of disaster and the data can be easily accessed from different locations	Velev and Zlateva (2011), Lorenz et al. (2015), Liu and Xu (2017), Landherr et al. (2016), Marilungo et al. (2017).
7.	Cyber-physical systems	It will help in secured and reliable communications and integrate the human with physical systems	Drath and Horch (2014), Hozdic (2015), Schlectendahl et al. (2015)

After exploring the application of different Industry 4.0 technologies for effective HSC, we would try to identify critical success factors for the application of these technologies in the next section.

2.4 Critical success factors for the application of Industry 4.0 technologies in humanitarian supply chains

The humanitarian supply chain is often treated as an amalgamation of abstract activities without any sense of coherence. Also, the connections between various stages of an HCM (Human Capital Management) are weak or nonexistent (Taylor and Petiit 2009). In the post-crisis situation, the wastage is up to 25% (Beresford and Rugamba 1996). The need for CSFs arises primarily because of the following reasons: Firstly, diversity actors (UNDP 1993). Second, high-stress condition and short-time crisis response. Third, organizations are reluctant to coordinate with other humanitarian aid organizations (Schulz and Heigh 2007). During a disaster, organizations must coordinate to provide prompt service to the needy. Efficient coordination ensures better handling of the natural disaster. Implementing industry 4.0 technologies in HSCs may greatly reduce the cost and increase the efficiency of different functions. After the identification of critical success factors (CSFs) from literature, we had a focused group discussion with experts to finalize these factors. Some of the factors were removed due to duplicity. Experts agreed with fifteen number of factors based on their significance These factors are summarized in Table 2.

In the present study, we have tried to bridge the research gaps in the existing literature. Major research objectives of the present study are:

(a) To identify the measures of coordination in the Humanitarian Supply Chains.
(b) To explore the applications of Industry 4.0 based technologies in different functions of humanitarian supply chains.
(c) To identify the critical success factors for the successful application of Industry 4.0 technologies in HSCs.
(d) To prioritize critical success factors from strategic perspectives for ensuring coordination in HSCs.

Table 2 Critical success factors for the application of Industry 4.0 technologies to improve coordination

S. No.	Critical success factors	References	Operational boundary and scope
T1	Presence of system architecture	Yang et al. (2011), Yan and Huang (2009), Balcik et al. (2010)	To coordinate the relief activities and rehabilitation of the end beneficiaries during and/or post-disaster period, the system architecture is required to support the permanent and temporary network established. It is required to support the use of technologies such as IoT, RFID, Sensors, etc
T2	Developing cross-functional teams	Torres (2018), Mrugalskaand Wyrwicka (2017), Daft (2015), Shamim et al. (2016)	Developing the cross-functional team is concerned with cross-functional skills of the actors to handle the various technologies and expertise required to fulfil the need of the victims. In the first phase of the disaster, it is difficult to develop or manage skilled people for humanitarian aid. Thus, the cross-functional team is required to fulfil the need like food, medicine, clothes, shelter, etc
T3	Systems integration	Kim et al. (2018), Wankmuller and Reiner (2020)	Lack of systems integration in the HSC is a major obstacle for coordination. Vertical and horizontal integration will reduce response time. From the initial to final phases of the disaster management, vertical and horizontal integration of the entire HSC is required
T4	Resource management	Pettit and Beresford (2009)	In the initial phase, HSC suffers due to lack of resources and in the later phase ineffective resource management due to a large number of entries of the donors. The collection of excessive relief items creates inventory management related problem at the site of the disaster which is termed as a second-tier disaster. Thus, resource management is crucial in both during the incident and post-disaster
T5	Accountability of the financial supply chain	Balcik et al.(2010), Agostinho (2013)	Donors always want to see the proper utilization of their contributions. They require complete transparency in the transaction. Therefore, all transactions should be on the record. The NGOs and other actors must be accountable for all the financial transactions

Table 2 (continued)

S. No.	Critical success factors	References	Operational boundary and scope
T6	Strategic planning	Daft and Marcic (2004), Kahn et al. (2012), Avison et al. (2004), Pettit and Beresford (2009)	The occurrence of the disasters are very frequent in the form of the man-made disasters, natural disasters and technological disasters. Strategic planning in HSC is required to meet the challenges. The extent of damages and way of occurrence may vary from one disaster to another. But the need of the end beneficiaries is very similar. Here, the strategic planning for better utilization of the available resources becomes very important
T7	The mechanism for coordination among organizations	Shinohara et al. (2017), Azmat and Kummer (2019), Pettit and Beresford (2009)	In a post-disaster scenario, a large number of volunteers, agencies and stakeholders respond to the disaster by being a part of rescue and relief operations. The coordination between these HSC actors becomes more challenging. There is the need for a mechanism of coordination between the organizations so the overall performance of the HSC could be improved
T8	Effective data processing and Information sharing	Walker (2014), Jabbour et al. (2017), Pettit and Beresford (2009), Dubey et al. (2019), Wagner and Thakur-Weigold (2018), Moshtari and Gonçalves (2016), Da Costa et al. (2014)	Effective data processing and information sharing are important for knowing about disaster's location, the intensity of the disasters, infrastructural damages and a reasonable estimate of causalities or affected people, so that the appropriate action may be initiated effectively
T9	Training support for the skilled workforce	Jabbour and Jabbour (2016), Fan et al. (2017), Sarkis et al. (2010), Shinohara et al. (2017), Prieto and Santana (2014), Pettit and Beresford (2009)	To implement the Industry 4.0 technologies in HSCs the employees/workforce are required to be skilled. Use of advance technologies such as IoT, cyber-physical systems, cloud computing requires a certain skill. Proper training is required to avoid the lake of skilled workforce

Table 2 (continued)

S. No.	Critical success factors	References	Operational boundary and scope
T10	Engaging the local community, NGOs, Govt. Agencies	Akhtar et al. (2012), Balcik et al. (2010)	Due to the high uncertainty of the occurrences of the disasters, planning to meet the need of the people becomes very important. In this situation, engagement of the local people, local government, NGOs, international NGOs, and military forces become more desirable
T11	Availability of real-time data	Yu et al. (2018)	When a disaster occurs, many reliable and unreliable sources of information emerges. In this situation, the application of Industry 4.0 tools to capture the real-time data becomes very important for resource management and further action plan. Reliable and real-time data are required to plan further action
T12	Traceability	Liu et al. (2013), Duan et al. (2017), Khan et al. (2018)	Traceability is real-time tracking of the location of items from origin to end. It helps in resource planning and giving prompt service to the needy
T13	Development of swift trust among stakeholders	Lewis and Weigert (1985), Hart and Saunders (1997), Dubey et al. (2019)	Mutual swift trust acts as a driving force in enhancing coordination amongst actors. Swift trust increases the willingness of the organizations to expand the amount of sharing of information and also enables them to explore more mutually beneficial arrangements
T14	Risk and needs assessment	Tjahjono et al. (2017), Akhtar et al. (2012), Azmat and Kummer (2019)	During the disaster, the efficient supply of the required items is achieved through risk and needs assessment. The organizations divide the affected areas and carry out the risk and need assessment individually and share the data collected with all to assess the situation

Table 2 (continued)

S. No.	Critical success factors	References	Operational boundary and scope
T15	Warehouse and logistics automation	Pettit and Beresford (2009), Balcik et al. (2010)	To meet the requirement of agile and flexible supply chain operations in HSC, warehouse and logistics services are to be automated to integrate with Industry 4.0 technologies. It will help in resource planning and network optimisation. Also, sometimes robotic or automated material handling is required in place of manual material handling

3 Research methodology

In this study, two sources of information are used as the input. The first source of information is the past literature through which various measures of post-disaster relief operations, critical success factors for the implementation of Industry 4.0 technologies in the humanitarian supply chain, and Industry 4.0 technologies are explored. The second source of information is the experts' opinion. The experts considered are the officials working in the NGOs, Govt. organizations, and some social volunteers working for the relief in the situation of emergency. The mapping between the measures of coordination criteria and the critical success factors is based on the opinion of these experts and the final ranking and outranking of the critical success factors are based on Analytic Hierarchy Process (AHP) and Elimination Et Choix Traduisant la Realité (ELimination Et Choice Translating Reality-ELECTRE). The Importance rating of the coordination measures is derived using AHP (Analytic Hierarchy Process) and then ELECTRE is used to outrank the less important CSF and rank the important CSF.

The AHP method is acclaimed by many researchers due to its strength in capturing qualitative and quantitative criteria. It has good traceability of the decisions and assures the quality of decisions by giving the consistency indices (Tscheikner-Gratle et al. 2017). However, AHP suffers from rank reversals in the case of addition and removal of the criteria and alternatives in the decision model. It can be prevented by prohibiting the user from the addition and removal of criteria and alternatives from the decision model (Hodgett 2016). Sabaei et al. (2015) mentioned that "the main advantage of the method is that ELECTRE can handle both quantitative and qualitative data for outranking alternatives with high uncertainty". It is also less sensitive to any changes in data compared to other methods. Considering the above fact, the combination of AHP and ELECTRE-model is found more accurate than the other individual model of ranking and it is used in this study.

Roy (1991) proposed ELECTRE as an outranking method to choose, rank, and short the various alternatives at a consultancy company (SEMA). Many applications of ELECTRE have been reported in the literature, including its fuzzy extended ELECTRE (Kumar et al. 2016, 2017a, b). The detailed procedure (Shanian and Savadogo 2006; Hashemi et al. 2014) for conducting the study is mentioned in Fig. 1.

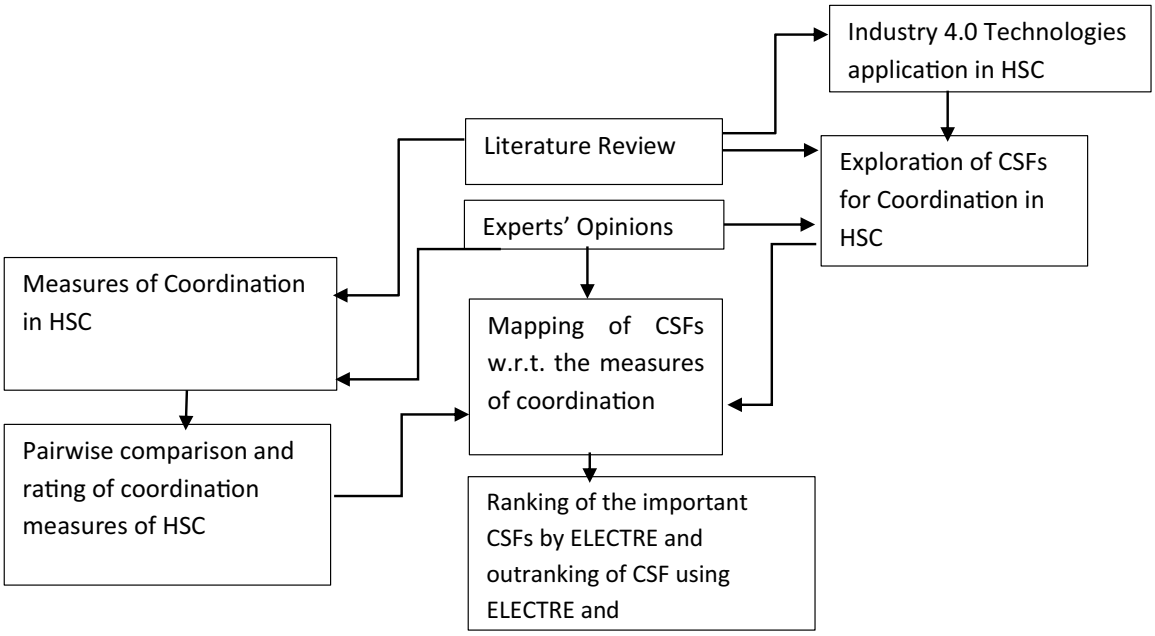


Fig. 1 A framework of the research methodology

4 Results

Experts are asked to give their opinion regarding the importance rating of the various coordination criteria/measures explored through the literature review. Pairwise comparison on a 9-point rating scale is used as input for the Analytic Hierarchy Process (AHP). For example, if criterion A is rated as 9 over criterion B, it means criterion A is extremely preferred over criterion B. Similarly, if criterion A is rated as 1/9 over criterion B, it means criterion B is extremely preferred over criterion A. Following basic steps are used to find the importance rating of the criterion.

(i) Prepare a pairwise comparison matrix C using Saaty's 9-point rating scale.

$$C = \left[c_{ij}\right]_{n\times n}$$

where c_{ij} is the importance rating of ith criterion with respect to jth criterion on the Saaty's nine-point rating scale. In this study, a total of nine (n = 9) criteria as the sustainability parameters of CE have been incorporated.

(ii) Find the normalized matrix C'. Divide each element (c_{ij}) of each column by the sum of all the elements of the corresponding column.

$$C' = \left[\frac{c_{ij}}{\sum_{i=1}^{n} c_{ij}}\right]$$

Here 'n' shows the number of criteria.

(iii) Find the weights of the criteria c_i.

$$C = \left[\sum_{j=1}^{n} \frac{c_{ij}}{\sum_{i=1}^{n} c_{ij}} \Big/ n\right] = \begin{bmatrix} c_1 \\ \cdots \\ c_i \\ \cdots \\ c_m \end{bmatrix}_{n\times 1}$$

Here, c_i denotes the importance rating of i^{th} criterion.

(iv) The consistency of the decision matrix is checked by the consistency ratio. The value of maximum eigenvalue ($\lambda_{\max}$) is required for finding the consistency index (CI). $\lambda_{\max}$ is calculated.

$$\lambda_{\max} = c_1 \sum_{i=1}^{n} c_{i1} + c_2 \sum_{i=1}^{n} c_{i2} + \cdots + c_n \sum_{i=1}^{n} c_{in}$$

(v) After finding the maximum eigenvalue of the matrix, the value of CI is calculated,

$$CI = \frac{\lambda_{\max} - n}{n - 1}$$

(vi) Finally, the consistency ratio is calculated

$$CR = \frac{CI}{RI(n)}$$

where $RI\ (n)$ is a random index, which varies with the value of n (Table 4). If the value of CR becomes greater than 0.1 then go to step 1and check the consistency of the decisions taken pairwise comparison.

List of random index values

n	2	3	4	5	6	7	8	9
RI(n)	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45

Based on AHP analysis, we found the following importance rating of the coordination measures of HSC.

C1-Improved information sharing-23.3%, C2-Flexibility in Humanitarian Operations-6.1%, C3-Service reliability-15.8%, C4-Response time-30.8%; C5-Reduced impact of diversity factors-6.4%, C6-Improved resource utilization and cost effectiveness-3.5%, C7-Logistics effectiveness-14.1%.

It can be observed that response time is the most important measure and improved information is the second most important measure of relief operations. After finding the importance rating of these measures, ELECTRE is used for the mapping of these measures with the critical success factors for the implementation of industry 4.0 technologies in the HSCs.

(i) Rate the critical success factors on a 5-point rating scale considering the various coordination measures and prepare a decision matrix as shown in Eq. (1).

$$X = \left[x_{ij}\right] \tag{1}$$

where, $1 \le i \le n$ indicates the coordination measures and, $1 \le j \le m$ indicated the critical success factors for implementing the Industry 4.0

(ii) Normalize the decision matrix using Eq. (2):

$$R = \left[r_{ij}\right] \tag{2}$$

where

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{j=1}^{m} x_{ij}^2}};$$

(iii) Multiply the elements of the normalized matrix with the weight of the corresponding criteria and prepare a weighted normalized matrix using Eq. (3). The weighted normalized matrix is shown in Table 3.

$$V = [v_{ij}] \tag{3}$$

Table 3 Weighted normalized matrix

	Coordination measures of relief operations in Humanitarian Supply Chain							
		C1	C2	C3	C4	C5	C6	C7
Critical Success Factors for implementation of Industry 4.0 in Humanitarian Supply Chain	T1	0.07116	0.01546625	0.032402	0.094781	0.020113	0.009354	0.043907
	T2	0.05693	0.01031083	0.032402	0.075825	0.015085	0.009354	0.043907
	T3	0.05693	0.01031083	0.043203	0.094781	0.015085	0.006236	0.032931
	T4	0.07116	0.01546625	0.043203	0.094781	0.015085	0.006236	0.032931
	T5	0.05693	0.02062166	0.043203	0.056869	0.010057	0.009354	0.021954
	T6	0.07116	0.01546625	0.043203	0.094781	0.015085	0.012472	0.032931
	T7	0.0427	0.01031083	0.032402	0.056869	0.020113	0.009354	0.032931
	T8	0.07116	0.01546625	0.043203	0.075825	0.015085	0.009354	0.043907
	T9	0.05693	0.01546625	0.043203	0.075825	0.010057	0.009354	0.032931
	T10	0.0427	0.01546625	0.043203	0.075825	0.015085	0.012472	0.032931
	T11	0.07116	0.01031083	0.043203	0.075825	0.025141	0.006236	0.043907
	T12	0.07116	0.01031083	0.032402	0.075825	0.020113	0.006236	0.043907
	T13	0.05693	0.02577708	0.043203	0.094781	0.015085	0.006236	0.032931
	T14	0.02847	0.01546625	0.054004	0.075825	0.015085	0.006236	0.032931
	T15	0.05693	0.02062166	0.032402	0.056869	0.015085	0.012472	0.032931

where

$$v_{ij} = r_{ij} \times w_i$$

(iv) Produce the concordance and discordance sets of coordination measures (criteria) using the conditions shown in Eqs. (4) and (5)

$$C_{kl} = \{I | v_{ik} \geq v_{il}\} \tag{4}$$

$$D_{kl} = \{I | v_{ik} < v_{il}\} = I - C_{kl} \tag{5}$$

where C_{kl} represents the set of those coordination measures (criteria k) for which the values of v_{ij} (as shown in Table 3) of k^{th} CSF is greater or equal to the values of v_{ij} of l^{th} CSF. The values v_{ij} for j^{th} critical success factor is compares for all the coordination measure ($i = k$ and l)**.** Similarly, D_{kl} shows that set of coordination measures for which the values of v_{ij} of k^{th} CSF is less than the values of v_{ij} of l^{th} CSF (i.e. complements of C_{kl}).

(v) Find the the weights of all the concordance coordination measures for each value v_{ij} and present in the form of the concordance matrix using Eq. (6).

$$C = [c_{kl}], \text{and}$$
$$c_{kl} = \sum_{i \in C_{kl}} w_i / \sum_{i=1}^{n} w_i; \text{where } 0 \leq c_{kl} \leq 1 \tag{6}$$

Here, c_{kl} is the sum of the weights of coordination measures (criteria) for which kth CSF is preferred over the lth CSF. The concordance matrix is shown in Table 4.

(vi) Find the value of d_{kl} using Eq. (7) and prepare the discordance matrix D.

$$D = [d_{kl}], \text{ and}$$
$$d_{kl} = \frac{\max\limits_{i \in D_{kl}} |v_{ik} - v_{il}|}{\max\limits_{i \in I} |v_{ik} - v_{il}|}; \text{where } 0 \leq d_{kl} \leq 1 \tag{7}$$

The numerator of Eq. (7) shows the maximum difference (numerically) between the kth and lth CSFs in the weighted normalized matrix for the coordination measure for which kth CSF is inferior to the lth CSF and denominator shows the maximum difference (numerically) between kth and lth CSFs for all the coordination measures. The discordance matrix is shown in Table 5.

(vii) Find the mean value $\bar{c}$ using Eq. (8).

$$\bar{c} = \sum_{k=1, k \neq l}^{m} \sum_{l=1, l \neq k}^{m} c_{kl} / m(m-1) \tag{8}$$

(viii) Based on the value of c_{kl}, prepare a Boolean Matrix E using Eq. (9).

$$\begin{array}{ll} e_{kl} = 1; & \text{for } c_{kl} \geq \bar{c} \\ 0; & \text{for } c_{kl} < \bar{c} \end{array} \tag{9}$$

Boolean matrix E is shown in Table 6.

(ix) Find the mean value $\bar{d}$ using Eq. (10).

Table 4 Concordance matrix

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
T1	–	1	0.842	0.842	0.781	0.807	1	0.842	0.842	0.807	0.778	1	0.781	0.842	0.904
T2	0.334	–	0.534	0.24	0.781	0.205	0.936	0.548	0.781	0.746	0.545	0.703	0.473	0.781	0.904
T3	0.466	0.824	–	0.706	0.904	0.671	0.965	0.53	0.904	0.904	0.562	0.562	0.939	0.781	0.904
T4	0.76	0.824	1	–	0.904	0.965	0.901	0.824	0.965	0.965	0.795	0.795	0.939	0.842	0.904
T5	0.254	0.487	0.487	0.254	–	0.219	0.795	0.318	0.551	0.452	0.254	0.254	0.426	0.329	0.76
T6	0.795	0.859	1	0.939	0.939	–	0.936	0.859	1	1	0.795	0.795	0.939	0.842	0.939
T7	0.257	0.318	0.301	0.24	0.548	0.205	–	0.099	0.24	0.438	0.096	0.318	0.24	0.473	0.671
T8	0.628	1	0.692	0.692	0.939	0.657	0.936	–	1	0.965	0.936	0.936	0.631	0.842	0.904
T9	0.254	0.795	0.628	0.395	0.875	0.36	0.936	0.562	–	0.901	0.562	0.562	0.567	0.778	0.84
T10	0.254	0.626	0.459	0.459	0.706	0.459	0.936	0.626	0.703	–	0.562	0.562	0.398	0.842	0.703
T11	0.596	0.965	0.692	0.631	0.904	0.596	0.965	0.904	0.904	0.904	–	1	0.631	0.781	0.904
T12	0.596	0.965	0.393	0.473	0.746	0.438	0.965	0.746	0.746	0.746	0.778	–	0.473	0.781	0.904
T13	0.527	0.825	1	0.767	0.965	0.732	0.901	0.591	0.965	0.965	0.562	0.562	–	0.842	0.965
T14	0.219	0.591	0.459	0.459	0.671	0.424	0.668	0.591	0.732	0.732	1	0.562	0.398	–	0.671
T15	0.254	0.551	0.534	0.301	0.842	0.301	0.936	0.16	0.534	0.534	0.096	0.254	0.473	0.534	–

Table 5 Discordance Matrix

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
T1	–	0	0.758	0.983	0.284	0.983	0	0.569	0.569	0.379	0.569	0	0.758	0.505	0.135
T2	1	–	1	1	0.491	1	0.265	1	1	0.758	1	1	1	0.758	0.543
T3	1	0.579	–	1	0.271	1	0.132	0.751	0.271	0.328	0.751	0.751	1	0.379	0.271
T4	1	0.579	0	–	0.135	1	0.132	0.579	0.164	0.219	0.579	0.579	0.724	0.252	0.165
T5	1	1	1	1	–	1	0.771	1	1	1	1	1	1	0.665	1
T6	1	0.579	0	0	0.135	–	0.176	0.579	0	0	0.579	0.579	0.724	0.252	0.135
T7	1	1	1	1	1	1	–	1	1	1	1	1	1	1	1
T8	1	0	1	1	0.234	1	0.176	–	0	0.109	1	0.465	1	0.252	0.164
T9	1	0.458	1	1	0.271	1	0.53	1	–	0.353	1	1	1	0.379	0.271
T10	1	1	1	1	0.75	1	0.265	1	1	–	1	1	1	0.758	0.75
T11	1	0.219	1	1	0.469	1	0.109	0.512	0.341	0.219	–	0	1	0.252	0.543
T12	1	0.219	1	1	0.491	1	0.109	1	0.758	0.379	1	–	1	0.505	0.543
T13	1	0.579	0	1	0.082	1	0.132	0.75	0.164	0.328	0.75	0.75	–	0.379	0.164
T14	1	1	1	1	1	1	0.658	1	1	1	1	1	1	–	1
T15	1	1	1	1	0.983	1	0.353	1	1	1	1	1	1	0.758	–

Table 6 Boolean matrix E

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
T1	–	1	1	1	1	1	1	1	1	1	1	1	1	1	1
T2	0	–	0	0	1	0	1	0	1	1	0	1	0	1	1
T3	0	1	–	1	1	0	1	0	1	1	0	0	1	1	1
T4	1	1	1	–	1	1	1	1	1	1	1	1	1	1	1
T5	0	0	0	0	–	0	1	0	0	0	0	0	0	0	1
T6	1	1	1	1	1	–	1	1	1	1	1	1	1	1	1
T7	0	0	0	0	0	0	–	0	0	0	0	0	0	0	0
T8	0	1	1	1	1	0	1	–	1	1	1	1	0	1	1
T9	0	1	0	0	1	0	1	0	–	1	0	0	0	1	1
T10	0	0	0	0	1	0	1	0	1	–	0	0	0	1	1
T11	0	1	1	0	1	0	1	1	1	1	–	1	0	1	1
T12	0	1	0	0	1	0	1	1	1	1	1	–	0	1	1
T13	0	1	1	1	1	1	1	0	1	1	0	0	–	1	1
T14	0	0	0	0	0	0	0	0	1	1	1	0	0	–	0
T15	0	0	0	0	1	0	1	0	0	0	0	0	0	0	–

Table 7 Boolean matrix F

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
T1	–	1	0	0	1	0	1	1	1	1	1	1	0	1	1
T2	0	–	0	0	1	0	1	0	0	0	0	0	0	0	1
T3	0	1	–	0	1	0	1	0	1	1	0	0	0	1	1
T4	0	1	1	–	1	0	1	1	1	1	1	1	0	1	1
T5	0	0	0	0	–	0	0	0	0	0	0	0	0	1	0
T6	0	1	1	1	1	–	1	1	1	1	1	1	0	1	1
T7	0	0	0	0	0	0	–	0	0	0	0	0	0	0	0
T8	0	1	0	0	1	0	1	–	1	1	0	1	0	1	1
T9	0	1	0	0	1	0	1	0	–	1	0	0	0	1	1
T10	0	0	0	0	0	0	1	0	0	–	0	0	0	0	0
T11	0	1	0	0	1	0	1	1	1	1	–	1	0	1	1
T12	0	1	0	0	1	0	1	0	0	1	0	–	0	1	1
T13	0	1	1	0	1	0	1	0	1	1	0	0	–	1	1
T14	0	0	0	0	0	0	1	0	0	0	0	0	0	–	0
T15	0	0	0	0	0	0	1	0	0	0	0	0	0	0	–

$$\bar{d} = \sum_{k=1,k\neq l}^{m} \sum_{l=1,l\neq k}^{m} d_{kl}/m(m-1) \tag{10}$$

(x) Based on the value of d_{kl}, prepare a Boolean Matrix F using Eq. (11).

$$\begin{matrix} f_{kl} = 0; & for\, d_{kl} \geq \bar{d} \\ 1; & for\, d_{kl} < \bar{d} \end{matrix} \tag{11}$$

Table 8 Aggregate dominance matrix G

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
T1	–	1	0	0	1	0	1	1	1	1	1	1	0	1	1
T2	0	–	0	0	1	0	1	0	0	0	0	0	0	0	1
T3	0	1	–	0	1	0	1	0	1	1	0	0	0	1	1
T4	0	1	1	–	1	0	1	1	1	1	1	1	0	1	1
T5	0	0	0	0	–	0	0	0	0	0	0	0	0	0	0
T6	0	1	1	1	1	–	1	1	1	1	1	1	0	1	1
T7	0	0	0	0	0	0	–	0	0	0	0	0	0	0	0
T8	0	1	0	0	1	0	1	–	1	1	0	1	0	1	1
T9	0	1	0	0	1	0	1	0	–	1	0	0	0	1	1
T10	0	0	0	0	0	0	1	0	0	–	0	0	0	0	0
T11	0	1	0	0	1	0	1	1	1	1	–	1	0	1	1
T12	0	1	0	0	1	0	1	0	0	1	0	–	0	1	1
T13	0	1	1	0	1	0	1	0	1	1	0	0	–	1	1
T14	0	0	0	0	0	0	0	0	0	0	0	0	0	–	0
T15	0	0	0	0	0	0	1	0	0	0	0	0	0	0	–

Boolean Matrix F is shown in Table 7.

(xi) Multiply the matrix E and the matrix F to find the aggregate dominance matrix G by using Eq. (12).

$$G = [g_{kl}] \tag{12}$$

where $g_{kl} = e_{kl} \times f_{kl}$

The final dominance matrix is used to find the dominance of one CSF over another. The Aggregate Dominance Matrix G is shown in Table 8.

The aggregate dominance matrix shows the dominance of the one CSFs over another. The value 1 in the matrix shows the dominance of column CSF over the row CSF. For example, the CSF T1 dominates T2, T5, T7, T8, T9, T10, T11, T12, T14, and T15. We may say that T1 outranks these CSFs. A summarized table of these outranking is prepared which is shown in Table 9.

It has been observed that the critical success factor T6 i.e. strategic planning is ranked as the most important factor of implementation of Industry 4.0 technologies in Humanitarian supply chains. Strategic planning is concerned with long-term planning to implement the Industry 4.0 technologies for enhancing coordination and collaboration among aid organizations in HSCs. It may help in gathering the quality information and communication among the actors for controlling the situation with minimum loss. Long (1997) suggested that the strategic plan can be used for the situation likely to occur in the future. For quick procurement of the items, the aid agencies can develop a strategic relationship with the suppliers (Kovacs and Spens 2007; Dignan 2005).

Resource planning is the second most important CSF for the implementation of Industry 4.0 technologies. For the implementation of Industry 4.0 technologies in the humanitarian supply chain, dedicated resources, skilled workforce for automated logistics and warehouse systems are very important. Therefore, management should focus on the training of the

Table 9 Ranking of critical success factors

CSF No	CSFs	Outranking CSFs	Ranks
T1	Presence of System Architecture	T2, T5, T7, T8, T9, T10, T11, T12, T14, T15	3
T2	Developing cross-functional teams	T5, T7, T15	8
T3	Systems integration	T2, T5, T7, T9, T10, T14, T15	6
T4	Resource planning	T2, T3, T5, T7, T8, T9, T10, T11, T12, T14, T15	2
T5	Accountability of the financial supply chain	–	OUTRANKED
T6	Strategic planning	T2, T3, T4, T5, T7, T8, T9, T10, T11, T12, T14, T15	1
T7	The mechanism for coordination among the organizations	–	OUTRANKED
T8	Effective data processing and information sharing	T2, T5, T7, T9, T10, T12, T14, T15	5
T9	Training support for a skilled workforce	T2, T5, T7, T10, T14, T15	7
T10	Engaging the local community, NGOs, Govt. agencies	T7	9
T11	Availability of real-time data	T2, T5, T7, T8, T9, T10, T12, T14, T15	4
T12	Traceability	T2, T5, T7, T10, T14, T15	7
T13	Development of swift trust among stakeholders	T2, T3, T5, T7, T9, T10, T14, T15	5
T14	Risk and needs assessment	–	OUTRANKED
T15	Warehouse and logistics automation	T7	9

workforce, technology awareness, financial support, supplier management, transportation management etc. Similarly, the third important CSF for the implementation of Industry 4.0 technologies is the presence of system architecture to support the technology. Technology like IoT (Internet of Things) enables the organization to collect and acquire information at any time and anywhere for the objects. Yang et al. (2011) have observed that the use of RFID (Radio Frequency Identification), two-dimensional bar code, camera sensor, cameras and sensor networks need to be implemented for the effective supply chain.

The fourth Important CSF is the availability of real-time data. Availability of real-time data is very crucial for planning and decisions of HSCs. This data may be collected effectively using Industry 4.0 technologies. It will help in the implementation of Industry 4.0 technologies for the digitization of the entire process to minimize the response time. The fifth important CSF is concerned with effective data processing and information sharing. After the collection of real-time data, its processing plays a crucial role in managing HSCs. Developing of swift trust among the various actors is also given equal importance. We know that many unknown actors are involved during the emergency to send relief to the sufferers as early as possible. But, handling diversity among stakeholders and beneficiaries and lack of swift trust are the major challenges in ensuring coordination in HSC. For successful operations and implementation of Industry 4.0 technologies, it is necessary to develop swift trust among all the actors engaged in HSC.

The sixth important CSF is systems integration. The systems of all stakeholders must be vertically as well as horizontally integrated for ensuring the availability of real-time data at all places. It will motivate everyone to be engaged in decision making and in planning for humanitarian operations. The next important CSF is the training of the workforce and traceability. Usually, workforces engaged in ground operations are not well equipped and skilled to use industry 4.0 technologies. Therefore, the workforce should be trained to implement the various technologies and also they should have cross-functional skills so that they can be employed in any work during the emergency. Traceability is real-time tracking of the location of items from origin to end. It is necessary for the tracking of the relief items sent to the different sites of the disasters. Developing the cross-functional team is also related to providing the training to the various actors of the HSC, which is the eighth important CSF.

Engaging the different local and international bodies like NGOs, INGOs, Govt. organizations, civil societies, and military forces are very important for the collaborative success of HSC. It is the ninth important CSF. Warehouse and logistics automation is also considered with the same importance for the HSC as it can help in supplying the relief items efficiently and effectively to the different sites of the disasters. Location of warehouses to cover the various sites of the disaster and to send the quick relief to the sufferers is also part of this CSF. Due to a different kind of uncertainties in HSCs, logistics and warehouse automation will help in the timely delivery of services to beneficiaries.

5 Discussion

Managing coordination in HSC had been a challenge. Findings of this study will help in improving the coordination in humanitarian supply chain. It has been observed that the humanitarian supply chain works in a chaotic environment and uncertainties. A large number of diversity actors works simultaneously and they have no proper linkage and interaction. In this situation, coordination among these actors becomes an important issue to

distribute relief to the end beneficiaries effectively. The theoretical and managerial implications of this study have been mentioned in the following paragraphs.

Strategic planning, resource planning and presence of System Architecture are found major critical success factors. Usually it is observed that relief items are not able to reach to needy people on time due to lack of planning for resources. Most of the organizations work in isolation so last moment coordination becomes difficult. Findings of the study, will contribute in improving coordination. Out of fifteen CSFs, three CSFs are of relatively less important and outranked by all the above mentioned CSFs. The outranked CSFs are the mechanism for coordination among the organization, risk and needs assessment, accountability of financial supply chain. Since the humanitarian supply chain for the specific disaster is temporary, the establishment of coordination among the organizations involved in the relief operations had not been the priority. The actors are not ready to coordinate on short term notice due to many constraints. After the disaster, there is a very challenging situation for managing HSC. Nothing is in order and at this time everyone tries to help the people in their way without considering the risks involved in the operations. Also, there is a lack of time of the need assessment. In this situation, the application of Industry 4.0 technologies for different humanitarian operations can help in delivering prompt service to the needy. The next outranked factor is the accountability of the financial supply chain. It is very difficult to manage the accountability of finance due to the urgent requirement of the aid without looking for the cost-effectiveness. Due to supply chain disruptions, many transporters are not ready to work in the relief operations and they demand higher charges unethically. Thus, the lack of ethics is also one of the barriers in accountability of the financial supply chain. Although three CSFs are outranked, the management should not ignore them completely.

Theoretical implications: This study may help stakeholders to correlate the critical success factors for the implementation of Industry 4.0 with the measures of coordination. Nowadays, Industry 4.0 technologies are very important to integrate the devices and data and also helps in controlling the situation from the remote location. This study may motivate the researchers to study different dimensions of coordination in HSCs in different context.

Managerial implications: In the recent past, the entire world has been facing different threats due to natural and man-made disasters. During this time, management of the humanitarian supply chain is very important to reduce the impact of a disaster. The application of Industry 4.0 technologies may be useful to improve coordination in HSCs. It may help in the efficient management of humanitarian operations and in analyzing the huge unstructured data collected from the disaster sites for long-term planning by different government agencies and NGOs. Resource planning and decision making can be made more effective. Despite many benefits of industry 4.0 technologies, implementation in HSC is not so easy. Therefore, all stakeholders including governments have to do long term planning for developing the integrated systems along with the competent workforce.

6 Conclusion

Managing post-disaster operations at a global level is the biggest challenge for decision-makers. The most important thing to reduce losses after any calamity or disaster is to ensure coordination among all stakeholders including beneficiaries. This study has tried to propose a strategic framework for the application of Industry 4.0 technologies in HSCs

for improving coordination. In this study, we have explored different measures of coordination and major critical success factors for the implementation of Industry 4.0 technologies in HSCs. The expertise required in handling the HSC operations is different from the commercial supply chain. In this study, the opinions are taken from the experts who are working for NGOs and disaster management authorities. Inputs of these experts helped in the ranking of the critical success factors for the implementation of Industry 4.0 in HSC. The importance weights of different coordination measures have been obtained using AHP. Again, the experts' opinion is used in the mapping of critical success factors with these measures on the 5-point rating scale. ELECTRE is used to rank the important CSFs and outrank the unimportant CSFs.

Strategic planning for the implementation of Industry 4.0 in HSC is the most important critical factor. It may help in formulating policies and guidelines for the implementation of technologies in different humanitarian operations. The second most important factor is the resource planning which leads to procurement and distribution of relief items. The system architecture is the third most important factor, that is concerned with the structure of the supply chain in which Industry 4.0 technologies can be easily implemented. Similarly, the availability of real-time data, effective data processing and information sharing, systems integration, training support for the workforce, traceability, developing cross-functional teams, and warehouse and logistics automation are other important factors. As sufficient research on the application of Industry 4.0 technologies in HSCs has not been done in the past, findings of this study may be very helpful in strategy formulation for application of Industry 4.0 technologies in HSCs. If the HSC operations are digitized and effective information sharing is used in the decision making, the response time can be reduced drastically and services can be given promptly to needy people.

7 Limitations and future scope

The analysis of this research is based on the inputs given by some experts working with government organizations or NGOs involved in the relief operations. Observations may also differ depending upon types of disasters and with time. Some more diversity actors can be involved in the study for their inputs and opinion. Also, stakeholders of the HSCs, who have worked for relief distribution in different types of distorters must be involved in the study. Therefore, despite significant insights from the study, findings cannot be generalized. Additionally, the authors have taken the crisp value for the analysis, which may not be able to capture the vagueness in the decision making of the experts. Some probabilistic or fuzzy model may be used for the extended work in the future. Future studies can be also done after collecting survey based data for further validation of findings.

Acknowledgements Authors would like to express sincere thanks to the Editor and reviewers for giving their constructive comments and suggestions while revising the paper.

References

Abidi, H., de Leeuw, S., & Klumpp, M. (2013). Measuring success in humanitarian supply chains. *International Journal of Business and Management Invention, 2*(8), 31–39.

Agostinho, F. (2013). Humanitarian logistics: How to help even more? *Presented at the 6th IFAC Conference on Management and Control of Production and Logistics, 6*(1), 206–210.

Akhtar, P., Marr, N. E., & Garnevska, E. V. (2012). Coordination in humanitarian relief chains: Chain coordinators. *Journal of Humanitarian Logistics and Supply Chain Management., 2*(1), 85–103.

Akoka, J., Comyn-Wattiau, I., & Laoufi, N. (2017). Research on big data-asystematic mapping study. *Computer Standards and Interfaces, 54,* 105–115.

Alexander, D. (1997). The study of natural disasters 1977–1997: Some reflections on a changing field of knowledge. *Disasters, 21*(4), 284–304.

Altay, N., & Green, W. G., III. (2006). OR/MS research in disaster operations management. *European Journal of Operational Research, 175*(1), 475–493.

Altay, N., Gunasekaran, A., Dubey, R., & Childe, S. J. (2018). Agility and resilience as antecedents of supply chain performance under moderating effects of organizational culture within the humanitarian setting: A dynamic capability view. *Production Planning and Control, 29*(14), 1158–1174.

Altay, N., & Pal, R. (2014). Information diffusion among agents: Implications for humanitarian operations. *Production and Operations Management, 23*(6), 1015–1027.

Altay, N., & Ramirez, A. (2010). Impact of disasters on firms in different sectors: Implications for supply chains. *Journal of Supply Chain Management, 46*(4), 59–80.

Arshinder, K., Kanda, A., & Deshmukh, S. G. (2008). Supply chain coordination: Perspectives, empirical studies and research directions. *International Journal of Production Economics, 115*(2), 316–335.

Avison, D., Jones, J., Powell, P., & Wilson, D. (2004). Using and validating the strategic alignment model. *The Journal of Strategic Information Systems, 13*(3), 223–246.

Azmat, M., & Kummer, S. (2019). Importance of key success factors for local and international NGOs in humanitarian supply chain. *LogForum, 15*(4), 545–555.

Balakrishnan, A., & Geunes, J. (2004). Collaboration and coordination in supply chain management and e-commerce. *Production and Operations Management, 13*(1), 1–2.

Balcik, B., & Beamon, B. M. (2008). Facility location in humanitarian relief. *International Journal of Logistics, 11*(2), 101–121.

Balcik, B., Beamon, B. M., Krejci, C. C., Muramatsu, K. M., & Ramirez, M. (2010). Coordination in humanitarian relief chains: Practices, challenges and opportunities. *International Journal of Production Economics, 126*(1), 22–34.

Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of Management, 17*(1), 99–120.

Bauer, A. (2016). *KUKA LBR Iiwa*. Augsburg: KukaRoboter GmbH.

Bealt, J., Fernández, B., Jair, C., & Mansouri, S. A. (2016). Collaborative relationships between logistics service providers and humanitarian organizations during disaster relief operations. *Journal of Humanitarian Logistics and Supply Chain Management, 6,* 118–144.

Beamon, B. M. (2004). Humanitarian relief chains: issues and challenges. In *Proceedings of 34th International Conference on Computers and Industrial Engineering, San Francisco, CA, USA*.

Beresford, A. K., & Rugamba, A. (1996). *Evaluation du secteur des transports au Rwanda*. CNUCED.

Bhattacharya, S., Hasija, S., & Van Wassenhove, L. N. (2014). Designing efficient infrastructural investment and asset transfer mechanisms in humanitarian supply chains. *Production and Operations Management, 23*(9), 1511–1521.

Blecken, A. (2010). Logistics in the context of humanitarian operations. *Advanced manufacturing and sustainable logistics* (pp. 85–93). Berlin: Springer.

Cozzolino, A. (2012). *Humanitarian logistics: Cross-sector cooperation in disaster relief management*. Berlin: Springer.

Da Costa, S. R. A., Bandeira, R. A. M., Mello, L., & Campos, V. B. G. (2014). Humanitarian supply chain: An analysis of response operations to natural disasters. *European Journal of Transport and Infrastructure Research, 14,* 290–310.

Daft, R. L. (2015). *Organization theory and design*. London: Cengage Learning.

Daft, R. L., & Marcic, D. (2004). *Understanding management* (4th ed.). Andover: Thomson.

Dai, J., Wang, S., & Yang, X. (1994). Computerized support systems for emergency decision making. *Annals of Operations Research, 51*(7), 313–325.

Dasaklis, T. K., & Pappis, C. P. (2018). Critical success factors for implementing cholera vaccination campaigns in humanitarian emergencies: A DEMATEL-based approach. *EURO Journal on Decision Processes, 6*(1–2), 1–20.

Datta, P. P., & Christopher, M. G. (2011). Information sharing and coordination mechanisms for managing uncertainty in supply chains: A simulation study. *International Journal of Production Research, 49*(3), 765–803.

De Camargo, J. A., Mendonça, P. S. M., de Oliveira, J. H. C., Jabbour, C. J. C., & de Sousa Jabbour, A. B. L. (2019). Giving voice to the silent: A framework for understanding stakeholders' participation in

socially-oriented initiatives, community-based actions and humanitarian operations projects. *Annals of Operations Research, 283*(1–2), 143–158.

Diabat, A., Jabbarzadeh, A., & Khosrojerdi, A. (2019). A perishable product supply chain network design problem with reliability and disruption considerations. *International Journal of Production Economics, 212,* 125–138.

Dignan, L. (2005). Tricky currents; tsunami relief is a challenge when supply chains are blocked by cows and roads don't exist. *Baseline, 1*(39), 30.

Drath, R., & Horch, A. (2014). Industrie 4.0: Hit or hype? [industry forum]. *IEEE Industrial Electronics Magazine, 8*(2), 56–58.

Duan, Y., Miao, M., Wang, R., Fu, Z., & Xu, M. (2017). A framework for the successful implementation of food traceability systems in China. *The Information Society, 33*(4), 226–242.

Dubey, R. (2019). Resilience and agility: the crucial properties of humanitarian supply chain. In *Handbook of Ripple Effects in the Supply Chain* (pp. 287–308). Springer, Cham.

Dubey, R., & Altay, N. (2018). Drivers of coordination in humanitarian relief supply chains. In *The Palgrave handbook of humanitarian logistics and supply chain management* (pp. 297–325). Palgrave Macmillan, London.

Dubey, R., Altay, N., & Blome, C. (2019). Swift trust and commitment: The missing links for humanitarian supply chain coordination? *Annals of Operations Research, 283*(1), 159–177.

Dubey, R., & Gunasekaran, A. (2016). The sustainable humanitarian supply chain design: Agility, adaptability and alignment. *International Journal of Logistics Research and Applications, 19*(1), 62–82.

Dubey, R., Gunasekaran, A., Altay, N., Childe, S. J., & Papadopoulos, T. (2016). Understanding employee turnover in humanitarian organizations. *Industrial and Commercial Training., 48*(4), 208–214.

Dubey, R., Luo, Z., Gunasekaran, A., Akter, S., Hazen, B. T., & Douglas, M. A. (2018). Big data and predictive analytics in humanitarian supply chains. *The International Journal of Logistics Management*.

Eckstein, D., Goellner, M., Blome, C., & Henke, M. (2015). The performance impact of supply chain agility and supply chain adaptability: The moderating effect of product complexity. *International Journal of Production Research, 53*(10), 3028–3046.

Ergun, Ö., Gui, L., HeierStamm, J. L., Keskinocak, P., & Swann, J. (2014). Improving humanitarian operations through technology-enabled collaboration. *Production and Operations Management, 23*(6), 1002–1014.

Falasca, M., & Zobel, C. (2012). An optimization model for volunteer assignments in humanitarian organizations. *Socio-Economic Planning Sciences, 46*(4), 250–260.

Fan, H., Li, G., Sun, H., & Cheng, T. C. E. (2017). An information processing perspective on supply chain risk management: Antecedents, mechanism, and consequences. *International Journal of Production Economics, 185,* 63–75.

Fawcett, A. M., & Fawcett, S. E. (2013). Benchmarking the state of humanitarian aid and disaster relief: A systems design perspective and research agenda. *Benchmarking An International Journal, 20*(5), 661–692.

Gattullo, M., Scurati, G. W., Fiorentino, M., Uva, A. E., Ferrise, F., & Bordegoni, M. (2019). Towards augmented reality manuals for industry 4.0: A methodology. *Robotics and Computer-Integrated Manufacturing, 56,* 276–286.

Gillmann, N. (2009). Interagency coordination during disaster strategic choices for the UN. NGOs, and other Humanitarian Actors in the Field. Baden-Baden: Nomos.

Global Assistance Report, (2015). Global Humanitarian Assistance Report 2015.

Goswami, S., Chakraborty, S., Ghosh, S., Chakrabarti, A., & Chakraborty, B. (2018). A review on application of data mining techniques to combat natural disasters. *Ain Shams Engineering Journal, 9*(3), 365–378.

Gunasekaran, A., Dubey, R., Wamba, S. F., Papadopoulos, T., Hazen, B. T., & Ngai, E. W. T. (2018). Bridging humanitarian operations management and organisational theory. *International Journal of Production Research, 56*(21), 6735–6740.

Gunasekaran, A., & Ngai, E. W. T. (2003). The successful management of a small logistics company. *International Journal of Physical Distribution and Logistics Management, 33*(9), 825–842.

Gustavsson, L. (2003). Humanitarian logistics: Context and challenges. *Forced Migration Review, 18*(6), 6–8.

Hart, P., & Saunders, C. (1997). Power and trust: Critical factors in the adoption and use of electronic data interchange. *Organization Science, 8*(1), 23–42.

Hashemi, S. S., Razavi Hajiagha, S. H., & Amiri, M. (2014). Decision making with unknown data: Development of ELECTRE method based on black numbers. *Informatica, 25*(1), 21–36.

Heaslip, G., Sharif, A. M., & Althonayan, A. (2012). Employing a systems-based perspective to the identification of inter-relationships within humanitarian logistics. *International Journal of Production Economics, 139*(2), 377–392.

Hodgett, R. E. (2016). Comparison of multi-criteria decision-making methods for equipment selection. *The International Journal of Advanced Manufacturing Technology, 85*(5–8), 1145–1157.

Hozdić, E. (2015). Smart factory for industry 4.0: A review. *International Journal of Modern Manufacturing Technologies, 7*(1), 28–35.

Institute, F. (2005). *Logistics and the effective delivery of humanitarian relief* (p. 12). San Francisco: Fritz Institute.

Ivanov, D., Dolgui, A., Sokolov, B., & Ivanova, M. (2017). Literature review on disruption recovery in the supply chain. *International Journal of Production Research, 55*(20), 6158–6174.

Jabbour, C. J. C., & de Sousa Jabbour, A. B. L. (2016). Green human resource management and green supply chain management: Linking two emerging agendas. *Journal of Cleaner Production, 112,* 1824–1833.

Jabbour, C. J. C., Mauricio, A. L., & Jabbour, A. B. L. D. S. (2017). Critical success factors and green supply chain management proactivity: Shedding light on the human aspects of this relationship based on cases from the Brazilian industry. *Production Planning and Control., 28*(6–8), 671–683.

Jabbour, C. J. C., Sobreiro, V. A., de Sousa Jabbour, A. B. L., de Souza Campos, L. M., Mariano, E. B., & Renwick, D. W. S. (2019). An analysis of the literature on humanitarian logistics and supply chain management: Paving the way for future studies. *Annals of Operations Research, 283,* 289–307.

Jahre, M., & Jensen, L. (2010). Coordination in humanitarian logistics through clusters. *International Journal of Physical Distribution and Logistics Management, 40*(8/9), 657–674.

Jeong, K. Y., Hong, J. D., & Xie, Y. (2014). Design of emergency logistics networks, taking efficiency, risk and robustness into consideration. *International Journal of Logistics Research and Applications, 17*(1), 1–22.

Jin, S., Jeong, S., Kim, J., & Kim, K. (2015). A logisticsmodel for the transport of disaster victims with various injuries and survival probabilities. *Annals of Operations Research, 230*(1), 17–33.

John, L., Gurumurthy, A., Soni, G., & Jain, V. (2019). Modelling the inter-relationship between factors affecting coordination in a humanitarian supply chain: A case of Chennai flood relief. *Annals of Operations Research, 283*(1), 1227–1258.

Kabra, G., & Ramesh, A. (2015). Segmenting critical factors for enhancing the use of IT in humanitarian supply chain management. *Procedia Social and Behavioral Sciences, 189,* 144–152.

Kahn, K. B., Barczak, G., Nicholas, J., Ledwith, A., & Perks, H. (2012). An examination of new product development best practice. *Journal of Product Innovation Management, 29*(2), 180–192.

Khan, S., Haleem, A., Khan, M. I., Abidi, M. H., & Al-Ahmari, A. (2018). Implementing traceability systems in specific supply chain management (SCM) through critical success factors (CSFs). *Sustainability, 10*(1), 204.

Kim, J., Pettit, S., Harris, I., & Beresford, A. (2018). Towards a better understanding of humanitarian supply chain integration. In *The Palgrave handbook of humanitarian logistics and supply chain management* (pp. 249–277). Palgrave Macmillan, London.

Kovacs, G., & Spens, K. M. (2007). Humanitarian logistics in disaster relief operations. *International Journal of Physical Distribution and Logistics Management., 37*(2), 99–114.

Kumar, P., Singh, R. K., & Kharab, K. (2017a). A comparative analysis of operational performance of cellular mobile telephone service providers in the Delhi working area using an approach of fuzzy ELECTRE. *Applied Soft Computing, 59,* 438–447.

Kumar, P., Singh, R. K., & Sinha, P. (2016). Optimal site selection for a hospital using a fuzzy extended ELECTRE approach. *Journal of Management Analytics, 3*(2), 115–135.

Kumar, P., Singh, R. K., & Vaish, A. (2017b). Suppliers' green performance evaluation using fuzzy extended ELECTRE approach. *Clean Technologies and Environmental Policy, 19*(3), 809–821.

Landherr, M., Schneider, U., & Bauernhansl, T. (2016). The application center industrie 4.0-industry-driven manufacturing, research and development. *Procedia CIRP, 57,* 26–31.

Lewis, J. D., & Weigert, A. (1985). Trust as a social reality. *Social Forces, 63*(4), 967–985.

Liu, H., Ke, W., Wei, K. K., & Hua, Z. (2013). The impact of IT capabilities on firm performance: The mediating roles of absorptive capacity and supply chain agility. *Decision Support Systems, 54*(3), 1452–1462.

Liu, Y., & Xu, X. (2017). Industry 4.0 and cloud manufacturing: A comparative analysis. *Journal of Manufacturing Science and Engineering, 139*(3), 118. https://doi.org/10.1115/1.4034667.

Long, D. (1997). Logistics for disaster relief: Engineering on the run. *IIE solutions, 29*(6), 26–30.

Long, D. C., & Wood, D. F. (1995). The logistics of famine relief. *Journal of Business Logistics, 16*(1), 213–239.

Lorenz, M., Rüßmann, M., Strack, R., Lueth, K. L., & Bolle, M. (2015). Man and machine in industry 4.0: How will technology transform the industrial workforce through 2025. *The Boston Consulting Group*, *2*.

Maghsoudi, A., Zailani, S., Ramayah, T., & Pazirandeh, A. (2018). Coordination of efforts in disaster relief supply chains: The moderating role of resource scarcity and redundancy. *International Journal of Logistics-Research and Applications, 21,* 407–430.

Maon, F., Lindgree, A., & Vanhamme, J. (2009). Developing supply chains in disaster relief operations through cross-sector socially oriented collaborations: A theoretical model. *Supply Chain Management An International Journal, 14*(2), 149–164.

Marilungo, E., Papetti, A., Germani, M., & Peruzzini, M. (2017). From PSS to CPS design: A real industrial use case toward industry 4.0. *Procedia CIRP., 64,* 357–362.

Masoni, R., Ferrise, F., Bordegoni, M., Gattullo, M., Uva, A. E., Fiorentino, M., et al. (2017). Supporting remote maintenance in industry 4.0 through augmented reality. *Procedia Manufacturing, 11,* 1296–1302.

McLachlin, R., & Larson, P. D. (2011). Building humanitarian supply chain relationships: Lessons from leading practitioners. *Journal of Humanitarian Logistics and Supply Chain Management, 1*(1), 32–49.

Moore, S., Eng, E., & Daniel, M. (2003). International NGOs and the role of network centrality in humanitarian aid operations: A case study of coordination during the 2000 Mozambique floods. *Disasters, 27*(4), 305–318.

Moshtari, M. (2016). Inter-organizational fit, relationship management capability, and collaborative performance within a humanitarian setting. *Production and Operations Management, 25*(9), 1542–1557.

Moshtari, M., & Gonçalves, P. (2016). Factors influencing interorganizational collaboration within a disaster relief context. *VOLUNTAS International Journal of Voluntary and Nonprofit Organizations, 28,* 1673–1694.

Mrugalska, B., & Wyrwicka, M. K. (2017). Towards lean production in Industry 4.0. *Procedia Engineering, 182,* 466–473.

Neugebauer, R., Hippmann, S., Leis, M., & Landherr, M. (2016). Industrie 4.0-From the perspective of applied research. 2–7.

Oloruntoba, R., & Gray, R. (2006). Humanitarian aid: An agile supply chain? *Supply Chain Management An International Journal, 11*(2), 115–120.

Papadopoulos, T., Gunasekaran, A., Dubey, R., Altay, N., Childe, S. J., & Fosso-Wamba, S. (2017). The role of big data in explaining disaster resilience in supply chains for sustainability. *Journal of Cleaner Production, 142,* 1108–1118.

Pettit, S., & Beresford, A. (2009). Critical success factors in the context of humanitarian aid supply chains. *International Journal of Physical Distribution and Logistics Management, 29*(6), 450–468.

Prieto, I. M., & Pérez-Santana, M. P. (2014). Managing innovative work behavior: The role of human resource practices. *Personnel Review, 43*(2), 184–208.

Rietjens, S. J., Voordijk, H., & De Boer, S. J. (2007). Co-ordinating humanitarian operations in peace support missions. *Disaster Prevention and Management An International Journal., 16*(1), 56–69.

Roy, B. (1991). The outranking approach and the foundations of ELECTRE methods. *Theory and decisions, 31*(1), 49–73.

Sabaei, D., Erkoyuncu, J., & Roy, R. (2015). A review of multi-criteria decision making methods for enhanced maintenance delivery. *Procedia CIRP, 37,* 30–35.

Samii, R., & Van Wassenhove, L. N. (2003). The United Nations Joint Logistics Centre: The Afghanistan crisis (No. 052003-5092). *Fontainebleau, France: INSEAD*.

Santarelli, G., Abidi, H., Klumpp, M., & Regattieri, A. (2015). Humanitarian supply chains and performance measurement schemes in practice. *International Journal of Productivity and Performance Management., 64*(60), 784–810.

Sarkis, J., Gonzalez-Torre, P., & Adenso-Diaz, B. (2010). Stakeholder pressure and the adoption of environmental practices: The mediating effect of training. *Journal of operations Management, 28*(2), 163–176.

Schlechtendahl, J., Keinert, M., Kretschmer, F., Lechler, A., & Verl, A. (2015). Making existing production systems Industry 4.0-ready. *Production Engineering, 9*(1), 143–148.

Schulz, S. F., & Heigh, I. (2007). Logistics performance management in action–design and piloting of a development indicator tool for regional logistics units of IFRC. In *Proceedings of the 1st Cardiff-Cranfield humanitarian logistics initiative conference, Shrivenham, November*.

Seaman, J. (1999). Malnutrition in emergencies: How can we do better and where do the responsibilities lie? *Disasters, 23*(4), 306–315.

Seybolt, T. B. (2009). Harmonizing the humanitarian aid network: Adaptive change in a complex system. *International Studies Quarterly, 53*(4), 1027–1050.
Shamim, S., Cang, S., Yu, H., & Li, Y. (2016, July). Management approaches for Industry 4.0: A human resource management perspective. In *2016 IEEE congress on evolutionary computation* (*CEC*) (pp. 5309–5316). IEEE.
Shanian, A., & Savadogo, O. (2006). ELECTRE I decision support model for material selection of bipolar plates for Polymer Electrolyte Fuel Cells applications. *Journal of New Materials for Electrochemical Systems, 9*(3), 191.
Shao, J., Wang, X., Liang, C., & Holguín-Veras, J. (2020). Research progress on deprivation costs in humanitarian logistics. *International Journal of Disaster Risk Reduction, 42,* 101343.
Shinohara, A. C., da Silva, E. H. D. R., de Lima, E. P., Deschamps, F., & da Costa, S. E. G. (2017). Critical success factors for digital manufacturing implementation in the context of Industry 4.0. In *IIE Annual Conference. Proceedings* (pp. 199–204). Institute of Industrial and Systems Engineers (IISE).
Simons, S., Abé, P., & Neser, S. (2017). Learning in the AutFab-the fully automated Industrie 4.0. *Learning factory of the University of Applied Sciences Darmstadt Procedia Manuf., 9,* 81–88.
Singh, R. K., Gupta, A., & Gunasekaran, A. (2018). Analysing the interaction of factors for resilient humanitarian supply chain. *International Journal of Production Research*. https://doi.org/10.1080/00207543.2018.1424373.
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* 1463–5771. https://doi.org/10.1108/bij-07-2018-0204.
Sinha, A., Kumar, P., Rana, N. P., Islam, R., & Dwivedi, Y. K. (2019). Impact of internet of things (IoT) in disaster management: A task-technology fit perspective. *Annals of Operations Research, 283*(1–2), 759–794.
Sodhi, M. S. (2016). Natural disasters, the economy and population vulnerability as a vicious cycle with exogenous hazards. *Journal of Operations Management, 45,* 101–113.
Soin, S. S. (2004). *Critical success factors in supply chain management at high technology companies* (Doctoral dissertation, University of Southern Queensland).
Srinivasan, R., & Swink, M. (2018). An investigation of visibility and flexibility as complements to supply chain analytics: An organizational information processing theory perspective. *Production and Operations Management, 27*(10), 1849–1867.
Tabaklar, T., Halldórsson, Á., & Kovács, G. (2015). Borrowing theories in humanitarian supply chain management. *Journal of Humanitarian Logistics and Supply Chain Management, 5*(3), 281–299.
Tatham, P., & Kovács, G. (2007). The humanitarian supply network in rapid onset disasters. In *Proceedings of the 19th annual conference for Nordic researchers in logistics, NOFOMA* (pp. 1059–1074).
Tatham, P., & Kovács, G. (2010). The application of "swift trust" to humanitarian logistics. *International Journal of Production Economics, 126*(1), 35–45.
Taylor, D., & Pettit, S. (2009). A consideration of the relevance of lean supply chain concepts for humanitarian aid provision. *International Journal of Services, Technology and Management, 12*(4), 430–444.
Thomas, A. (2003). Why logistics? *Forced Migration Review, 18*(4), 4.
Tjahjono, B., Esplugues, C., Ares, E., & Pelaez, G. (2017). What does industry 4.0 mean to supply chain? *Procedia Manufacturing, 13,* 1175–1182.
Tomasini, R. M., & Van Wassenhove, L. N. (2009). From preparedness to partnerships: Case study research on humanitarian logistics. *International Transactions in Operational Research, 16*(5), 549–559.
Torres Saenz, A. (2018). Identifying Challenges and success factors towards Implementing Industry 4.0 technologies in the Shipbuilding Industry. http://resolver.tudelft.nl/uuid:958e592d-0316-4ca8-9029-3c0961f63843.
Trunick, P. A. (2005). Special report: Delivering relief to tsunami victims. *Logistics today, 46*(2), 1–3.
Tscheikner-Gratl, F., Egger, P., Rauch, W., & Kleidorfer, M. (2017). Comparison of multi-criteria decision support methods for integrated rehabilitation prioritization. *Water, 9*(2), 68.
UNDP. (1993). *Logistics* (1st ed.). Geneva: United Nations Development Programme.
Vaidya, S., Ambad, P., & Bhosle, S. (2018). Industry 4.0–a glimpse. *Procedia Manufacturing, 20,* 233–238.
Van Lopik, K., Sinclair, M., Sharpe, R., Conway, P., & West, A. (2020). Developing augmented reality capabilities for industry 4.0 small enterprises: Lessons learnt from a content authoring case study. *Computers in Industry, 117,* 103208.
Van Wassenhove, L. N. (2006). Humanitarian aid logistics: Supply chain management in high gear. *Journal of the Operational Research Society, 57*(5), 475–489.
Van Wassenhove, L. N. (2017). Humanitarian aid logistics: Supply chain management in high gear. *Journal of the Operational Research Society, 57,* 475–489.

Velev, D., & Zlateva, P. (2011). Principles of cloud computing application in emergency management. *IPEDR, IACSIT Press, Singapore, 25,* 119–123.

Wagner, S. M., & Thakur-Weigold, B. (2018). Supporting collaboration in humanitarian supply chains–insights from a design science project. *Production Planning and Control, 29*(14), 1130–1144.

Wagner, S. M., Thakur-Weigold, B., Gatti, F., & Stumpf, J. (2020). Measuring and improving the impact of humanitarian logistics consulting. *Production Planning and Control*, 1–21.

Walker, S. J. (2014). Viktor mayer-schönberger and kennethcukier—big data: A revolution that will transform how we live, work, and think. *International Journal of Advertising, 33*(1), 181.

Wankmüller, C., & Reiner, G. (2020). Coordination, cooperation and collaboration in relief supply chain management. *Journal of Business Economics, 90*(2), 239–276.

Wentz, L. (2006). Information and Communication Technologies for Civil-Military Coordination in Disaster Relief and Stabilization and Reconstruction In. *Defense and Technology Paper*, *31*. Center for Technology and National Security Policy. National Defense University.

Whiting, M. C., & Ayala-Öström, B. E. (2009). Advocacy to promote logistics in humanitarian aid. *Management Research Review, 32*(11), 1081.

Whybark, D. C., Melnyk, S. A., Day, J., & Davis, E. D. (2010). Disaster relief supply chain management: New realities, management challenges, emerging opportunities. *Decision Line, 41*(3), 4–7.

Witkowski, K. (2017). Internet of things, big data, industry 4.0–innovative solutions in logistics and supply chains management. *Procedia Engineering, 182,* 763–769.

Yahyaei, M., & Bozorgi-Amiri, A. (2019). Robust reliable humanitarian relief network design: An integration of shelter and supply facility location. *Annals of Operations Research, 283*(1), 897–916.

Yan, B., & Huang, G. (2009). Supply chain information transmission based on RFID and internet of things. In *2009 ISECS International colloquium on computing, communication, control, and management* (Vol. 4, pp. 166–169). IEEE.

Yang, T., Wen, Y. F., & Wang, F. F. (2011). Evaluation of robustness of supply chain information-sharing strategies using a hybrid Taguchi and multiple criteria decision-making method. *International Journal of Production Economics, 134*(2), 458–466.

Yang, L., Yang, S. H., & Plotnick, L. (2013). How the internet of things technology enhances emergency response operations. *Technological Forecasting and Social Change, 80*(9), 1854–1867.

Yu, W., Chavez, R., Jacobs, M. A., & Feng, M. (2018). Data-driven supply chain capabilities and performance: A resource-based view. *Transportation Research Part E Logistics and Transportation Review, 114,* 371–385.

Zander, J., Mosterman, P. J., Padir, T., Wan, Y., & Fu, S. (2015). Cyber-physical systems can make emergency response smart. *Procedia Engineering, 107,* 312–318.

Zhang, Q., Zhu, C., Yang, L. T., Chen, Z., Zhao, L., & Li, P. (2017). An incremental CFS algorithm for clustering large data in industrial Internet of Things. *IEEE Transactions on Industrial Informatics, 13*(3), 1193–1201.

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

J Mater Sci: Mater Electron

Band gap and gate metal engineering of novel hetero-material InAs/GaAs-based JLTFET for improved wireless applications

Samriti Sharma[1] and Rishu Chaujar[1,*]

[1] *Department of Applied Physics, Delhi Technological University, Delhi, India*

Received: 13 August 2020
Accepted: 7 December 2020

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

ABSTRACT

This paper investigates the reliability of a dual metal gate-hetero-material junctionless tunnel FET (DMG-HJLTFET), by using a novel combination of III–V compound semiconducting materials, InAs (lower bandgap) in the source region and GaAs (higher bandgap) in the channel and drain regions. We applied bandgap engineering and dual material gate engineering to improve the linearity metrics and distortion parameters by optimizing an appropriate lower work function tunnel gate toward the source and higher work function supplementary gate toward the drain and compared all the results with SMG (single metal gate)-HJLTFET and Si-JLTFET. The DMG-HJLTFET showed marked improvements in terms of I_{ON}, I_{ON}/I_{OFF}, SS, g_m, g_{m3}, VIP2, VIP3, and 1-dB compression point. The input power, IIP3 of DMG-HJLTFET, is 158% greater than SMG-HJLTFET and is 154.7% greater than Si-JLTFET. The distortion power, IMD3 of DMG-HJLTFET, is 171.8% and 20.5% lower than SMG-HJLTFET and Si-JLTFET, respectively, thereby making it suitable for low-power distortion-free wireless communication systems.

1 Introduction

The expansion of modern wireless technology has increased the urge to downscale the dimensions of the modern communication systems. To effectuate higher speed operations and faster calculations in mobile communication industry, the present technology necessitates innovations and high-performance communication systems. For advanced systems, the progress stipulates cost-effective and highly efficient radio frequency (RF) and linearity circuit designs, which can handle weak signals. To ameliorate the signal-to-noise ratio and improve the quality of weak signals, high-power and low-noise amplifiers are the two main contenders. However, due to the higher-order harmonics in the non-linear circuits, loss of information may occur because of the cross-modulation [1–3].

For linearity and RF applications, intermodulation distortion and device linearity demand proper attention. The most critical issue in wireless application devices is power dissipation. One of the ways to

Address correspondence to E-mail: Chaujar.rishu@dtu.ac.in

https://doi.org/10.1007/s10854-020-05064-1
Published online: 01 January 2021

control power dissipation is the scaling down of supply voltage and the other way is to design transistors, which have a low subthreshold slope (SS). On downscaling the threshold voltage, the off current increases dramatically, which is theoretically related to the fact that the SS of metal oxide semiconductor field effect transistor (MOSFET) cannot be lower than 60 mV/decade [4, 5]. The MOSFET suffers from many other serious issues like static power dissipation, high leakage current, and short channel effects [6]. The Tunnel Field Effect Transistors (TFETs) are considered as a potential candidate to deal with issues related to MOSFETs due to the fundamental technique of current switching, which makes a TFET more efficient than MOSFET. A MOSFET works on the mechanism of the thermionic emission of carriers over the potential barrier, while TFET inhibits the quantum band-to-band tunneling of carriers through the barrier [7]. Therefore, a TFET proves to be the better alternative to MOSFET, attaining lower SS and low OFF current; however, lower ON current is still an issue with TFET. To resolve the aforementioned issue, a new more promising architecture named junctionless TFET (JLTFET) has been explored in previous research, which unites the benefits of a junctionless FET (higher ON current) and a TFET (steeper SS) [8, 9]. Moreover, JLTFET also enhances the immunity toward the short channel effects due to non-existence of p–n doping junctions [10, 11].

The primary mechanism of JLTFET is the charge plasma method to convert a highly doped $n^+ - n^+ - n^+$ substrate into the conventional $p^+ - i - n^+$ substrate by using two gates (polar gate and control gate) with different work functions over the source and channel region and thus reducing the fabrication complexity, scalability issues, and random doping fluctuations [11–13]. Another factor responsible for low ON current in conventional TFET is the semiconducting material. The silicon (Si) is an indirect semiconducting material with large energy bandgap, which degrades the carrier mobility and band-to-band tunneling probability. Thus, the Si-based homojunction TFET attains a lower ON current [14–16]. Many researchers have recently explored the III–V compound semiconducting materials-based heterostructure TFETs by using a lower energy gap material than Si in the source region to increase the tunneling of carriers across the junction [12, 13, 17, 18].

Against this backdrop, this work focuses on the linearity performance assessment of our proposed device, dual metal gate-hetero-material-junctionless TFET (DMG-HJLTFET), having dual metal gate engineering (DMG) applied on a junctionless TFET. We used a novel combination of III–V compound semiconducting materials Indium Arsenide (InAs)/Gallium Arsenide (GaAs) to incorporate the hetero-material bandgap engineering benefits along with junctionless configuration; this is facilitated by the direct bandgap nature and very high electron and hole mobility of InAs/GaAs at the tunneling interface. The InAs [17] with a lower work function than Si is used as the source region material and GaAs [19] with greater work function than Si is employed as the channel and drain material. The material properties of InAs and GaAs employed in our simulation framework are displayed in Table 1. Before applying the DMG, we proposed single metal gate-hetero-material-junctionless TFET (SMG-HJLTFET) device using the same materials and dimensions. We then added the DMG to study the effect of dual metal work functions on the device linearity metrics. The control gate (CG) is divided into a tunnel gate (TG) and a supplementary gate (SG) with the work function of TG being smaller than SG. In this device, TG is located near the source/channel (S/C) interface and SG is located near the channel/drain (C/D) interface. A polar gate (PG) with fixed work function is used over the source region to induce the p-type polarity. The linearity and distortion characteristics of DMG-HJLTFET are also compared with the SMG-HJLTFET and the conventional Si-JLTFET (homo-material).

Table 1 Material properties of InAs and GaAs used in our simulation [22, 24]

Material parameters	InAs	GaAs
Relative permittivity	14.6	13.2
Energy bandgap (eV)	0.35	1.42
Electron mobility (cm^2/Vs)	$\leq 4 \times 10^4$	1200
Hole mobility (cm^2/Vs)	$\leq 5 \times 10^2$	61
Electron affinity (eV)	4.9	4.07

2 Device structure

Figure 1a, b show the schematic cross-sectional view of the simulation structure of DMG-HJLTFET and SMG-HJLTFET, respectively. First, we designed the Si-JLTFET (conventional), which has the same design and channel material parameters (bandgap, lattice constant, and electron affinity) as in [20]. We then incorporated the novel amalgamation of III–V compound semiconducting hetero-materials (InAs/GaAs) and DMG to compare homo-material Si-JLTFET and hetero-material JLTFETs (SMG-HJLTFET and DMG-HJLTFET) for various result characteristics. We used a lateral n-type JLTFET with the same type of doping incorporated at the source, channel, and drain regions so that a junction is induced between two semiconductor materials at S/C interface, instead of formation of doping junctions as in regular TFET devices. Since the configuration is junctionless, the charge plasma concept is used to induce the relative carrier concentration in the source and channel regions with the help of PG and CG [13]. The doping of the body is kept uniform with concentration of $1 \times 10^{19}\text{cm}^{-3}$ and the thickness of the body (T_{BO}) is 3 nm. The length of the channel (L_{CG}) is 20 nm, while the source and the drain are kept at 15 nm length each and we kept the isolation (L_{ISO}) between PG and CG at 2 nm. The work function of PG is 5.93 eV, and the work function of CG is 4.5 eV. A high-k dielectric material Al_2O_3 with dielectric thickness (T_{OX}) of 2 nm is used. The work functions of TG (TG ($\phi_{TG} = 4.1$ eV)) and SG (SG ($\phi_{SG} = 4.5$ eV)) are chosen by tuning them for optimized linearity performance. The length of TG (L_{TG}) and SG (L_{SG}) is kept at 4 nm and 16 nm, respectively, so that the total CG length ($L_{CG} = L_{TG} + L_{SG}$) remains constant.

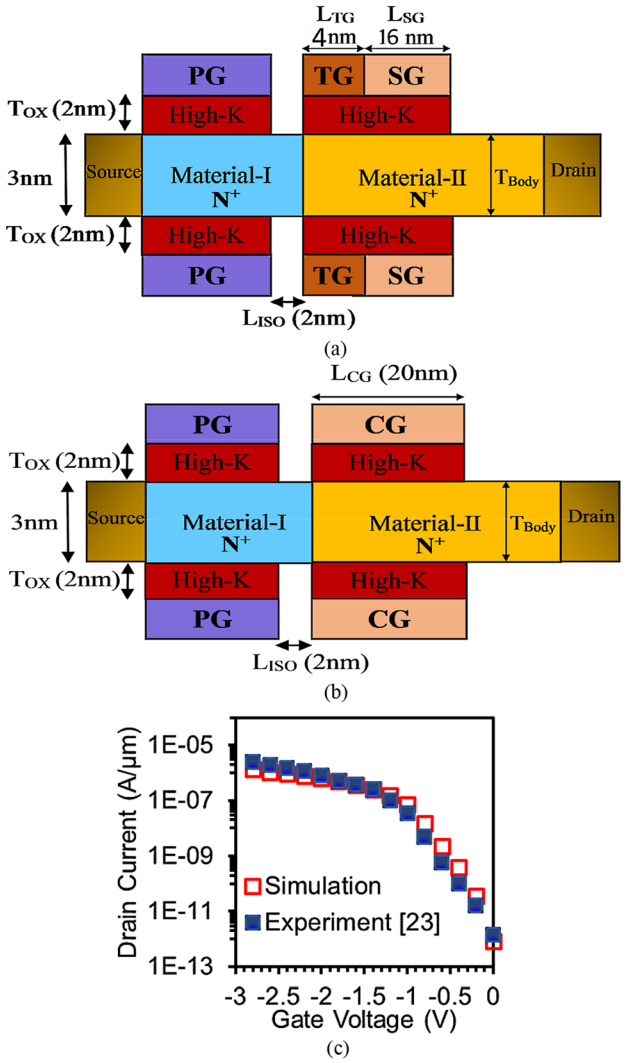


Fig. 1 Schematic structure of **a** Dual metal gate-heterojunctionless TFET (DMG-HJLTFET) and **b** Single metal gate-heterojunctionless TFET (SMG-HJLTFET). **c** Calibration of device with the published results

3 Model validation

3.1 Simulation methodology

To employ the effects of electron tunneling from valence band of the source to conduction band of the channel at S/C interface, the non-local band-to-band tunneling (BTBT) model is applied. We have also implemented the quantum tunneling meshing in the tunneling region to administer the generation rate and tunneling rate of carriers at each mesh node of the S/C interface. To include the effect of traps between the top of the valence band and the bottom of the conduction band, a trap-assisted tunneling model is applied. For simulation accuracy, the quantum confinement model given by Hansch [21] is also invoked. The bandgap narrowing model, the Shockley–Read–Hall recombination model, and the Auger recombination model are also enabled. The Fermi Dirac statistics is also invoked during the simulation. The CVT model (Lombardi) is also enabled to include concentration-dependent mobility and mobility due to parallel and perpendicular electric field and temperature. To perform all the mathematical carrier transport equations, we have employed the Gummel and Newton's numerical

method in our simulations. All the models are incorporated in the simulation setup using a TCAD simulator, Silvaco-ATLAS [22].

3.2 Calibration

As InAs/GaAs hetero-material combination has not been applied in TFET so far, we first calibrated the simulation models, specifically the most important non-local band-to-band tunneling model, with previously reported experimental work based on hetero-material SiGe, Ge, and Si for HS-TFET [23] by keeping all the device parameters and biasing conditions precisely as same as in [23] (Fig. 1c). We extracted the transfer characteristics of the experimental work reported in [23] and performed the simulation in Silvaco-ATLAS TCAD software; the two sets of results are depicted in Fig. 1c, which validate the model parameters due to a close proximity between the two set of results. After developing the calibrated simulation models, we executed them on our proposed device (SMG-HJLTFET and DMG-HJLTFET) by fine-tuning the hetero-material properties of InAs and GaAs (shown in Table 1) in the simulation models.

4 Results and discussion

Figure 2a illustrates the electrical transfer characteristics as a function of gate bias, ($V_{GS} = 1.5\text{V} \text{and} V_{DS} = 1.5\text{V}$) of our proposed device (DMG-HJLTFET) in comparison with SMG-HJLTFET and Si-JLTFET. The ON current of DMG-HJLTFET attains the highest value due to the utilization of InAs, which has the energy band gap of 0.34 eV as the source material and lower work function TG toward the S/C interface. The valence band of the source overlaps with the conduction band of the channel, which reduces the potential barrier width and hence larger electrons tunnel through the barrier ensues. It results in an exponential rise in the ON current so that DMG-HJLTFET achieves ON current 268% higher as compared to SMG-HJLTFET and 1164% higher than Si-JLTFET. The SS of three devices, DMG-HJLTFET, SMG-HJLTFET, and Si-JLTFET, is displayed in Fig. 2b, which shows a steeper SS value of DMG-HJLTFET than the Si-JLTFET due to the application of lower bandgap material toward the source and DMG. Figure 2c illustrates the comparison of ON current peak value for hetero-material tunnel FET having the CG work function equal to 4.1 eV and 4.5 eV for V_{GS} = 1.5 V and V_{DS} = 1.5 V. The results are also compared with DMG hetero-material tunnel FET which comprises of TG with work function of 4.1 eV and SG with work function of 4.5 eV.

Figure 2d, e also depict the contrast between set of results, expressed in the form of OFF current and current switching ratio (I_{ON}/I_{OFF}) for CG work function 4.1 eV and 4.5 eV, and DMG which has TG of 4.1 eV and SG of 4.5 eV. The CG with 4.1 eV work function appears to attain the highest ON current of 1.3×10^{-04} A/μm, while the CG with work function 4.5 eV attains the least ON current of 2.4×10^{-05} A/μm. When we set the CG work function to 4.1 eV, the leakage current of SMG design rises to 1.9×10^{-10} A/μm; in contrast, when the DMG design is invoked, the leakage current reduces to 2.9×10^{-16} A/μm. The DMG device (4.1 TG + 4.5 SG), appears to attain the maximum current switching ratio (3.1×10^{11}) which is 4.6×10^{5} higher than 4.1 CG SMG design and 2.2 orders higher than 4.5 CG SMG design. The reason behind the maximum current switching ratio and the lower leakage current flow in the DMG configuration is the application of a low work function TG toward the source and a high work function SG toward the drain. When we applied the gate bias of 1.5 V, the TG increases the electric field at the S/C intersection and decreases the tunneling barrier thickness and hence helps pulling down the conduction band. The valence band at the source overlaps with the conduction band at the channel. The electrons can tunnel very quickly through the junction thereby leading to higher ON current. At the no bias state, the broader energy gap material GaAs toward the channel side leads to higher tunneling barrier potential and reduction in tunneling of electrons diminishes so that low leakage current ensues.

It is critical to study the linearity parameters of a device to analyze its behavior in the high-frequency regime. The modern communication system must achieve minimum signal distortion and better linearity along with high speed. To improve the linearity of the system, there should be a uniformity of transconductance, g_m as a function of applied gate voltage. However, g_m of MOSFET and TFET is dependent on the applied gate voltage, signifying the non-linear properties of the two devices. Here, we examine the significance of using the III–V

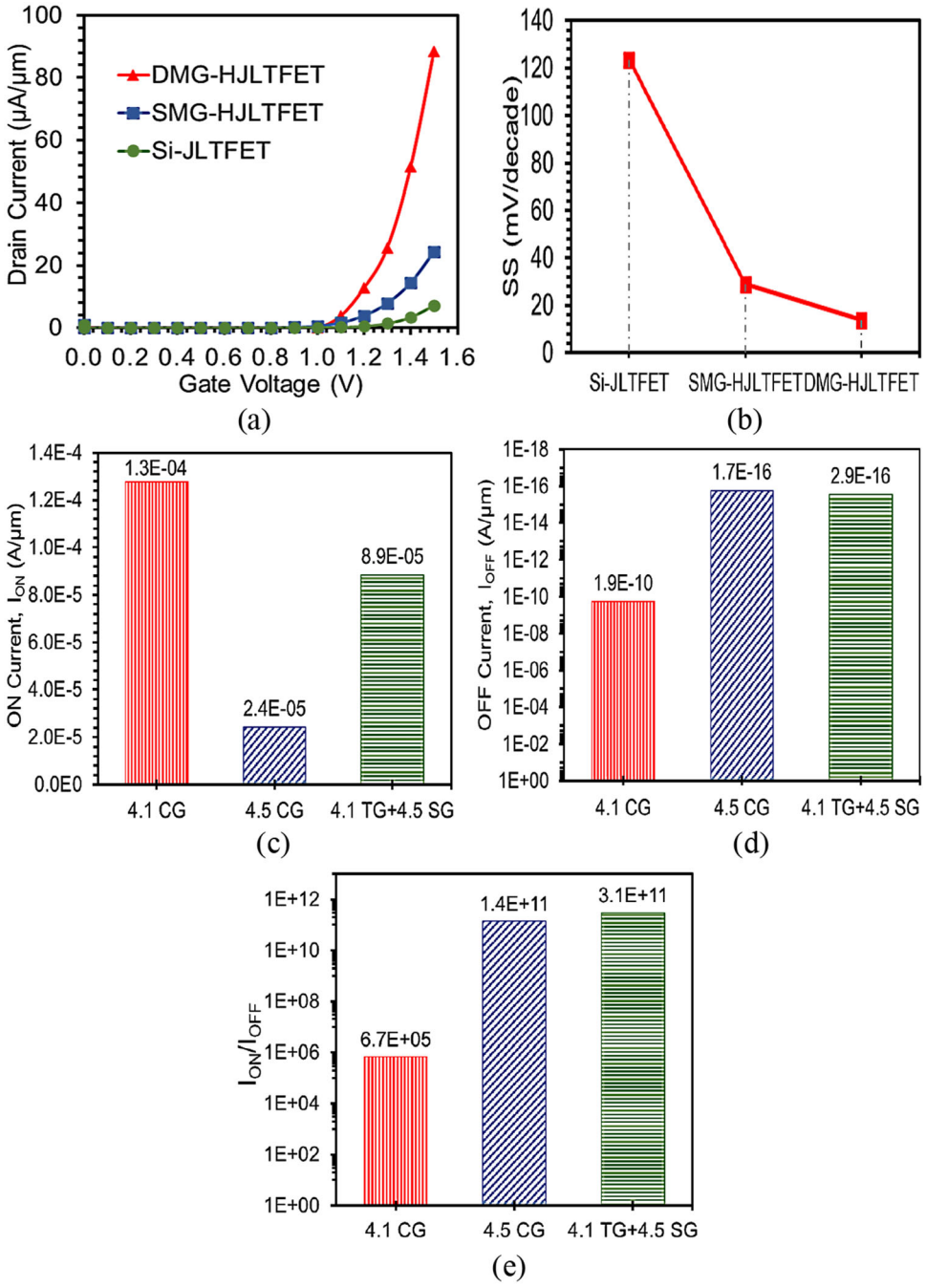


Fig. 2 **a** Transfer characteristics and **b** SS of DMG-HJLTFET, SMG-HJLTFET, and Si-JLTFET for $V_{GS} = 1.5V, V_{DS} = 1.5V$; peak value of **c** ON current, **d** OFF current, and **e** Current switching ratio (I_{ON}/I_{OFF}) for single metal gate configuration with 4.1 eV and 4.5 eV control gate work function and dual metal gate configuration consisting of 4.1 eV tunnel gate and 4.5 eV supplementary gate

compound semiconducting materials (InAs/GaAs) and DMG engineering on the linearity performance and intermodulation distortion in the form of VIP2, VIP3, IIP3, IMD3, 1-dB compression point, g_m, and g_{m3} as below:

$$g_{mn} = \frac{1}{n!}\frac{d^n I_{DS}}{dV_{GS}^n}, \quad \text{where } n = 1,2,3 \tag{1}$$

$$\text{VIP2} = 4 \times \frac{g_{m1}}{g_{m2}} \tag{2}$$

$$\text{VIP3} = \sqrt{24 \times \frac{g_{m1}}{g_{m3}}} \tag{3}$$

$$\text{IIP3} = \frac{2}{3} \times \frac{g_{m1}}{g_{m3} \times R_S} \tag{4}$$

$$\mathrm{IMD3} = R_S[4.5 \times (\mathrm{VIP3})^3 \times g_{m3}]^2 \quad (5)$$

$$1-\mathrm{dB} = 0.22\sqrt{\frac{g_{m1}}{g_{m2}}} \quad (6)$$

In Eq. (4), $R_S = 50$ is imputed owing to its practical applications in RF systems. The VIP2 is extrapolated input voltage at which first- and second-order harmonic voltages are equal and the VIP3 is extrapolated input voltage at which first- and third-order harmonic voltages are equal [25]. For high linearity, VIP2 and VIP3 must be high. The third-order intercept input power (IIP3) is defined as extrapolated input power at which the first- and third-order harmonic powers are equal. The third-order intermodulation distortion power (IMD3) is defined as the third-order intermodulation distortion corresponding to extrapolated intermodulation power at which the first- and third-order intermodulation harmonic powers are equal [25]. In Eq. (3–5), the third-order transconductance coefficient g_{m3} determines the lower limits on the distortion, and hence the amplitude of g_{m3} must be low for better linearity and low distortion applications. Equation (6) defines the 1 dB compression point which is the input power due to which the gain drops by 1 dB.

Figure 3a represents the comparison between g_m of DMG-HJLTFET, SMG-HJLTFET, and Si-JLTFET as a function of gate voltage $V_{GS} = 1.5$V. The transconductance (g_m) is an important parameter for wireless device applications, which provides the gain of the device as a function of V_{GS} and its value must be high to achieve high frequency and linearity characteristics of the device. In Fig. 3a, we found that the g_m of DMG-HJLTFET is approximately 274% higher than SMG-HJLTFET and 902% higher than Si-JLTFET with Si-JLTFET having the lowest value of g_m among the three devices. Due to the DMG configuration employed in the DMG-HJLTFET, TG comprising of the lower work function metal gives rise to the tunneling of charge carriers at the S/C intersection. It improves the current driving capability of the device, and hence the g_m increases, which also helps to decide the optimum gate bias for the device mechanism. For wireless applications, the third-order transconductance parameter is more dominant than the second-order transconductance because it leads to non-linearity due to the frequency interference and is responsible for output signal distortions. Figure 3b illustrates the variation of g_{m3} with V_{GS}. Among the three devices, DMG-HJLTFET has the lowest g_{m3} at $V_{GS} = 1.5$V and its amplitude is 9 times lower than SMG-HJLTFET and 20.6 times lower than Si-JLTFET. However, above a certain gate bias, there is a considerable variation in g_{m3} with gate voltage for DMG-HJLTFET in comparison with other devices. The large variation of g_{m3} with gate voltage in DMG-HJLTFET represents the non-linear behavior of g_{m3} due to the harmonic distortion. The increment in gate voltage beyond a certain point gives rise to harmonic distortion generated by g_{m3}, which results in reduced

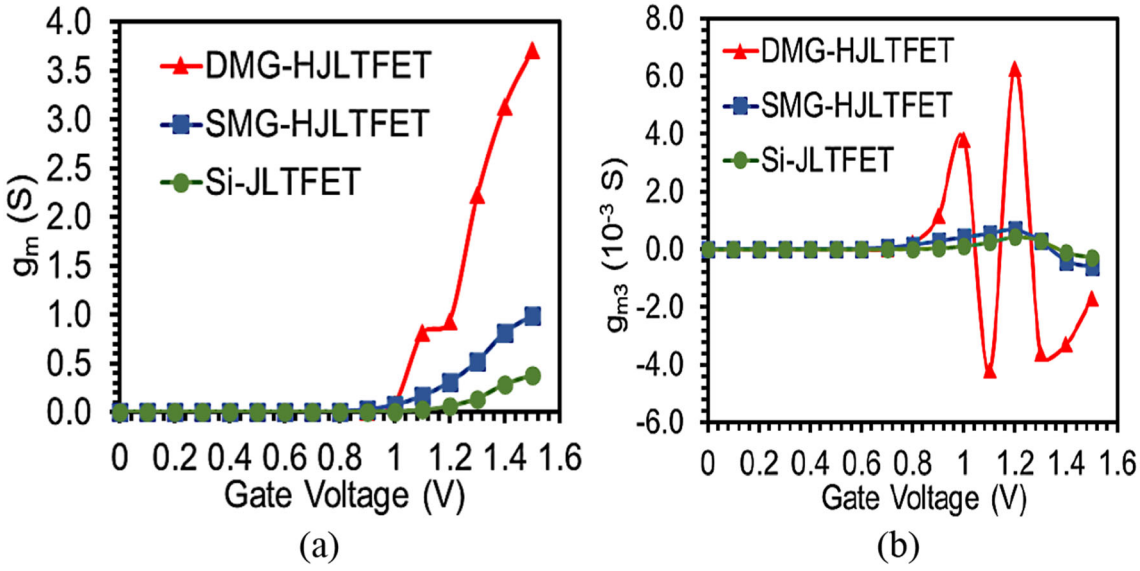


Fig. 3 **a** g_m and **b** g_{m3} of DMG-HJLTFET, SMG-HJLTFET, and Si-JLTFET

gate control over the channel; this, however, can be avoided by setting the gate bias close to g_{m3} zero-crossover point (mentioned in Fig. 7c). Because of almost symmetrical variation of g_{m3} around the zero crossover point (Fig. 3b), the harmonic distortion generated from it can be nullified for a small signal [26].

In Fig. 4a, b, DMG-HJLTFET attains maximum peak value of VIP2 and VIP3 at very low V_{GS}, which signifies that in order to preserve high linearity, low driving voltage is required. Si-JLTFET has the lowest value of the VIP3 peak among three devices. The DMG and the lower bandgap compound semiconducting material in the source and higher bandgap material in the channel and drain in our proposed device fosters higher current driving proficiency and hence leads to higher VIP2 and VIP3 peak values toward the lower gate voltage side. DMG-HJLTFET attains 220 times higher value of VIP2 than SMG-HJLTFET and 276 times higher value than Si-JLTFET and it also achieves 19.3 times higher value of VIP3 than SMG-HJLTFET and 21.9 times higher value than Si-JLTFET. Figure 5a, b signifies the disparity of VIP2 and VIP3 as a function of gate voltage for TG work function variation. The maximum peak value of VIP2 and VIP3 gets shifted toward the lower gate voltage for DMG configuration by using TG of 4.1 eV work function toward the source region, whereas when the TG work function increases to 4.3 eV, the peak values of VIP2 and VIP3 decrease. The 4.5 eV TG work function attains the lowest VIP2 and VIP3, shifting the peaks toward the higher gate voltage.

Figure 6a illustrates the IIP3 peak value of DMG-HJLTFET, SMG-HJLTFET, and Si-JLTFET at $V_{GS} = 0.6$V. IIP3 of DMG-HJLTFET is 158.4% higher than SMG-HJLTFET and 154.7% higher than Si-JLTFET. Si-JLTFET has the lowest value (− 17.33 dBm) among the three devices, which indicates the potential of DMG-HJLTFET for high linearity applications. The difference between the IIP3 peak values of DMG-HJLTFET and SMG-HJLTFET is the consequence of DMG design, as discussed before. It aids in attaining the maximum peak value of IIP3 at lower gate voltage, indicating its higher linearity attributes at low gate bias. We observe the same in Fig. 6b, where the disparity of IIP3 as a function of gate voltage is shown by varying the TG work function. On applying lower work function TG of 4.1 eV toward the source and higher work function SG of 4.5 eV toward the drain, the maximum peak value of IIP3 can be achieved at just 0.6 V gate bias. For better linearity and distortion-free operations, IIP3 needs to be high and IMD3 needs to be low. IIP3 shows approximately similar variation for TG work function 4.1 eV and 4.3 eV and when we increase the work function to 4.5 eV, the peak value is attained at 1.3 V gate voltage, which is not favorable when compared with lower work function values. It implies that the optimized value of the TG work function is 4.1 eV.

Figure 6c illustrates the disparity of IMD3 for DMG-HJLTFET, SMG-HJLTFET, and Si-JLTFET as a function of gate bias. DMG-HJLTFET exhibits the lowest value of IMD3 as − 212.9 dBm, whereas SMG-HJLTFET attains 1.24 dBm and Si-JLTFET attains 10.4

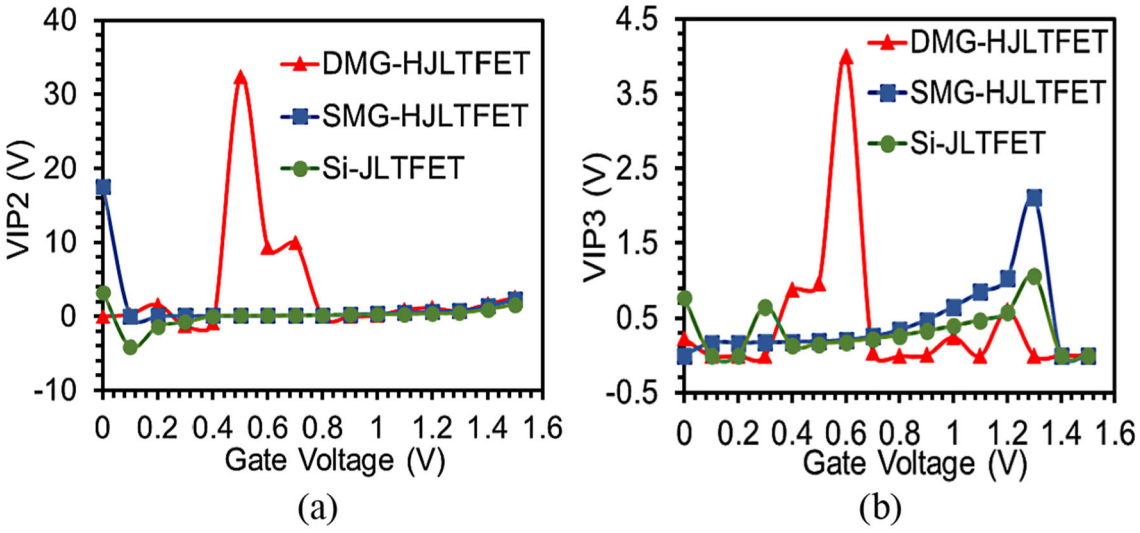


Fig. 4 Disparity of **a** VIP2 and **b** VIP3 as a function of gate voltage for DMG-HJLTFET, SMG-HJLTFET, and Si-JLTFET

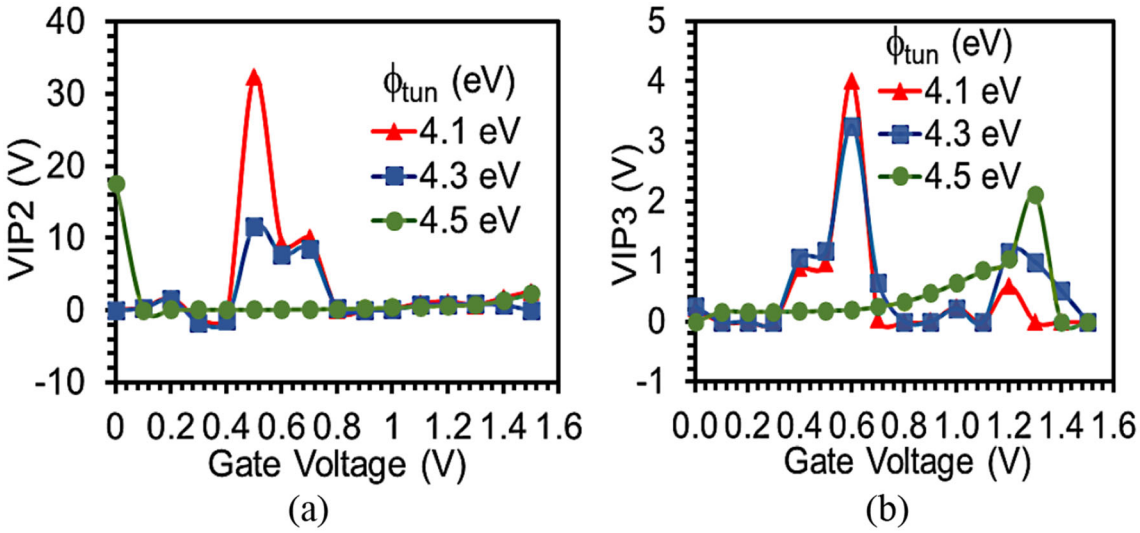


Fig. 5 Plot of **a** VIP2 and **b** VIP3 as a function of gate voltage for TG work function variation

dBm with IMD3 of DMG-HJLTFET being 171.8 orders lower than SMG-HJLTFET and 20.5 orders lower than Si-JLTFET, which is essential for a distortionless communication system. Figure 6d clarifies that the hetero-material JLTFET with 4.1 eV TG work function exhibits lowest value of IMD3 as compared to 4.3 eV and 4.5 eV TG work function. The 4.1 eV TG work function gives IMD3 approximately 16.4% lower than 4.3 eV TG and 225.8% lower than 4.5 eV TG work function. Thus, the optimized TG work function value of 4.1 eV reduces the amplitude distortion at low gate bias.

In Fig. 7a, 1-dB value is improved in the case of DMG-HJLTFET than SMG-HJLTFET and Si-JLTFET, due to the reduced signal distortion and higher value of g_m. At lower gate voltage (0.6 V), DMG-HJLTFET exhibits the maximum (25.25 dBm) value of 1 dB compression point, whereas SMG-HJLTFET exhibits 16.38 dBm as the maximum 1 dB compression point value. Si-JLTFET has the least 1 dB compression point value. Out of three devices, DMG-HJLTFET shows 54% improved value of 1 dB compression point than SMG-HJLTFET and 57% higher than Si-JLTFET. Figure 7b demonstrates the disparity of 1 dB compression point with gate voltage of 1.5 V for TG work function variation of 4.1 eV, 4.3 eV, and 4.5 eV. If we apply the 4.1 eV TG work function, (DMG), the maximum peak value of 1 dB compression point shifts toward the lower gate voltage and becomes 8.5% higher than 4.3 eV TG and 75% higher than 4.5 eV TG work function. The maximum value of 1 dB compression point is still observed at higher gate bias on changing of TG work function to 4.5 eV. Thus, the optimized value of TG work function is 4.1 eV. Figure 7c depicts the disparity of zero-crossover point for DMG-HJLTFET, SMG-HJLTFET, and Si-JLTFET. The choice of optimum gate bias point can help in suppressing the non-linearity in terms of g_{m3}, which can be determined through the zero-crossover point of g_{m3}. It is observed from Fig. 7c that in DMG-HJLTFET, the zero-crossover point gets shifted toward lower gate bias as compared to SMG-HJTFET and Si-JLTFET, due to the application of DMG. Table 2 displays the comparison of result characteristics of our proposed device with previously published hetero-material-junctionless structures based on different materials in the literature, where our proposed device appears to have promising result characteristics in terms of I_{ON}, SS, and g_m. Moreover, the bandgap and DMG engineering incorporated in DMG-HJLTFET leads to improved intermodulation distortion figure of merits (FOMs) such as VIP2, VIP3, IIP3, IMD3, and 1 dB compression point, making the transistor-level linearization more efficient and less time consuming than the expensive experimental efforts and optimization through device fabrication. These improved parameters maintain the linearity of the wireless communication systems even when receiving a weak signal, making the device highly suitable for wireless applications such as low-noise amplifiers (LNA) for cellular transceivers, space communications, and radio astronomy [27]. Moreover, the other improved performance parameters

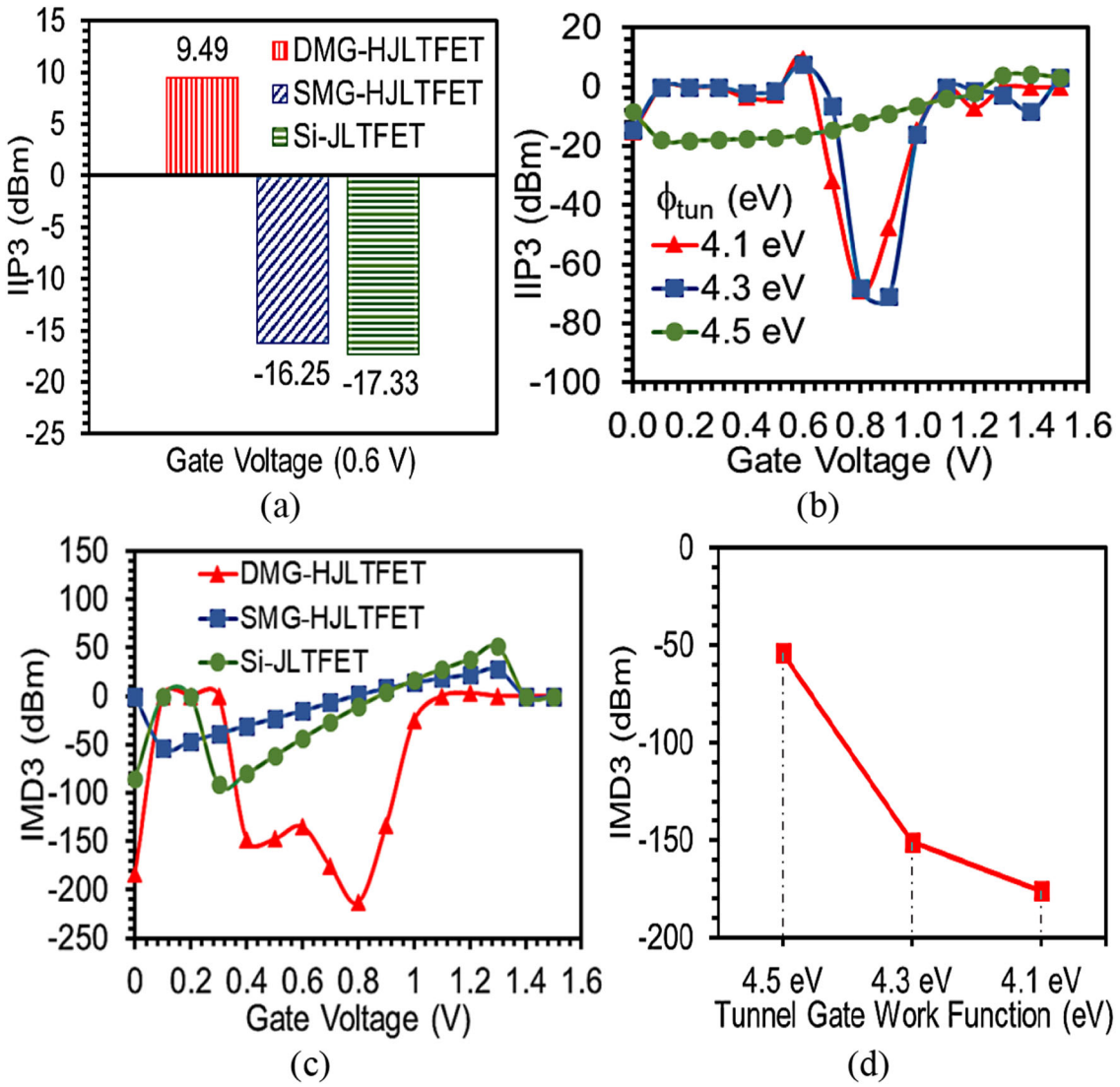


Fig. 6 Disparity of **a** IIP3 and **c** IMD3 as a function of gate bias for DMG-HJLTFET, SMG-HJLTFET, and Si-JLTFET. Disparity of **b** IIP3 and **d** IMD3 as a function of gate bias TG work function variation

such as I_{ON}, I_{ON}/I_{OFF}, I_{OFF}, steeper SS, and g_m of DMG-HJLTFET as compared to its counter devices make it desirable for fast switching performance and analog applications.

5 Conclusion

In this paper, we explored the potential of our recommended device in terms of linearity performance metrics and intermodulation distortion parameters by applying the bandgap engineering using the novel combination of compound semiconducting materials InAs/GaAs. The dual material gate engineering utilized in our proposed device improves the ON current up to 8.9×10^{-05} A/μm, which is 3.7 times higher than SMG-HJLTFET and 12.7 times higher than Si-JLTFET, whereas the current switching ratio rises to 2.2 times higher than SMG-HJLTFET and 1.7×10^5 times higher than Si-JLTFET. The implementation of lower work function TG toward the source region enhances the carrier passage efficiency and transconductance, thereby leading to more significant outcomes in terms of VIP2, VIP3, and IIP3 linearity metrics of DMG-HJLTFET in comparison to SMG-HJLTFET and Si-JLTFET. The transfer characteristics get modified because of the band overlapping at the S/C interface, which reduces the

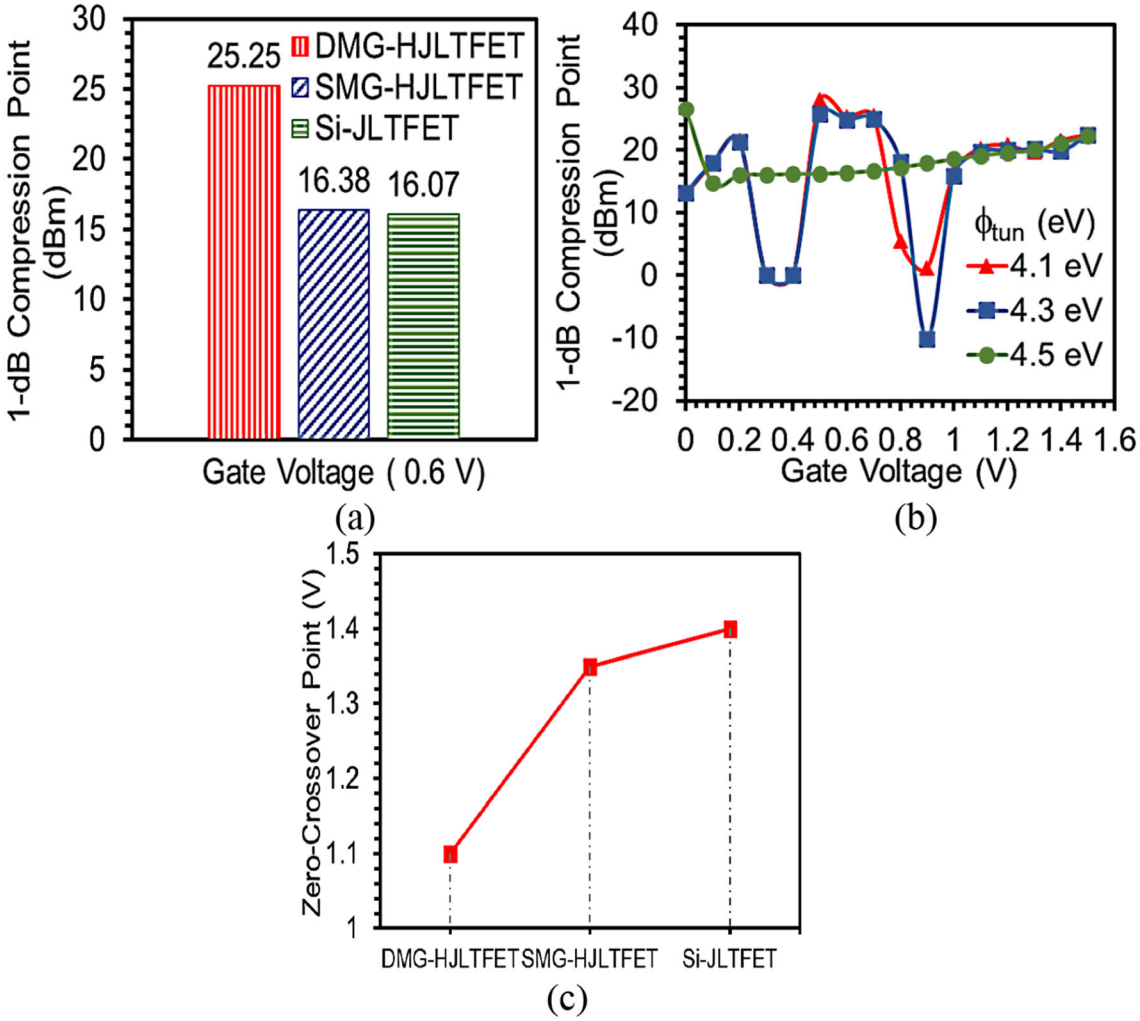


Fig. 7 Plot of 1-dB compression point as a function of gate bias for **a** DMG-HJLTFET, SMG-HJLTFET, and Si-JLTFET and **b** TG work function variation. **c** Variation of zero-crossover point with gate bias for DMG-HJLTFET, SMG-HJLTFET, and Si-JLTFET

Table 2 Comparison of this work with previously published devices

	DMMG-HJLTFET [28]	HJLTFET [20]	DMG-HJLTFET [12]	DMG-HJLTFET (this work)
I_{ON} (A/μm)	10×10^{-6}	6×10^{-6}	10×10^{-4}	88.5×10^{-6}
I_{OFF} (A/μm)	6.5×10^{-19}	1.72×10^{-18}	7.8×10^{-14}	2.89×10^{-16}
I_{ON}/I_{OFF}	1.5×10^{13}	2.3×10^{12}	1.2×10^{10}	3.1×10^{11}
SS (mV/decade	52	> 60	29	13.8
g_m (S)	1.2×10^{-5}	5.5×10^{-6}	2.8×10^{-4}	3.7×10^{-4}

tunneling barrier width and leads to the superior zero-crossover point and 1 dB compression point. The DMG-HJLTFET is found to have better g_m, g_{m3}, VIP2, VIP3, IIP3, IMD3, and 1 dB compression point in comparison with SMG-HJLTFET and Si-JLTFET. The 1 dB compression point of DMG-HJLTFET is 55% superior to SMG-HJLTFET and 57% superior to Si-JLTFET. Therefore, DMG-HJLTFET promises superior performance than SMG-HJLTFET and Si-JLTFET and can be considered an alternative in the modern wireless and distortionless communication systems.

Acknowledgements

The authors acknowledge the Microelectronics Research Lab, Delhi Technological University to carry out this research work.

Funding

No funding was received for conducting this study.

Compliance with ethical standards

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

1. R. Upasana, R. Narang, M. Saxena, M. Gupta, Linearity and analog performance realization of energy-efficient TFET-based architectures: an optimization for RFIC design. IETE Tech. Rev. **33**(1), 23–28 (2016)
2. B. Baral, S.M. Biswal, D. De, A. Sarkar, Effect of gate-length downscaling on the analog/RF and linearity performance of InAs-based nanowire tunnel FET. Int. J. Numer. Model. Electron. Netw. Devices Fields **30**(3–4), 2186 (2017)
3. H.S. Jhon, J. Jeon, M. Kang, W.Y. Choi, A sub-0.5 V operating RF low noise amplifier using tunneling-FET. Jpn. J. Appl. Phys. **56**(2), 020303 (2017)
4. H.K. Lee, W.Y. Choi, Linearity of hetero-gate-dielectric tunneling field-effect transistors. J. Semicond. Technol. Sci. **13**(6), 551–555 (2011)
5. A.M. Ionescu, H. Riel, Tunnel field-effect transistors as energy-efficient electronic switches. Nature **479**(7373), 329 (2011)
6. B. Kathy, A.M. Ionescu, Double-gate tunnel FET with high-$\kappa $ gate dielectric. IEEE Trans. Electron Devices **54**(7), 1725–1733 (2007)
7. J. Madan, R. Chaujar, Numerical simulation of N+ source pocket PIN-GAA-tunnel FET: impact of interface trap charges and temperature. IEEE Trans. Electron Devices **64**(4), 1482–1488 (2017)
8. J.P. Colinge, C.W. Lee, A. Afzalian, N.D. Akhavan, R. Yan, I. Ferain, A.M. Kelleher, Nanowire transistors without junctions. Nat. Nanotechnol. **5**(3), 225 (2010)
9. C.W. Lee, A. Afzalian, N.D. Akhavan, R. Yan, I. Ferain, J.P. Colinge, Junctionless multigate field-effect transistor. Appl. Phys. Lett. **94**(5), 053511 (2009)
10. M.J. Kumar, S. Janardhanan, Dopingless tunnel field effect transistor: design and investigation. IEEE Trans. Electron Devices **60**(10), 3285–3290 (2013)
11. B. Ghosh, M.W. Akram, Junctionless tunnel field effect transistor. IEEE Electron Device Lett. **34**(5), 584–586 (2013)
12. H. Aghandeh, S.A.S. Ziabari, Gate engineered heterostructure junctionless TFET with Gaussian doping profile for ambipolar suppression and electrical performance improvement. Superlattices Microstruct. **111**, 103–114 (2017)
13. P.K. Asthana, B. Ghosh, Y. Goswami, B.M.M. Tripathi, High-speed and low-power ultradeep-submicrometer III-V heterojunctionless tunnel field-effect transistor. IEEE Trans. Electron Devices **61**(2), 479–486 (2014)
14. K. Boucart, A.M. Ionescu, Double gate tunnel FET with high-k gate dielectric. IEEE Trans. Electron Devices **54**(7), 17251733 (2007)
15. A.C. Seabaugh, Q. Zhang, Low voltage tunnel transistors for beyond CMOS logic. Proc. IEEE **98**(12), 2095–2110 (2010)
16. J. Madan, R.S. Gupta, R. Chaujar, Analytical drain current formulation for gate dielectric engineered dual material gate-gate all around-tunneling field effect transistor. Jpn. J. Appl. Phys. **54**(9), 094202 (2015)
17. T. Vasen, P. Ramvall, A. Afzalian, G. Doornbos, M. Holland, C. Thelander, M. Passlack, Vertical gate-all-around nanowire GaSb-InAs core-shell n-type tunnel FETs. Sci. Rep. **9**(1), 1–9 (2019)
18. M. Luisier, G. Klimeck, Atomistic full-band design study of InAs band-to-band tunneling field-effect transistors. IEEE Electron Device Lett. **30**(6), 602–604 (2009)
19. M. Vadizadeh, Characteristics of GaAs/GaSb tunnel field-effect transistors without doping junctions: numerical studies. J. Comput. Electron. **17**(2), 745–755 (2018)
20. T. Han, H. Liu, S. Wang, S. Chen, W. Li, X. Yang, K. Yang, Design and Investigation of the junction-less TFET with Ge/Si0.3Ge0.7/Si heterojunction and heterogeneous gate dielectric. Electronics **8**(5), 476 (2019)
21. W. Hänsch, T. Vogelsang, R. Kircher, M. Orlowski, Carrier transport near the Si/SiO2 interface of a MOSFET. Solid-State Electron **32**(10), 839–849 (1989)
22. ATLAS User's Guide, SILVACO Int., Santa Clara, CA, USA, Version 5.26.1.R (2018)
23. Y. Zhao, C. Wu, Q. Huang, C. Chen, J. Zhu, L. Guo, R. Huang, A novel tunnel FET design through adaptive bandgap engineering with constant sub-threshold slope over 5 decades of current and high I_{ON}/I_{OFF} ratio. IEEE Electron Device Lett. **38**(5), 540–543 (2017). https://doi.org/10.1109/LED.2017.2679031
24. I. Vurgaftman, J.A. Meyer, L.A. Ram-Mohan, Band parameters for III–V compound semiconductors and their alloys. J. Appl. Phys. **89**(11), 5815–5875 (2001)

25. P. Ghosh, H. Subhasis, R.S. Gupta, M. Gupta, An investigation of linearity performance and intermodulation distortion of GME CGT MOSFET for RFIC design. IEEE Trans. Electron Devices **59**(12), 3263–3268 (2012)
26. N.B. De Carvalho, J.C. Pedro, Large-and small-signal IMD behavior of microwave power amplifiers. IEEE Trans. Microw. Theory Tech. **47**(12), 2364–2374 (1999)
27. I. Akhchaf, S. Khoulji, M. Essaidi, M.L. Kerkeb, A novel and single chip tri-band low-noise amplifier for WLAN, WIFI and WiMAX receivers. Int. J. Comput. Sci. Inform. Technol. **4**(6), 51 (2012)
28. H. Xie, H. Liu, S. Wang, S. Chen, T. Han, W. Li, Improvement of electrical performance in heterostructure junctionless TFET based on dual material gate. Appl. Sci. **10**(1), 126 (2020)

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

Clean Technologies and Environmental Policy
https://doi.org/10.1007/s10098-020-02008-5

ORIGINAL PAPER

Big data analytics application for sustainable manufacturing operations: analysis of strategic factors

Narender Kumar[1] · Girish Kumar[1] · Rajesh Kumar Singh[2]

Received: 27 May 2020 / Accepted: 4 December 2020
© The Author(s), under exclusive licence to Springer-Verlag GmbH, DE part of Springer Nature 2021

Abstract

In the present era of Industry 4.0, organizations are transforming from traditional production systems to digital production systems. This transformation is in terms of additional deployment of technologies that lead to digitization and integration of products and services, business processes and customers, etc. A high volume of unstructured data is being created across different processes due to digitization. The digitization captures the data that includes text, images, multimedia, etc., due to multiplicity of platforms, e.g., machine-to-machine communications, sensors networks, cyber-physical systems, and Internet of Things. Managing this huge data generated from different sources has become a challenging task. Big data analytics (BDA) may be helpful in managing this unstructured data for effective decision making and sustainable operations. Many organizations are struggling to integrate BDA with their manufacturing processes for sustainable operations. The application of BDA from a sustainability perspective is not extensively researched in the current literature. Therefore, firstly this study explores the contribution of BDA in sustainable manufacturing operations. It further identifies strategic factors for the successful application of BDA in manufacturing for sustainable operations. For a detailed analysis of strategic factors in manufacturing, a hybrid approach comprising the analytic hierarchy process, fuzzy TOPSIS and DEMATEL is used. Results revealed that development of contract agreement among all stakeholders, engagement of top management, capability to handle big data, availability of quality and reliable data, developing team of knowledgeable, and capable decision-makers have emerged as major strategic factors for the application of BDA in the manufacturing sector for sustainable operations. Major contribution of this study is in analyzing BDA benefits for manufacturing sector, identifying major strategic factors in implementation and categorization of these factors into cause and effect group. These findings may be used by managers as guidelines for successful implementation of BDA across different functions in their respective organization to achieve sustainable operations goal. The results of this study will also motivate industry professionals to integrate BDA with their manufacturing functions for effective decision making and sustainable operations.

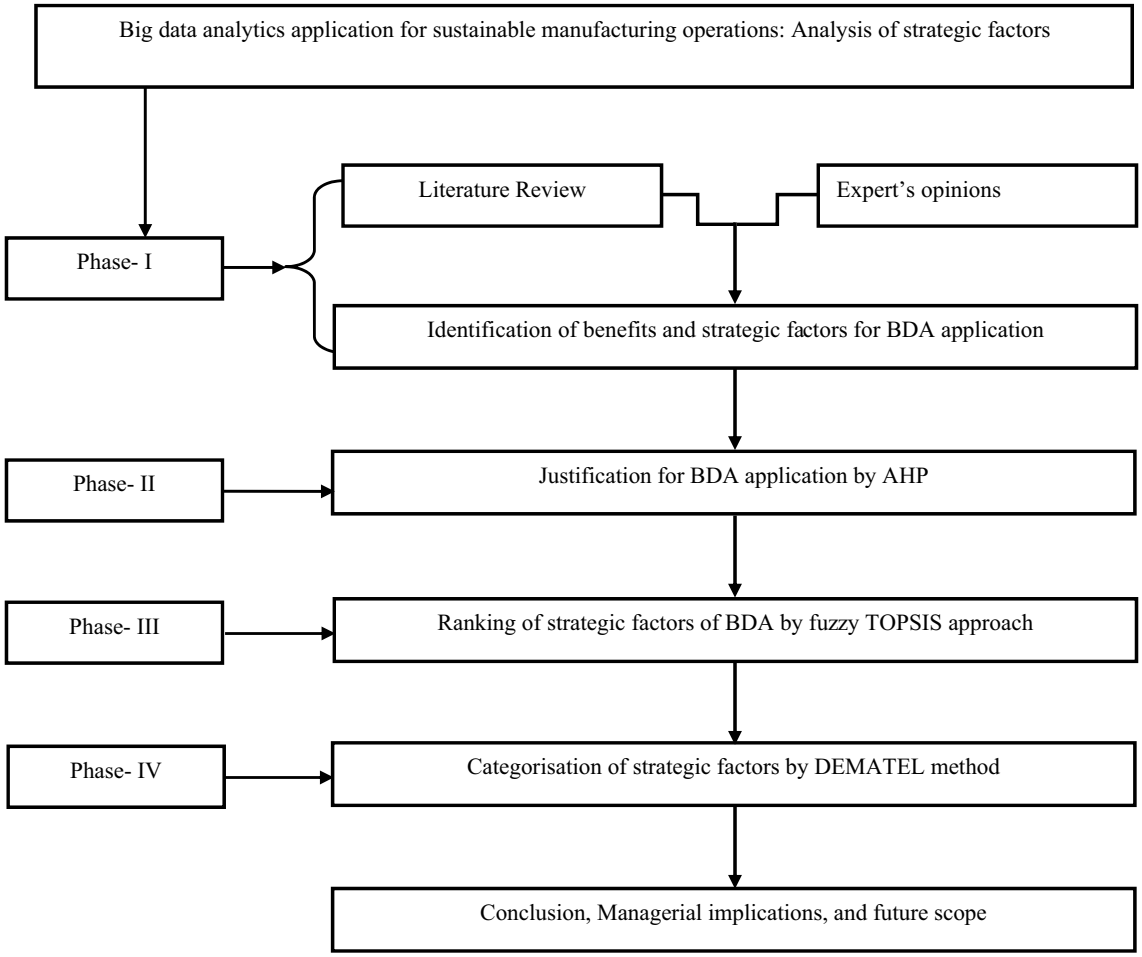
✉ Girish Kumar
girish.kumar154@gmail.com

[1] Delhi Technological University, New Delhi, India

[2] Management Development Institute, Gurgaon, India

Published online: 22 January 2021

Graphic abstract



Keywords Sustainable operations · Manufacturing sector · Big data analytics · Strategic factors

Introduction

Manufacturing sector is generating a huge amount of unstructured data due to the use of various digital machines, electronic devices, and sensors on shop floors and production lines (Zhong et al. 2015). Managing this unstructured huge data is becoming a herculean task for industry professionals. Brinch et al. (2017) have observed that big data analytics can help in streamlining of this huge data for decision making and planning of operations. Gong et al. (2018) have found that BDA applications are getting popular across different functions of the supply chain. BDA is a new age technology to manage and integrate data to improve manufacturing performance (Bi and Cochran 2014). In addition, BDA can help plant automation in the era of the fourth industrial revolution (Telukdarie et al. 2018; Tseng et al. 2019). The term BDA refers to the use of advanced analytic techniques such as applied mathematical analysis, predictive analytics, data processing, etc.. These techniques provide better insight about the processes that help in timely and right decision making and improving business processes. In the context of BDA, 5 V stands for the following characteristics of data: volume (dimension of data), velocity (flow rate of data), variety (different formats of data), veracity (uncertainty of data), and value (quality of data)" (Tao et al. 2018). In general, BDA is associated with typical data analysis and mining approaches, employed on a big volume of data. Data usually comes from various sources and needs to go through a sequence of procedures for meaningful analysis (Chen et al. 2014). BDA may be applied across the different functions of the supply chain, such as sourcing, manufacturing, distribution, and marketing (Sanders 2016).

BDA plays a crucial role in improving business benefits and operational efficiency. A business benefit is a quantifiable outcome of an activity or decision that helps meet

business goals. For example: cost savings, revenue growth, improved quality, risk reduction, etc. Application of BDA in the manufacturing sector reduces processing defects to get better quality of items produced in a manufacturing firm, saves time and money, as it aids in better and real-time decision making. The real-time data capture and instant processing through advanced algorithms in BDA will help in better and real-time decision making. It has the ability to analyze equipment failures, production bottlenecks, supply chain deficiencies, etc., that enables better decision making. BDA also gives manufacturers a better insight by identifying patterns, measuring impact, and predicting outcomes. BDA is helpful to improve product lifecycle management (Li et al. 2015). In addition, BDA is useful to other areas such as performance management, health care, and governmental services (Elgendy and Elragal 2014). BDA provides improved data-driven decision making, which lead to deeper insight into business operations and as a result better performance management. In the health industry, a large amount of data is generated to control and monitor the various processes of treatment, protection, and management of patient's medical records, and regulatory requirements. BDA helps in health care as online patient scheduling, to prescribe treatments, and make clinical decisions with greater accuracy. BDA assists the government in fighting crime, enhancing transparency, improving transportation, etc. by applying predictive analytics and machine learning to big data and providing real-time access. In addition to this, with the high volume of data from sensors and satellite images, government can forecast in advance and quickly take action to minimize losses due to natural disaster. Higher productivity goals may be also achieved if predictive analytic tools are employed for monitoring and forecasting of workers' performance. Zhong et al. (2016) has observed that BDA plays a significant role in the efficient management of the health care industry and the digitization of records. Here, the digitalization of record means all the available data in a text-searchable format to retrieve the content easily for specific information. The data can be shared instantaneously to all the relevant parties, without managing a large volume of paper documents.

BDA reduces the processing time of structured and unstructured data (Barlow 2013). Overall business performance is getting impacted by BDA as BDA-enabled organizational activities assist to lower operating costs, improve product quality, and improve product delivery (Lin et al. 2018). It helps organizations for the better forecast; as a result, organizations realize operational process benefits in the form of cost reductions, better operations planning, lower inventory levels, and elimination of waste. Thus it improves the overall performance of the organization, i.e., profitability, productivity, efficiency, etc. (Gunasekaran et al. 2017). Wamba et al. (2020) have observed that BDA can improve supply chain agility, adaptability, and operational excellence. For a manufacturing organization, BDA increases comparative advantage by strengthening the decision-making process making faster and better decisions, and increasing the capability of organizations (Dubey et al. 2016). It promotes innovation by providing the valuable information to improve the manufacturing operations and offers a means to manage environmental uncertainty; thus, it improves the overall performance of the organization.

Big data predictive analytics also facilitates in the sustainability of supply chains (Gandomi and Haider 2015). Predictive analytics tools of BDA help to get insights from data about what can happen in the future from available big data by discovering meaningful patterns of data using data mining tool like machine learning, artificial intelligence, and pattern repository system. Artificial intelligence and machine learning assist to identify inefficiencies, automate decision making, and improve the customer buying experience, which will improve sustainability of supply chain. Thus, big data predictive analysis assists in identifying and prioritizing the most important environmental and social challenges for improving the sustainability supply chain. According to Dubey et al. (2016), manufacturing organizations should adopt BDA for sustainable manufacturing operations. Continuous monitoring and analysis of operational data in real time helps in removing bottlenecks. Big data analytics helps in defect detection, predict machine failure, minimize risks, improve performance, and reduce downtime. BDA could help manufacturing firms to apply sustainable practices in more effective manner to enhance productivity. Big data support and improve sustainability measure in various operations by better management of sustainability practices of reduce, reuse and recycle. BDA also ensures lean and green manufacturing with resources optimization (Gunasekaran et al. 2017; Ji-fan Ren et al. 2017; Doolun et al. 2018). Sustainable manufacturing operations improve overall firm performance in terms of resources saving and reducing negative environmental impact (Braganza et al. 2017). Bag et al. (2020) have observed in context to the South Africa mining industry that big data analytics significantly contribute to green product development and sustainable supply chain outcomes. The environmental impact of manufacturing processes can influence the reputation of the organization in the long term (Wood et al. 2016).

The manufacturing organizations realize operational benefits in the form of cost reductions, better operations planning, lower inventory levels, better organization of the labor force and elimination of waste, while they leverage improvements in operations effectiveness and customer service as a result of BDA implementation. Popovič et al. (2018) established that the implementation of BDA reduces the consumption of materials (10–15%), energy (about 5%), scrap and rework (about 15%), and manual labor (about 20%). Also, their study revealed that implementation of BDA

reduced the maintenance and waste costs for about 12.5% on a year-to-year basis. On the other hand, particularly concerning the delivery of goods to the customer, it increases the satisfaction of customers. Companies incorporate effective decision-making process on the basis of meaningful information derived through data analytics which help them to run smarter, more agile and efficient businesses (Demirkan and Delen 2013).

Belhadi et al. (2019) have observed that researchers have not significantly explored the capabilities of BDA for sustainable manufacturing processes. Apart from it, studies on application of BDA in manufacturing sector in context to devloping countries like India are limited. Many organizastions are still working on silos and not automating their processes as per philosophy of Industry 4.0. This provides a research gap for the problem of current study, i.e., to investigate the strategic factors in implementation of I4.0 technologies such as BDA to achieve sustainable operations goal. The findings of this study will positively impact the successful implementation of BDA in manufacturing operations and will motivate the industry professionals to prioritize their resources to invest in BDA. Manufacturing organizations are facing severe challenges to adopt new technology for sustainable operations (Singh et al. 2019). Therefore, this study investigates strategic factors for implementing BDA in the manufacturing sector to make manufacturing operations more sustainable. The main objectives of this study are as follows:

- To justify the applications of BDA to ensure sustainable manufacturing operations.
- To identify strategic factors and rank them for implementing BDA in the manufacturing sector.
- To categories strategic factors in causes and effects from strategic perspective.

The structure of this paper is as follows: "Literature review" section deals with the literature review on big data applications for sustainable manufacturing operations, and the strategic factors for its implementation. In "Research methodology" section, the research methodology is discussed. "Results and discussion" section presents the obtained results and discussion. Finally, the conclusion, limitations, future scope, and managerial implications of the study are provided in "Conclusion, limitations, future scope, and managerial implications" section.

Literature review

In this section, literature review is presented on BDA applications for sustainable manufacturing operations and identification of strategic factors for BDA implementation in the manufacturing sector.

BDA applications for sustainable manufacturing operations

In the prevailing business environment of economic slowdown, manufacturers are aiming to reduce waste and improve value. Customers are looking for high-quality and low-cost products (Dubey et al. 2016; Fercoq et al. 2016). Therefore, there is a challenge in front of manufacturing organizations to meet these expectations. BDA may be a possible solution to this problem due to its many benefits. In this subsection, the authors identify the various benefits of BDA applications in the manufacturing sector from a sustainability perspective from the existing literature which are summarized in Table 1. The main sustainability benefits of BDA application in manufacturing operations are enhanced production recovery/reuse, energy-efficient and safe processes, improved customer satisfaction, improvement in profit margin, waste minimization, resource optimization, developing sustainable capabilities. Here, developing sustainable capabilities means the ability of firms to respond to their short-term financial objectives as well as future goals for profound changes in the current production systems. These benefits are further classified into three aspects, i.e., economic, social, and environmental. The benefit "Enhanced production recovery/reuse" refers to the increase in the production rate of a manufacturing system that is achieved through the effective implementation of various techniques such as lean, Kaizen, six sigma, cloud-based enterprise resource planning, etc., with the help of BDA. Reuse infers that things are utilized by the second user without prior operations or as originally designed. The energy consumption, impact on the environment, and cost, would be reduced by accurate and timely decisions taken with the support of data analytics (Hazen et al. 2016; Raut et al. 2019). "Improved customer satisfaction" is the continuous change due to optimal forecast to meet organization targets and customer requirements. Utilizing BDA, the customer may be effectively involved with green purchasing practices, cleaner production, eco-labelling, and eco-design feedback (Raut et al. 2019). For instance, optimization and machine learning have been used to select suppliers with low carbon emissions and supply chain carbon maps are generated using BDA for identifying hot spots of carbon emission so that they can be reduced (Acquaye et al. 2014; Singh et al. 2018). Additionally, customer loyalty can be improved with BDA analysis of sentiment (Dubey et al. 2016). The firm should have a BDA-based data offering structure to its customers.

Tseng et al. (2019) stated that to build successful sustainable manufacturing operations, firms should upgrade the synchronization of financial-related decisions, obtain cost information, focus on service and quality of the product, and ensure improved customer satisfaction. BDA helps in reducing costs of manufacturing processes along with final

Table 1 Big data analytics sustainability benefits for manufacturing organizations

Benefits	Description	References	Sustainability aspects		
			Economic	Social	Environmental
Enhanced production recovery and reuse (EPRR)	Refers to increase in the production rate of a manufacturing system that is achieved through the implementation of various techniques	Amui et al. (2017), Lee et al. (2015), ElMaraghy et al. (2017)	√		√
Energy-efficient and safe processes (EESP)	Reduce the amount of energy required to provide products and services	Raut et al. (2019), Wang et al. (2019), Das et al. (2020)	√	√	√
Improved customer satisfaction (ICS)	Continuous change in expected performance by accurate forecast to meet organization targets	Dubey et al. (2016), Raut et al. (2019), Gawankar et al. (2020)		√	
Improvement in profit margin (IPM)	Refer to increasing the amount of profit made from the sale	Gawankar et al. (2020),Wang et al. (2019)	√		
Waste minimization (WM)	Systematic method for the minimization of waste within a manufacturing system without sacrificing productivity, which can cause problems	Song et al. (2019), Manavalan and Jayakrishna (2019), Cui et al. (2020)	√		√
Resources optimization (RO)	Refers capacity to intricate resources in an efficient way to accomplish a sustainable goal	Amui et al. (2017), Singh and El-Kassar (2019), Song et al. (2019)	√		√
Developing sustainable capabilities (DSC)	Ability of firms to respond to their short-term financial objectives as well as future goals	Singh and El-Kassar (2019), Amui et al. (2017)	√	√	√

costs of products and components. Using predictive analytics, the manufacturer can schedule predictive maintenance that allows manufacturers to prevent costly asset breakdowns and avoid unexpected downtime which lead to reduce operational costs.

Sustainable manufacturing focuses on resource optimization without compromising with the productivity or effectiveness of manufacturing operations. Resource optimization refers to the optimal usage of available resources and to reduce carbon dioxide emissions, harmful materials, etc., from different manufacturing processes (Piyathanavong et al. 2019). Sustainable natural resource management requires complete thought of different factors with the goal that available resources may meet the requirement of contemporary society along with future generations (Mustapha et al. 2017). Sustainable capability is the ability of firms to respond to their short-term financial objectives as well as future goals. The capacity to integrate intricate resources in a successful way to accomplish sustainable goals, convey sustained values to its stakeholders, and gain sustainable competitive advantage constitute a firm's sustainable capabilities. Subsequently, by coordinating green human resources management, and green supply chain management, etc., firms can develop such sustainable capabilities that lead to enhancing performance at the ecological, environmental and social levels (Amui et al. 2017). Singh and El-Kassar (2019) have observed that organizations should have an environmental policy in place and their management should support implementing such environment-friendly practices.

Table 1 categorizes these sustainability benefits from three perspectives of the triple bottom line approach, i.e., social, economical and environmental. In manufacturing organization, minimization of manufacturing time and increasing reuse of components belong to social aspects of sustainability. Economic aspects include reduction of manufacturing cost, maintenance, and recycling. Environmental aspects refer to reduction in carbon dioxide emission, electric consumption, component packaging, and component weight (Raut et al. 2019).

Identification of strategic factors for big data analytics application

In this subsection, authors have reviewed the literature to identify strategic factors of BDA application for sustainable manufacturing. Strategic factors are defined as the attributes required to ensure overall success for an enterprise. In other words, strategic factors include issues vital to an organization's current activities. Based on the review of the literature, several strategic factors are identified. These strategic factors are listed in Table 2 with their brief description. Many organizations in developing countries are constrained with

limited access to technology, finances, infrastructure and skilled manpower. However, this is not the case with most of the organizations located in the developed countries (Kumar et al. 2020). Therefore, the stratetegic factors may be different for developing countries from the developed countries. The literature review reveals that a lot of work has been reported on the application of BDA particularly in developed countries (Mathias Kalema and Mokgadi 2017). Many organizations in developing countries are still struggling to leverage the benefits of BDA application for improving their performance from a sustainability perspective. Manufacturing organizations are showing reluctance for technological changes happening in the market with the same pace.

Research methodology

This study comprises four phases. The first phase is the identification of benefits and strategic factors for BDA application in the manufacturing sector. The second phase is the justification of BDA application in manufacturing, the third phase is the ranking of strategic factors, and the fourth phase is on the categorization of strategic factors into cause and effect. The research framework for all phases of the study is shown in Fig. 1. Based on the literature review, the benefits of BDA application and strategic factors for its implementation are identified. AHP is used for justification of the BDA application. Fuzzy TOPSIS and DEMATEL approaches are applied for the ranking and categorization of strategic factors for BDA application.

Analytic hierarchy process (AHP)

AHP-based methodology is proposed for justification of BDA application for sustainable manufacturing operations. It is a multi-criteria decision making (MCDM) approach used to solve complex decision making problems. The AHP approach was developed in 1972 (Saaty 1980). This is selected as it is easy to use and has high applicability in MCDM procedures. MCDM procedure is a methodology for decision making where various alternates are ranked on the basis of different criteria. Popular MCDM tools are AHP, TOPSIS,VIKOR, etc. AHP was used for decision making for flexible manufacturing system supply chain justification in Small and medium enterprises (SMEs), prioritizing the factors for coordinated supply chain and microalgae cultivation systems (Singh 2012, 2013; Tan et al. 2017).

As AHP is a hierarchical process, 3 levels are considered for this work. The goal of problem, benefits, and alternatives, i.e., big data-enabled manufacturing (BDM) and without big data-enabled manufacturing (WBDM). These are placed at the first, second, and third level of the hierarchy, respectively. The solution procedure passes through structural hierarchy development, construction and development of comparative judgments, and synthesis of priorities and consistency calculation.

In structural hierarchy development, the authors built analytic hierarchy model for a given problem as shown in Fig. 2. The goal of this problem is to justify the application of BDA in manufacturing sector, and it is placed at top level. The main factors in the present context are benefits of BDA application in manufacturing sector are placed on second level of hierarchy. The justification of BDA is analyzed on the basis of its benefits. At level 3 which is the last level of hierarchy, two alternatives namely big data-enabled manufacturing (BDM) and without big data-enabled manufacturing (WBDM) are positioned as these are the final outcome.

In the construction and development of comparative judgments, the priorities of elements are determined at every level. The pairwise comparisons are done in terms of which benefit dominates the other. For this purpose, responses from experts are obtained in linguistic terms and converted into crip values. A pairwise comparison matrix (nxn), P_1 for all benefits of BDA is constructed on the basis of Satty's nine-point scale (1994) as given in Appendix 1, and it is expressed as: $P_1 = \left[a_{ij}\right]_{nxn}$, where the matrix element a_{ij} represents the relative importance *i*th factor w.r.t. *j*th factor and $a_{ij*}a_{ji} = 1$ when $i \neq j$(symmetric elements are reciprocal) and $a_{ij} = 1$ when $i \equiv j$(diagonal elements are reciprocal).

Next, the matrix P_1 is normalized by dividing each entry in a column by the sum of column entries. Let b_{ij} be the normalized value of *i*th criteria with respect to *j*th criteria, and it is expressed as: $b_{ij} = \frac{a_{ij}}{\sum_{i=1}^{n} a_{ij}}$ for all $j = 1$ to n. Further, the normalized matrix (n X n) is used to obtain a priority vector matrix (principal matrix), P_2 (n X 1) by taking the average of each row elements. The matrix, P_2 is expressed as: $P_2 = [c_{ij}]_{nx1}$ where $c_{ij} = \frac{\sum_{j=1}^{n} b_{ij}}{n}$ for all $i = 1$–n. The matrix P_2 is a column vector where the element indicates the weight of each benefit.

The consistency ratio (CR) of the pairwise comparison matrix is determined to check whether the expert responses are consistent. To find the consistency ratio, P_3 and P_4 column vectors are evaluated using the matrices P_1 and P_2 as per Eq. (1) (Singh 2012).

$$P_3 = P_1 * P_2 \text{ and } P_4 = P_3 / P_2 \tag{1}$$

Next, λ_{max} is evaluated by taking average of the P_4 vector, i.e., $\lambda_{\max} = \frac{\sum_{j=1}^{n} p_j}{n}$ where p_j is the element of column vector P_4. Finally, consistency index (CI) is calculated as:

$$\text{CI} = \frac{\lambda \max - n}{n - 1}$$

where n is the size of the matrix.

Table 2 Strategic factors for big data analytics application

Strategic factors	Description	Reference
Development of contract agreement among all stakeholders	Ensures confidence of big data usage among stakeholders and defines responsibilities and procedures for better communication	Janssen et al. (2017), de Camargo Fiorini et al. (2018)
Commitment and engagement of top management	Top management will guarantee resource mobilization and its attention in BDA implementation	Gupta et al. (2018), de Camargo Fiorini et al. (2018), Dubey et al. (2016), Ivanov et al. (2019)
Development of capability for handling big data	Deployment of accurate tool and techniques for analysis, visualization, and processing of big data	Janssen et al. (2017), Yaqoob et al. (2016), Dubey et al. (2019a), Wilcox et al. (2019)
Robust cybersecurity system	Refers to the maintenance of privacy and security of big data. Its absence may lead to financial loss and damage to a firm's reputation	Lee (2017), Ivanov et al. (2019)
Coordination among big data stakeholders	Cooperation is needed to change the mindset of all stakeholders against the notions that BDA tools require huge investment and extra efforts	de Camargo Fiorini et al. (2018), Wilcox et al. (2019)
Problem identification and solving capacity	Capabilities in terms of resources including human, technology, capital, etc., to ensure conversion of inputs into higher value outputs	Janssen et al. (2017), Dubey et al. (2019a)
Process integration and institutionalization	Capacity to coordinate procedures to institutionalize assignments/information that brings improvement in the big data chain. In other words, this is an imperative condition for institutionalizing and reutilizing the utilization of big data	Janssen et al. (2017), Gupta et al. (2018)
Flexible digital infrastructure	In-house resources for data acquisition, data processes, and data analysis	Cui et al. (2020), Duan et al. (2019), Ivanov et al. (2019)
Strategy development for BDA	The process or activities that ensure the strategic planning is actually executed	Janssen et al. (2017), Gupta et al. (2018), Gupta et al. (2020)
Availability of quality and reliable big data	Ensuring adequate and correct data for the intended purpose	Janssen et al. (2017), Duan et al. (2019), Wilcox et al. (2019)
Knowledgeable and capable decision-makers	For quick and better decision, the experienced decision-makers are required, who easily understand and analyze the big data	Del Fabbro and Santarossa (2016), Cui et al. (2020)
Data-driven organization culture	Changing a culture is a vast scale undertaking, and in the end, the majority of the authoritative apparatuses for changing personalities should be placed in play	Janssen et al. (2017), Duan et al. (2019), Weerakkody et al. (2017), Cui et al. (2020)
Process monitoring and control	Hardware and software system related to data processing	Dubey et al. (2019b), Duan et al. (2019)
Integrating customers' requirements with performance framework	Data help to improve the customer relationship with the performance of the industry, decision making capability and reduce fraud	Dubey et al. (2019a), Janssen et al. (2017), Gupta et al. (2020)
Responsive information sharing framework	The relation between information system framework and industry performance is an evaluation of the roundabout advantages of information technology	Duan et al. (2019), Dubey et al. (2019b)

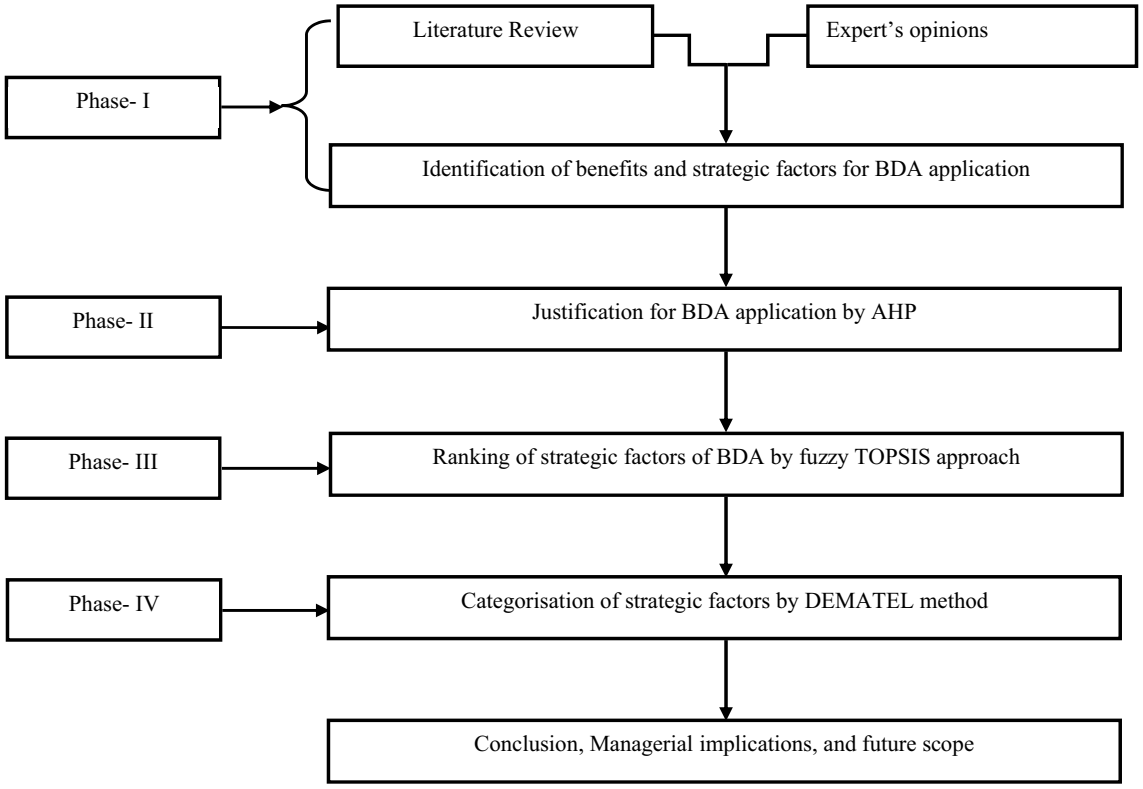


Fig.1 Proposed research framework

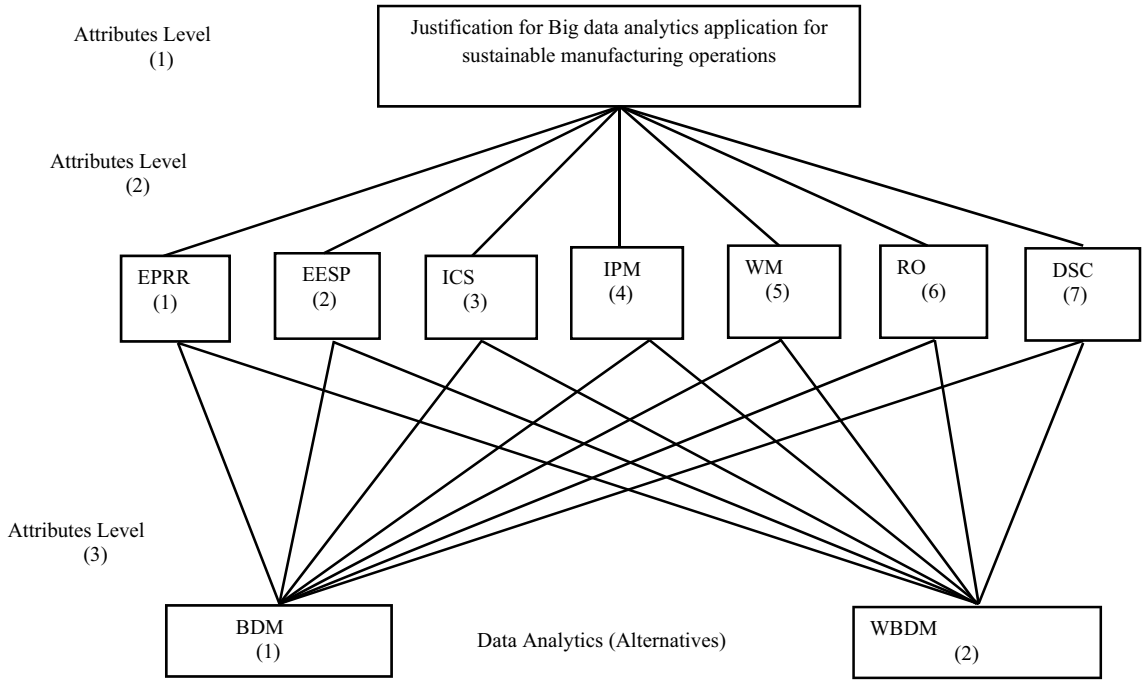


Fig. 2 Schematic of the analytic hierarchy process

The ratio of consistency index (CI) to random consistency index (RCI) is known as the consistency ratio which is expressed as:

$$\mathrm{CR} = \mathrm{CI}/\mathrm{RCI} \tag{3}$$

where RCI is random consistency index and RCI value taken as per Appendix 2.

If CI is less than 0.1 decisions are considered as consistent. For CR value more than 0.1, the nature of decisions ought to be revised till CR value reaches in a consistent range.

The acceptable CR depends on the size of the matrix, and it is 0.1 for matrix size 4 × 4 and larger (Saaty 2000). If the value of consistency ratio is equal to or less than the permissible value, it suggests that the assessment within the matrix is satisfactory or shows a good level of consistency in the relative decisions. Similar procedure is followed for last hierarchy for computing the weights of BDM and WBDM for each benefit. The list of variables/symbols used in the expressions/ equations in this work is given in Appendix 9.

Fuzzy TOPSIS approach

Fuzzy TOPSIS is used for the ranking of strategic factors for BDA application for sustainable manufacturing operations. Fuzzy TOPSIS was employed for analyzing disposition strategies in reverse supply chains, disposition decision in reverse logistics, prioritizing strategic factors for reverse logistics, and environmental sustainability (Agrawal et al. 2016; Singh and Agrawal 2018; Samaie et al. 2020). This approach was also implemented to support outsourcing of logistics service, evaluation of 3PL in a supply chain, and supplier selection (Kumar and Singh 2012; Lima Junior et al. 2014).

In the fuzzy TOPSIS method, the data is collected in linguistic terms (later converted on a fuzzy scale) for selected alternatives for the chosen criteria, unlike the TOPSIS approach where crisp values are gathered. Fuzzy TOPSIS is a simple, realistic form of modeling and compensatory method which includes and excludes alternative solutions based on hard cutoff (Singh and Agrawal 2018). Additionally, it is a computation process that can be easily programmed into a spreadsheet that contains a data on a scalar value that represents both the best and worst alternatives at the same time, a sound logic that represents the rationale of human choice and the performance measures of all alternatives can be visualized on polyhedron (Kim et al. 1997). Moreover, the integration of fuzzy will further increases its strength as it is capable of handling vague and uncertain information (Zimmermann 1985). These benefits of fuzzy TOPSIS make it a better choice among MCDM approaches. Therefore, this approach is employed for this work, and the step-by-step procedure of fuzzy TOPSIS is detailed below:

Step 1 Collect the data through the survey method in the form of linguistics terms. The experts should be asked to select the best option. The options are expressed in linguistic terms for a given question. A 5-point scale with the linguistic terms low (L), fairly low (FL), medium (M), fairly high (FH), and high (H) is generally used in the questionnaire. Once the data is collected in linguistic terms, the same is converted into fuzzy numbers.

Step 2 A fuzzy decision matrix is derived based on the data collected in step 1 and converted into triangular fuzzy numbers.

$$D = \begin{bmatrix} Y_{11} & Y_{12} & \dots & Y_{1j} & \dots & Y_{1n} \\ Y_{21} & Y_{22} & \dots & Y_{2j} & \dots & Y_{21} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ Y_{i1} & Y_{i2} & \dots & Y_{ij} & \dots & Y_{in} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ Y_{m1} & Y_{m2} & \dots & Y_{mj} & \dots & Y_{mn} \end{bmatrix}$$

where $Y_{ij} = (d_{ij}, e_{ij}, f_{ij})$ is a triangular fuzzy number for the linguistic term allocated by the *i*th respondent to the *j*th factor. i = 1, 2, …, m are the number of respondents ,and $j = 1$, 2, …, n are the number of factors (Strategic factors). Table 3 shows the scale in terms of triangular fuzzy numbers used for each linguistic term.

Step 3 A fuzzy decision matrix (D) is converted into a fuzzy un-weighted matrix (R) using the following relationship (Singh and Agrawal 2018).

$$R = [r_{ij}]_{m\times n;} = \left(\frac{d_{ij}}{c_{j^*}^*}, \frac{e_{ij}}{c_j^*}, \frac{f_{ij}}{c_j^*} \right) \text{ and } R = [r_{ij}]_{m\times n;} = \left(\frac{d_j^-}{d_{ij}}, \frac{d_j^-}{e_{ij}}, \frac{d_j^-}{f_{ij}} \right) \tag{4}$$

For benefit criteria, $c_{j^*}^* = \max_i cand$ for cost criteria, $d_j^- = \min_i d_{ij}$.

Step 4 Evaluate the weighted normalized decision matrix (V) using Eq. 5 (Singh and Agrawal 2018).

$$\mathrm{V} = \mathrm{R} * \mathrm{W} \tag{5}$$

where W is the weight of vector criteria as evaluated with AHP and $V = [v_{ij}]_{m\times n;}$ i = 1, 2,… m; j = 1,2,…n

Table 3 Scale for Linguistic terms

Linguistic terms	Fuzzy numbers
Very low	(0.0, 0.1, 0.3)
Low	(0.1, 0.3, 0.5)
Medium	(0.3, 0.5, 0.7)
High	(0.5,0.7,0.9)
Very high	(0.7, 0.9, 1)

Step 5 Generate the ideal and negative ideal solution for the strategic factors using Eq. 6 (Singh and Agrawal 2018).

$$A^{+} = \left\{V^{*}_{w1}, V^{*}_{w2} \ldots V^{*}_{wn}\right\} \text{ and } A^{-} = \left\{V^{-}_{w1}, V^{-}_{w2} \ldots V^{-}_{wn}\right\} \tag{6}$$

The values as per Eq. 7 are considered for the ideal and negative ideal solution.

$$V^{*} = (1, 1, 1) \text{ and } V^{-} = (0, 0, 0) \tag{7}$$

Step 6 Compute the total of distances from fuzzy positive ideal solution (FPIS) and fuzzy negative ideal solution (FNIS) for each factor using Eq. 8 (Singh and Agrawal 2018).

$$D^{+} = \frac{\sum_{i=1}^{m} d\left(V - V^{+}\right)}{m} \tag{8}$$

$d\left(V - V^{*}\right)$ is the distance between two fuzzy numbers which is determined using Eq. 9

$$d(V_1 - V_2) = \sqrt{\frac{1}{3}\left[\left(g_2 - g_1\right)^2 + \left(h_2 - h_1\right)^2 + \left(i_2 - i_1\right)^2\right]} \tag{9}$$

On the similar lines, the distance from the negative ideal solution is evaluated using Eq. 10

$$D^{-} = \frac{\sum_{i=1}^{m} d\left(v - v^{-}\right)}{m} \tag{10}$$

Step 7 Compute the relative closeness to the ideal solution using Eq. 11

$$C = D^{-}/D^{+} + D^{-} \tag{11}$$

Step 8 Rank the strategic factors based on the order of the values of C.

DEMATEL approach

DEMATEL method was applied for evaluating direct and indirect influences among criteria in multi-criteria decision scenarios (Gandhi et al. 2015). DEMATEL was applied for prioritizing of investment projects portfolio, agri-food supply chains for sustainable initiatives, and analysis of cloud adoption determinants (Hidayanto et al. 2015; Altuntas and Dereli 2015; Mangla et al. 2018). This method was also applied to evaluate green supply chain management practices and enablers in solar power developments (Lin 2013; Luthra et al. 2016). Singh et al. (2019) have used DEMATEL for analyzing the application of ICT in SMEs from the food industry. The steps used in AHP, Fuzzy TOPSIS and DEMATEL approach are described in the following subsection. The step by step process of DEMATEL approach is given below:

Step 1 Collect expert response and evaluate their average to obtain average matrix Z.

Consider '*m*' experts and '*n*' factors for the analysis. Expert opinion is taken based on pairwise comparison to get the degree of direct influence between two factors. x_{ij} denotes the degree of influence of factor ion factor *j* as per expert view. The integer scores of 0, 1, 2, 3, and 4 denote no influence, low influence, medium influence, high influence, and very high influence, respectively. An *n x n* nonnegative matrix, $X^{k} = [x^{k}_{ij}]$ is obtained from each expert. The average matrix $Z = [zij]$ is obtained as per Eq. 12 and represents the aggregate of all responses.

$$Z_{ij} = \frac{1}{m}\sum\nolimits_{i=1}^{m} x^{k}_{ij} \tag{12}$$

Step 2 Generate the normalized initial direct-relation matrix, N

The matrix, $N = [nij]$, where the value of each element in matrix N is ranged between [0, 1], and it is evaluated using Eq. 13.

$$\mathrm{N} = \lambda * \mathrm{Z}, \text{ or} \left[n_{ij}\right]_{\mathrm{nxn}} = \lambda \left[z_{ij}\right]_{\mathrm{nxn}} \tag{13}$$

$$\text{where } \lambda = Min\left[\frac{1}{\max 1 \le i \le n \sum_{j=1}^{m} |Z_{ij}|}, \frac{1}{\max 1 \le i \le n \sum_{j=1}^{n} |Z_{ij}|}\right].$$

Step 3 Develop the total relation matrix Y.

Total relation matrix *Y* is derived using Eq. 14 and its individual element represents the indirect effect of factor *i* on factor *j*. Matrix Y shows the total relationship between each pair of strategic factors.

$$Y = N(I - N)^{-1} \tag{14}$$

where I is the Identity matrix.

Step 4: Determine the sums of rows and columns of Total relation matrix Y.

The sums of rows and columns of matrix Y are denoted by vectors S_R and S_C, and these are evaluated using Eq. 15.

$$S_R = \left[r_i\right]_{nx1} = \left(\sum\nolimits_{j=1}^{n} y_{ij}\right)_{nx1} \text{ and } S_C = \left[c_j\right]'_{1xn} = \left(\sum\nolimits_{j=1}^{n} y_{ij}\right)'_{1xn} \tag{15}$$

The values of S_R and S_C indicate the total given and received both, directly and indirectly, effects that factor i has on the other factors.

Step 5 Develop a cause and effect relationship.

The cause and effect diagram is constructed in a coordinate plane using the values of $S_R + S_C$ and S_R—S_C as abscissa and ordinate, respectively. Interrelationships among system factors are established using the cause and effect diagram. Based on the values of (S_R—S_C), strategic factors

are classified into cause and effect group. If the score of (S_R—S_C) is positive, strategic factors fall in cause group and directly affect other strategic factors. On the other hand, if the score of (S_R—S_C), is negative, such strategic factors belong to the effect group, and these are influenced by the other strategic factors.

Results and discussion

As discussed in the research methodology framework (Fig. 1), this study is carried out in four phases. The first phase of identification of BDA benefits and strategic factors for its implementation has been done through literature review. The remaining three phases are analyzed based on using different MCDM techniques as discussed in the preceding section. AHP method is used for the justification of the application of BDA. Fuzzy TOPSIS approach is employed for the ranking of strategic factors, and the DEMATEL tool is utilized for the analysis of cause and effect of strategic factors. This study is applicable to the Indian manufacturing sector. The findings of this work will help the managers to take decision for BDA application for sustainable manufacturing operations in different fields under manufacturing sector. Therefore, the study will motivate to industry professionals to invest in BDA applications across different functions to achieve the goal of sustainable manufacturing operations. Results are discussed in the following subsections.

Justification of BDA application for sustainable manufacturing operations

This subsection evaluates the priority vector in terms of relative weights for the applications identified and the global desirability index (GDI) for two alternatives. The first alternative is big data-enabled manufacturing (BDM), and the second alternative is without big data-enabled manufacturing (WBDM). A higher value of GDI indicates a better alternative. Justification of BDA in the manufacturing sector is analyzed based on the framework developed in "Analytic hierarchy process" section (Fig. 2) using AHP. Initially, a pairwise comparison matrix (P_1) for seven applications of BDA at level 2 of the AHP model is developed as shown in Table 4. Each element of this matrix signifies the relative importance. For example, $p_{23} = 5$ signifies benefit at second row (Energy efficient and safe processes) has strong importance over benefit at the third column (Improved customer satisfaction). Element p_{32} is the reciprocal of p_{23} and interpreted accordingly as per the Saaty scale (refer to Appendix 1). Further, the priority vector is determined as per the procedure in "Analytic hierarchy process" section for all seven benefits and it signifies the relative weight of each benefit. The priority vector is shown in the last column of Table 4. Also, the CR is evaluated following the procedure given in "Analytic hierarchy process" section to examine the degree of consistency in the pairwise comparison of seven applications (Appendix 3), Results of the pairwise comparison are shown in Table 4. The evaluated value of CR is 0.0923 which is less than 0.1. This signifies the good level of consistency in the relative decision about applications.

Subsequently, priority vector is evaluated for both the alternatives, i.e,. big data-enabled manufacturing (BDM) and without big data-enabled manufacturing (WBDM) for each sustainability benefit. The results are shown in Table 5. For example, for benefit, PF the value of PV is 0.889 and 0.111 for BDM and WBDM, respectively. A higher value of PV in the case of BDM shows that big data-enabled manufacturing is justified when performance forecasting benefit is considered. It is observed from results (Table 5) that in terms of all seven benefits, manufacturing organizations with BDA have more priority vector value in comparison with manufacturing organizations without BDA.

Global weight for each alternative is calculated by multiplying the local weight of each alternative to the weight of each benefit, and these evaluated global weights for each alternative are shown in Table 6. Subsequently, the global

Table 4 Pairwise comparison matrix of sustainability benefits (Level 2)

	EPRR	EESP	ICS	IPM	WM	RO	DSC	Priority Vector (PV)
EPRR	1	0.143	0.25	0.2	0.2	0.33	0.12	0.025
EESP	7	1	5	3	2	6	0.33	0.2260
ICS	4	0.2	1	0.33	0.2	5	0.25	0.0765
IPM	5	0.33	3	1	0.33	4	0.2	0.1030
WM	5	0.5	5	3	1	7	0.5	0.1940
RO	3	0.167	0.2	0.25	0.14	1	0.16	0.0395
DSC	8	3	4	5	2	6	1	0.3360
Total	33	5.34	18.45	12.8	5.87	29.33	2.5	

Table 5 Pairwise comparison judgment matrices

Applications	Alternative investigation	BDM	WBDM	Total
EPRR	BDM	1	0.125	1.125
	WBDM	8	1	9
	PV	0.889	0.111	
EESP	BDM	1	0.125	1.125
	WBDM	8	1	9
	PV	0.889	0.111	
ICS	BDM	1	0.1429	1.1429
	WBDM	7	1	8
	PV	0.875	0.125	
IPM	BDM	1	0.1429	1.1429
	WBDM	7	1	8
	PV	0.875	0.125	
WM	BDM	1	0.125	1.125
	WBDM	8	1	9
	PV	0.889	0.111	
RO	BDM	1	0.1667	1.1667
	WBDM	6	1	7
	PV	0.857	0.143	
DSC	BDM	1	0.1429	1.1429
	WBDM	7	1	8
	PV	0.875	0.125	

desirability index (GDI) is obtained by summing the all values of corresponding alternatives. The GDI value for BDM and WBDM is 0.8811 and 0.1189, respectively, as given in Table 6. A higher value of GDI justifies the application of BDA for sustainable manufacturing operations.

Ranking of strategic factors for the application of BDA

Strategic factors for BDA application in the manufacturing sector are identified through literature. After discussion with experts, fifteen factors are finalized for their ranking from a strategic perspective. These factors are listed in Table 2. A questionnaire-based survey was conducted in the context of big data analytics application for sustainable manufacturing operations in the Indian manufacturing sector, and the experts were selected from industry and academia. The experts from industries and academia were requested to give the responses in the questionnaire designed for this study. The experts team comprises two production managers, one marketing manager, one operation engineer, one logistics manager, and two academicians. The experts from industries have more than ten years of experience in their domain, whereas the experts from academia have more than fifteen years of experience in teaching and research. In total, seven experts were requested to provide their responses for rating of all 15 strategic factors in linguistic terms. A 5-point scale having the linguistic terms as very low (VL), low (L), medium (M), high (H), and very high (VH) was used for this purpose. The procedure described in "Fuzzy TOPSIS approach" section is applied here for ranking. The authors collected the responses from the experts in linguistic terms and converted them into crisp values by referring scale given in Table 3. Thus, the matrix so obtained is called fuzzy decision matrix D, and it is shown in Appendix 4. Then matrix D is converted into an un-weighted fuzzy matrix, R using Eq. 4, and the same is given in Appendix 5. Further the weighted normalized matrix is evaluated using Eq. 5. This evolves the product of un-weighted fuzzy decision matrix R (Appendix 5) and PV value for sustainability benefits given in Table 4. The same is shown in Table 7.

Next, the distance of the rating of each factor from a positive ideal solution is evaluated using Eq. 8. This is given in

Table 6 Weights of applications for alternatives and global desirability index

Sr. no	Applications	Weight of each benefit	Alternatives			
			BDM local weight	WBDM local weight	BDM global weight	WBDM global weight
1	EPRR	0.025	0.889	0.111	0.0223	0.0027
2	EESP	0.2260	0.889	0.111	0.2009	0.0251
3	ICS	0.0765	0.875	0.125	0.0669	0.0094
4	IPM	0.1030	0.875	0.125	0.0901	0.0127
5	WM	0.1940	0.889	0.111	0.1724	0.0214
6	RO	0.0395	0.875	0.143	0.0345	0.0056
7	DSC	0.3360	0.875	0.125	0.294	0.0420
Total global weight (GDI)					0.8811	0.1189

Appendix 6. Similarly, the distance of the rating of each factor from a negative ideal solution is evaluated using Eq. 10 and shown in Appendix 7. Further, the total distance of each factor is calculated from the positive and negative ideal solution. These are represented by D^+ and D^-, and the same is given in Appendices 6 and 7. Subsequently, the relative closeness with respect to ideal solution A^+ is evaluated using Eq. 11, and the same is used in the ranking of performance. The biggest value of closeness is ranked "1" and the lowest value of closeness is ranked "15." By following this closeness value, all the strategic factors are ranked and tabulated in Table 8. Commitment and engagement of top management, strategy development for BDA, and development of capability for handling big data are prioritized as 1st, 2nd, and 3rd in their relative importance, and these are very crucial for BDA implementation. Without commitment and support from top management, such kind of high-cost initiatives cannot be successful. Management should also develop a trained workforce to manage huge data through BDA. Responsive information sharing framework and development of contract agreement among all stakeholders are ranked 14th and 15th, respectively, and these factors have relatively less impact on the implementation of BDA.

Categorization of strategic factors in terms of cause and effect

As mentioned earlier, the DEMATEL approach is applied to categorize the strategic factors into two classes, viz. cause and effect. This is implemented by the evaluation of direct and indirect influences among strategic factors. Following the procedure given in "DEMATEL approach" section, the influence matrix data is recorded in terms of the 15 strategic factors from 7 experts and an average influence matrix, Z, is constructed using Eq. 12 (Appendices 3, 4, 5, 6, 7, and 8). Further, a normalized initial direct-influence matrix, N, is derived using Eq. 13. Average influence matrix, Z, and normalized initial direct-influence matrix, N, are given in Table 9. The steps of the DEMATEL approach were evaluated according to the above-mentioned procedure in "DEMATEL approach" section.

Subsequently in the DEMATEL approach the total influence matrix, Y, is calculated using Eq. 14, and results are shown in Table 10. Further, the rows sum vector (S_R), columns sum vector (S_C), $S_R + S_C$ vector, and S_R—S_C vector of matrix Y are calculated using Eq. 15 and these are shown in Table 11. Finally, the ranking of strategic factors is obtained on the basis of their (S_R–S_C) values and this is also presented in Table 11. Further, the results of the DEMATEL approach are also summarized in Fig. 3 in cause and effect groups.

Following the criteria of the positive score of (S_R–S_C), eight strategic factors falling in the category of cause group are identified, which directly affect other strategic factors in

Table 7 Weighted normalized fuzzy matrix

	1			2			3			4			5			6			7		
C-1	0.016	0.02	0.022	0.02	0.06	0.1	0.047	0.06	0.067	0.027	0.045	0.063	0.017	0.052	0.086	0.003	0.01	0.017	0.029	0.088	0.147
C-2	0.016	0.02	0.022	0.141	0.181	0.201	0.047	0.06	0.067	0.045	0.063	0.081	0.052	0.086	0.121	0.017	0.024	0.031	0.147	0.206	0.265
C-3	0.016	0.02	0.022	0.141	0.181	0.201	0.02	0.033	0.047	0.045	0.063	0.081	0.052	0.086	0.121	0.017	0.024	0.031	0.147	0.206	0.265
C-4	0.007	0.011	0.016	0.06	0.1	0.141	0.007	0.02	0.033	0.045	0.063	0.081	0.017	0.052	0.086	0.01	0.017	0.024	0.088	0.147	0.206
C-5	0.007	0.011	0.016	0.1	0.141	0.181	0.047	0.06	0.067	0.009	0.027	0.045	0.052	0.086	0.121	0.01	0.017	0.024	0.029	0.088	0.147
C-6	0.011	0.016	0.02	0.1	0.141	0.181	0.047	0.06	0.067	0.045	0.063	0.081	0.121	0.155	0.172	0.017	0.024	0.031	0.206	0.265	0.294
C-7	0.016	0.02	0.022	0.1	0.141	0.181	0.047	0.06	0.067	0.045	0.063	0.081	0.052	0.086	0.121	0.017	0.024	0.031	0.088	0.147	0.206
C-8	0.011	0.016	0.02	0.1	0.141	0.181	0.033	0.047	0.06	0.027	0.045	0.063	0.017	0.052	0.086	0.024	0.031	0.035	0.088	0.147	0.206
C-9	0.016	0.02	0.022	0.141	0.181	0.201	0.033	0.047	0.06	0.027	0.045	0.063	0.121	0.155	0.172	0.017	0.024	0.031	0.147	0.206	0.265
C-10	0.011	0.016	0.02	0.141	0.181	0.201	0.02	0.033	0.047	0.027	0.045	0.063	0.017	0.052	0.086	0.017	0.024	0.031	0.147	0.206	0.265
C-11	0.011	0.016	0.02	0.141	0.181	0.201	0.007	0.02	0.033	0.045	0.063	0.081	0.052	0.086	0.121	0.017	0.024	0.031	0.088	0.147	0.206
C-12	0.007	0.011	0.016	0.1	0.141	0.181	0.033	0.047	0.06	0.027	0.045	0.063	0.052	0.086	0.121	0.01	0.017	0.024	0.088	0.147	0.206
C-13	0.011	0.016	0.02	0.06	0.1	0.141	0.033	0.047	0.06	0.045	0.063	0.081	0.086	0.121	0.155	0.01	0.017	0.024	0.088	0.147	0.206
C-14	0.011	0.016	0.02	0.06	0.1	0.141	0.02	0.033	0.047	0.027	0.045	0.063	0.052	0.086	0.121	0.01	0.017	0.024	0.088	0.147	0.206
C-15	0.002	0.007	0.011	0.06	0.1	0.141	0.007	0.02	0.033	0.027	0.045	0.063	0.017	0.052	0.086	0.003	0.01	0.017	0.088	0.147	0.206

Table 8 Closeness coefficient matrix and ranking of strategic factors

Strategic factors	Strategic factors for BDA implementation	D^+	D^-	C	Ranking
C-1	Development of contract agreement among all stakeholders	6.67038	0.3521	0.050139	15
C-2	Commitment and engagement of top management	6.33759	0.75235	0.106115	1
C-3	Development of capability for handling big data	6.39095	0.72115	0.101398	3
C-4	Robust cybersecurity system	6.64735	0.42696	0.060353	13
C-5	Coordination among big data stakeholders	6.4504	0.62222	0.087976	8
C-6	Problems identification and solving capabilities	6.44129	0.62757	0.088779	7
C-7	Process Integration and institutionalization	6.43448	0.63192	0.089427	6
C-8	Flexible digital infrastructure	6.46666	0.612	0.086457	9
C-9	Strategy development for BDA	6.36313	0.73731	0.10384	2
C-10	Availability of quality and reliable big data	6.39781	0.71702	0.100779	4
C-11	Knowledgeable and capable decision-makers	6.42569	0.70278	0.098587	5
C-12	Data-driven organization culture	6.47577	0.60665	0.085656	10
C-13	Process monitoring and control	6.58347	0.46553	0.066042	11
C-14	Integrating customers' requirements with performance framework	6.61085	0.44798	0.05957	12
C-15	Responsive information sharing framework	6.65639	0.42163	0.063464	14

Table 9 Average direct-influence matrix, Z, and the normalized initial direct-influence matrix, N

		C-1	C-2	C-3	C-4	C-5	C-6	C-7	C-8	C-9	C-10	C-11	C-12	C-13	C-14	C-15
C-1	Z	0	2	2	2.5714	1.5714	3	2	2.7143	1.7143	1.8571	2.4286	3	2.5714	2.5714	2.1429
	N	0	0.066	0.066	0.0849	0.0519	0.0991	0.066	0.0896	0.0566	0.0613	0.0802	0.0991	0.0896	0.0849	0.0708
C-2	Z	2.2857	0	2.1429	2	2.2857	2.5714	2.4286	3.1429	2.7143	2.5714	1.7143	2.1429	2.5714	1.7143	2.1429
	N	0.0755	0	0.0708	0.066	0.0755	0.0849	0.0802	0.1038	0.0896	0.0849	0.0566	0.0708	0.0849	0.0566	0.0708
C-3	Z	2	2.4286	0	1.7143	2.1429	2	1.5714	2.7143	1.4286	1.2857	2	1.4286	2	1.4286	2
	N	0.066	0.0802	0	0.0566	0.0708	0.066	0.0519	0.0896	0.0472	0.0425	0.066	0.0472	0.066	0.0472	0.066
C-4	Z	2	1.2857	2.4286	0	2.4286	2.2857	2.1429	2.4286	2.1429	2.5714	2.1429	0.8571	2.4286	1	2.4286
	N	0.066	0.0425	0.0802	0	0.0802	0.0755	0.0708	0.0802	0.0708	0.0849	0.0708	0.0283	0.0802	0.033	0.0802
C-5	Z	1.5714	2.1429	1.8571	1.5714	0	2.8571	1.5714	2	2	2.4286	2	3	2.2857	2.8571	2.5714
	N	0.0519	0.0708	0.0613	0.0519	0	0.0943	0.0519	0.066	0.066	0.0802	0.066	0.0991	0.0755	0.0943	0.0849
C-6	Z	2.1429	2.4286	1.8571	2.5714	2.1429	0	2.1429	2.2857	1.8571	2.2857	2.5714	2.4286	2.7143	1.7143	2
	N	0.0708	0.0802	0.0613	0.0849	0.0708	0	0.0708	0.0755	0.0613	0.0755	0.0849	0.0802	0.0896	0.0566	0.066
C-7	Z	1.8571	1.8571	1.7143	2.4286	2	1.4286	0	2	1.2857	1.5714	1.7143	2.5714	2.1429	1.2857	2
	N	0.0613	0.0613	0.0566	0.0802	0.066	0.0472	0	0.066	0.0425	0.0519	0.0566	0.0849	0.0708	0.0425	0.066
C-8	Z	2.2857	2.1429	2.2857	2.7143	1.8571	2.1429	1.4286	0	2.1429	1.8571	2	1.4286	2.5714	1.8571	2.8571
	N	0.0755	0.0708	0.0755	0.0896	0.0613	0.0708	0.0472	0	0.0708	0.0613	0.066	0.0472	0.0849	0.0613	0.0943
C-9	Z	1.7143	2	2	2	2.2857	1.1429	2	2.7143	0	2.5714	1.5714	1.5714	2.7143	2.1429	1.4286
	N	0.0566	0.066	0.066	0.066	0.0755	0.0377	0.066	0.0896	0	0.0849	0.0519	0.0519	0.0896	0.0708	0.0472
C-10	Z	2.2857	2.7143	2	2.2857	2.2857	2.1429	2.4286	2.7143	2.5714	0	2.5714	2.5714	0.5714	2	3
	N	0.0755	0.0896	0.066	0.0755	0.0755	0.0708	0.0802	0.0896	0.0849	0	0.0849	0.0849	0.0189	0.066	0.0991
C-11	Z	2.2857	2	2.1429	2.2857	2.1429	2.7143	2	2.1429	1.7143	1.8571	0	2.5714	2.1429	1.8571	2.2857
	N	0.0755	0.066	0.0708	0.0755	0.0708	0.0896	0.066	0.0708	0.0566	0.0613	0	0.0849	0.0708	0.0613	0.0755
C-12	Z	2.2857	0.8571	1.2857	1.8571	2.8571	2.2857	2.2857	2	1.5714	2.7143	2.7143	0	2.4286	1.1429	1.4286
	N	0.0755	0.283	0.0425	0.0613	0.0943	0.0755	0.0755	0.066	0.0519	0.0896	0.0896	0	0.080 2	0.0377	0.0472
C-13	Z	2	2.8571	1.1429	2	2.5714	1.7143	2.4286	2.1429	2.4286	0.7143	1.4286	2.4286	0	1.8571	1.8571
	N	0.066	0.0943	0.0377	0.066	0.0849	0.0566	0.0802	0.0708	0.0802	0.0236	0.0472	0.0802	0	0.0613	0.0613
C-14	Z	2	1.8571	1.5714	1.7143	2.8571	1.2857	1.2857	1.7143	2	2.1429	2.1429	1.2857	1.8571	0	2
	N	0.066	0.0613	0.0519	0.0566	0.0943	0.0425	0.0425	0.0566	0.066	0.0708	0.0708	0.0425	0.0613	0	0.066
C-15	Z	2.1429	3	1.57.14	1.4286	2.5714	2	1.8571	2.8571	1.7143	2.1429	1.8571	1.2857	2.7143	2.1429	0
	N	0.0708	0.0991	0.0519	0.0472	0.0849	0.066	0.0613	0.0943	0.0566	0.0708	0.0613	0.0425	0.0896	0.0708	0

Table 10 Total direct-influence matrix (Y)

	C-1	C-2	C-3	C-4	C-5	C-6	C-7	C-8	C-9	C-10	C-11	C-12	C-13	C-14	C-15
C-1	1.8935	2.0044	1.7778	1.9927	2.1296	2.0385	1.8768	2.2602	1.8575	1.9292	1.9633	1.9719	2.167	1.7746	2.0463
C-2	1.9778	1.9592	1.7961	1.9914	2.1652	2.0409	1.9033	2.2908	1.9012	1.9643	1.9564	1.9629	2.1787	1.7648	2.0627
C-3	1.6271	1.6818	1.4196	1.6362	1.7845	1.6735	1.5494	1.8843	1.5372	1.588	1.6224	1.6013	1.7861	1.4493	1.7002
C-4	1.7587	1.7861	1.6138	1.7161	1.9375	1.8166	1.6929	2.0283	1.6835	1.7548	1.7582	1.7166	1.9421	1.5558	1.8515
C-5	1.8599	1.9244	1.6978	1.8786	1.9897	1.9485	1.7848	2.1446	1.7872	1.8648	1.8683	1.8901	2.0622	1.71	1.9718
C-6	1.9049	1.9615	1.725	1.9375	2.0858	1.8927	1.8295	2.1863	1.8105	1.8878	1.9124	1.9028	2.1064	1.7019	1.9861
C-7	1.5977	1.6376	1.4496	1.6309	1.7537	1.6308	1.4766	1.834	1.5085	1.572	1.5897	1.6093	1.7617	1.4211	1.6732
C-8	1.8205	1.8644	1.6577	1.8517	1.9806	1.8673	1.7234	2.0157	1.7348	1.7878	1.8072	1.784	2.0061	1.6278	1.9184
C-9	1.7021	1.7552	1.557	1.7288	1.8814	1.733	1.6422	1.98	1.5725	1.7075	1.6933	1.6884	1.8961	1.5451	1.7717
C-10	1.9674	2.0289	1.7831	1.9879	2.1535	2.0187	1.8923	2.2666	1.8856	1.8779	1.9711	1.9633	2.1099	1.7631	2.0761
C-11	1.8516	1.8903	1.681	1.8711	2.0228	1.9162	1.7697	2.116	1.7508	1.8191	1.7774	1.8496	2.0276	1.6547	1.9338
C-12	1.7237	1.7266	1.5404	1.7306	1.9025	1.7733	1.6561	1.964	1.6252	1.7169	1.733	1.6475	1.8935	1.5209	1.7769
C-13	1.6955	1.7628	1.5185	1.7137	1.8738	1.7351	1.6411	1.9467	1.632	1.6417	1.6744	1.6993	1.8014	1.5235	1.7666
C-14	1.6028	1.6409	1.4473	1.611	1.7798	1.6281	1.5182	1.8278	1.5316	1.5904	1.6032	1.5743	1.7539	1.3842	1.6756
C-15	1.8091	1.8819	1.63	1.8077	1.9927	1.8559	1.7286	2.0931	1.7164	1.7887	1.7954	1.7751	2.0014	1.6307	1.8249

Table 11 Categorization of strategic factors into the cause and effect groups

Strategic factors	Description	R	C	$(S_R + S_C)$	$(S_R - S_C)$	Ranking of strategic factors based on $(S_R - S_C)$	Causes and effects
C-1	Development of contract agreement among all stakeholders	29.68337	26.79219	56.47556	2.891181	Second	Cause
C-2	Commitment and engagement of top management	29.91558	27.50613	57.42171	2.40945	Third	Cause
C-3	Development of capability for handling big data	24.5409	24.29446	48.83536	0.246444	Seventh	Cause
C-4	Robust cyber security system	26.61256	27.08624	53.6988	−0.47367	Ninth	Effect
C-5	Coordination among big data stake holders	28.38261	29.43312	57.81573	−1.05052	Twelfth	Effect
C-6	Problems identification and solving capabilities	28.83097	27.56897	56.39995	1.262001	Fourth	Cause
C-7	Process Integration and institutionalization	24.14625	25.68487	49.83112	−1.53861	Thirteenth	Effect
C-8	Flexible digital infrastructure	27.44729	30.83838	58.28567	−3.39109	Fourteenth	Effect
C-9	Strategy development for BDA	25.85444	25.53452	51.38896	0.319924	Sixth	Cause
C-10	Availability of quality and reliable big data	29.74539	26.49093	56.23631	3.254461	First	Cause
C-11	Knowledgeable and capable decision-makers	27.93184	26.72567	54.65751	1.206164	Fifth	Cause
C-12	Data driven organization culture	25.93108	26.63639	52.56747	−0.70532	Eleventh	Effect
C-13	Process monitoring and control	25.62615	29.49395	55.1201	−3.86779	Fifteenth	Effect
C-14	Integrating customers' requirements with performance framework	24.16918	24.02741	48.19659	0.141777	Eighth	Cause
C-15	Responsive information sharing framework	27.33152	28.03592	55.36745	−0.7044	Tenth	Effect

effect group. These cause group factors are: development of contract agreement among all stakeholders, commitment and engagement of top management, development of capacity for handling big data, problems identification and solving capabilities, strategy development for BDA, availability of quality and reliable data, knowledgeable and capable

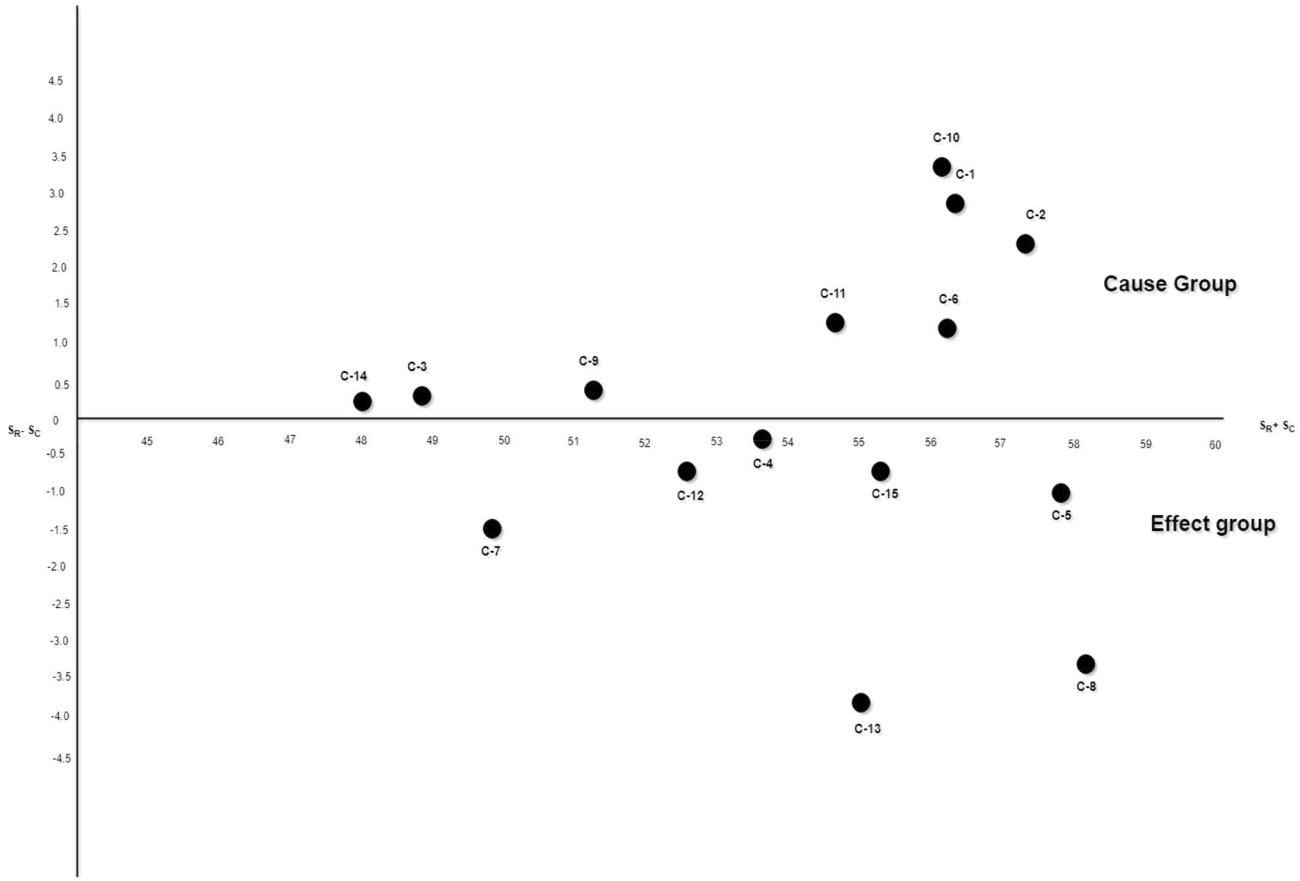


Fig. 3 The Causal diagram for strategic factors

decision-makers and integrating customers' requirements with performance framework. These cause group strategic factors are treated as independent factors and these have direct influence on the organization. These strategic factors require more attention for the successful implementation of BDA in the manufacturing sector.

The availability of quality and reliable big data has the highest (S_R–S_C) score, and it is the most important cause that has the highest direct impact on the other strategic factors. Based on (S_R–S_C) score development of contract agreement among all stakeholders and Commitment and engagement of top management are placed at the second and third highest positions in cause group strategic factors. This suggests that the development of contract agreement among all stakeholders and the engagement of top management are necessary for the implementation of BDA in manufacturing organizations. Problems identified and solving capabilities, with (S_R–S_C) score of 1.262 have fourth position pointing its importance on BDA in the manufacturing sector. Further, the knowledgeable and capable decision-makers with (S_R–S_C) score of 1.2016 are an important factor that will help in taking the right decisions timely for the organization. Next, strategy development for BDA, with (S_R–S_C) score of 0.3199 will aid in development of strategy. Development of capability for handling big data with (SR–SC) score of 0.2464 is another crucial factor that will ensure data handling. Integrating customer requirements with performance framework has the eighth rank with the smallest (S_R–S_C) score of (0.14178).

There is similarity up to a certain extent in drawing inferences from Fuzzy TOPSIS and DEMATEL approaches. The top five strategic factors as ranked by Fuzzy TOPSIS (refer to Table 8) are also falling in cause group as identified by the

DEMATEL approach. Therefore, management should give more attention to these strategic independent factors as they have a crucial role in implementing BDA for manufacturing organizations.

Further, based on the negative (S_R–S_C) values seven strategic factors fall in the effect group. There effect group factors are: robust cybersecurity system, coordination among big data stakeholders, process integration and institutionalization, flexible digital infrastructure, data-driven organization culture, process monitoring and control, and responsive information sharing framework. These strategic factors were most affected by the other strategic factors.

Conclusion, limitations, future scope, and managerial implications

In the present uncertain business environment, organizations are exploring application of emerging technologies for decision making and efficient production planning to ensure sustainable operations. BDA can help organization in managing unstructured huge data generated from digitization of operations. So far, the application of BDA for sustainable operations in the manufacturing sector is not extensively researched specifically in context to developing countries like India. Apart from it, the successful application of BDA in manufacturing processes is another big challenge for most of organizations. Major contribution of present study is that it has explored the application of BDA in manufacturing from sustainability perspective. From the literature, it has been observed that BDA application improves process and energy efficiency, product recovery, resource optimization, waste minimization, sustainable capabilities, etc. In spite of many benefits, several manufacturing organizations are still reluctant to integrate BDA with their processes due to heavy investment and risk of failure. Therefore, firstly based on industry inputs, authors justify BDA applications in manufacturing functions by using AHP. It has been observed from the results that BDA integrated manufacturing processes are more sustainable. This result is in line with the findings of other studies done in different context. Researchers have observed that BDA reduces manufacturing waste (Lee et al. 2013), increases equipment availability for the manufacturing process (Munirathinam and Ramadoss 2014), and improves insights into identification of faulty products and thus prevents returns and rework (Lavalle et al. 2011).

Further, strategic factors for BDA implementation are ranked by fuzzy TOPSIS**.** It has been found that commitment and engagement of top management, development of capability for handling big data, strategy development for BDA, knowledgeable and capable decision-makers, availability of quality and reliable data, process integration, and institutionalization are the major strategic factors for successful implementation of BDA in manufacturing. Commitment and engagement of top management is the most important factor as the top management plays an important role in implementation of BDA and other supporting technologies that may ensure the benefits identified in the study.

Finally, the DEMATEL approach is used to categorize strategic factors in terms of cause and effect. Availability of quality and reliable big data, commitment, and engagement of top management, development of contract agreement among all stakeholders are major factors in the cause category. This information of cause factors will help managers to prioritize the actions for the implementation of BDA. Although findings of the study will be highly motivating to industry professionals to invest in BDA applications to achieve the goal of sustainable operations, but results cannot be generalized. For further validation, empirical study along with few case studies can be carried out to validate the findings. The importance of these factors may vary as per sector and country, so in the future studies may be carried out to compare the findings in different contexts. Generally, industries have operational and business data about their inventories, products, human resources, and distributors. Companies may use big data analytics to capitalize on this and many other sources of data to optimize manufacturing operations. Therefore, companies should formulate their policies and channelize the resources as per priority of strategic factor for a smooth implementation of BDA in manufacturing sector.

Author contributions All authors contributed extensively to the work presented in this paper.

Funding There has been no financial support for this work that could have influenced its outcome.

Data availability The data that support the findings of this study are collected through the expert's opinion and available in Appendix.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Appendix 1

See Table 12.

Table 12 Thomas Saaty's scale for pairwise comparison of criteria (Saaty 1994)

Importance intensity	Terminology	Explanation
1	Equal importance	Allocate dissimilar value to each element depending on the significance of factor on another factor. If two factors have equal significance, then intensity of significance should be unity and one is allocated to both factors. Therefore, allocate value 3, 5, 7, 9 or the value of two adjoining judgments, i.e., 2, 4, 6, 8 is depending on the significance of each factor
3	Weak importance of one over another	
5	Essential or strong importance	
7	Highly strong importance	
9	Extreme/supreme importance	
2, 4, 6, 8	Intermediate, middle values of two adjoining judgments	
Reciprocals	Allocate any one number from previously mentioned value to activity i when it is compared with j, and allocated the corresponding of its value when the activity j compared to i	

Appendix 2

See Table 13.

Table 13 Average random index value (Saaty 2000)

Size of matrix (n)	1	2	3	4	5	6	7	8	9	10
Random Consistency Index (RCI)	0	0	0.52	0.9	1.1	1.25	1.35	1.4	1.45	1.49

Appendix 3

Let P_1 be the pairwise comparison matrix and P_2 principal vector matrix.

$$P_1 = \begin{bmatrix} 1 & 0.143 & 0.25 & 0.2 & 0.2 & 0.33 & 0.125 \\ 7 & 1 & 5 & 3 & 2 & 6 & 0.33 \\ 4 & 0.2 & 1 & 0.33 & 0.2 & 5 & 0.25 \\ 5 & 0.33 & 3 & 1 & 0.33 & 4 & 0.2 \\ 5 & 0.5 & 5 & 3 & 1 & 7 & 0.5 \\ 3 & 0.167 & 0.2 & 0.25 & 0.143 & 1 & 0.167 \\ 8 & 3 & 4 & 5 & 2 & 6 & 1 \end{bmatrix}$$

$$P_2 = \begin{bmatrix} 0.025 \\ 0.2260 \\ 0.0765 \\ 0.1030 \\ 0.1940 \\ 0.0395 \\ 0.3360 \end{bmatrix} \quad P_3 = P_1 * P_3 = \begin{bmatrix} 0.19087 \\ 1.8283 \\ 0.5750 \\ 0.8213 \\ 1.568 \\ 0.2421 \\ 2.660 \end{bmatrix} \quad \text{and } P_4 = P_3/P_2 = \begin{bmatrix} 0.19087 \\ 1.8283 \\ 0.5750 \\ 0.8213 \\ 1.568 \\ 0.2421 \\ 2.660 \end{bmatrix}$$

λ max, Average of the element of $P_4 = 7.7481$.

Now, consistency Index (CI) $= \frac{\lambda \max - n}{n-1} = (7.7481–7)/(7–1) = 0.12468$.

And, consistency ratio (CR) = CI/RCI = (n Appendix 2).

CR = 0.12468/1.35 0.0923, i.e., CR < 0.1. So, result is consistent.

Appendix 4

See Table 14.

Table 14 Fuzzy decision matrix D

	1	2	3	4	5	6	7
C-1	(0.7,0.9,1.0)	(0.1,0.3,0.5)	(0.7,0.9,1.0)	(0.3,0.5,0.7)	(0.1,0.3,0.5)	(0.1,0.3,0.5)	(0.1,0.3,0.5)
C-2	(0.7,0.9,1.0)	(0.7,0.9,1.0)	(0.7,0.9,1.0)	(0.5,0.7,0.9)	(0.3,0.5,0.7)	(0.5,0.7,0.9)	(0.5,0.7,0.9)
C-3	(0.7,0.9,1.0)	(0.7,0.9,1.0)	(0.3,0.5,0.7)	(0.5,0.7,0.9)	(0.3,0.5,0.7)	(0.5,0.7,0.9)	(0.5,0.7,0.9)
C-4	(0.3,0.5,0.7)	(0.3,0.5,0.7)	(0.1,0.3,0.5)	(0.5,0.7,0.9)	(0.1,0.3,0.5)	(0.3,0.5,0.7)	(0.3,0.5,0.7)
C-5	(0.3,0.5,0.7)	(0.5,0.7,0.9)	(0.7,0.9,1.0)	(0.1,0.3,0.5)	(0.3,0.5,0.7)	(0.3,0.5,0.7)	(0.1,0.3,0.5)
C-6	(0.5,0.7,0.9)	(0.5,0.7,0.9)	(0.7,0.9,1.0)	(0.5,0.7,0.9)	(0.7,0.9,1.0)	(0.5,0.7,0.9)	(0.1,0.3,0.5)
C-7	(0.7,0.9,1.0)	(0.5,0.7,0.9)	(0.7,0.9,1.0)	(0.5,0.7,0.9)	(0.3,0.5,0.7)	(0.5,0.7,0.9)	(0.3,0.5,0.7)
C-8	(0.5,0.7,0.9)	(0.5,0.7,0.9)	(0.5,0.7,0.9)	(0.3,0.5,0.7)	(0.1,0.3,0.5)	(0.7,0.9,1.0)	(0.3,0.5,0.7)
C-9	(0.7,0.9,1.0)	(0.7,0.9,1.0)	(0.5,0.7,0.9)	(0.3,0.5,0.7)	(0.7,0.9,1.0)	(0.5,0.7,0.9)	(0.5,0.7,0.9)
C-10	(0.5,0.7,0.9)	(0.7,0.9,1.0)	(0.3,0.5,0.7)	(0.3,0.5,0.7)	(0.1,0.3,0.5)	(0.5,0.7,0.9)	(0.5,0.7,0.9)
C-11	(0.5,0.7,0.9)	(0.7,0.9,1.0)	(0.1,0.3,0.5)	(0.5,0.7,0.9)	(0.3,0.5,0.7)	(0.5,0.7,0.9)	(0.3,0.5,0.7)
C-12	(0.5,0.7,0.9)	(0.5,0.7,0.9)	(0.5,0.7,0.9)	(0.3,0.5,0.7)	(0.3,0.5,0.7)	(0.3,0.5,0.7)	(0.3,0.5,0.7)
C-13	(0.5,0.7,0.9)	(0.3,0.5,0.7)	(0.5,0.7,0.9)	(0.5,0.7,0.9)	(0.5,0.7,0.9)	(0.3,0.5,0.7)	(0.3,0.5,0.7)
C-14	(0.5,0.7,0.9)	(0.3,0.5,0.7)	(0.3,0.5,0.7)	(0.3,0.5,0.7)	(0.3,0.5,0.7)	(0.3,0.5,0.7)	(0.3,0.5,0.7)
C-15	(0.1,0.3,0.5)	(0.3,0.5,0.7)	(0.1,0.3,0.5)	(0.3,0.5,0.7)	(0.1,0.3,0.5)	(0.1,0.3,0.5)	(0.3,0.5,0.7)

Appendix 5

See Table 15.

Table 15 Un-weighted fuzzy matrix R

	1			2			3			4			5			6			7		
C-1	0.7	0.9	1	0.1	0.3	0.5	0.7	0.9	1	0.3	0.5	0.7	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
C-2	0.7	0.9	1	0.7	0.9	1	0.7	0.9	1	0.5	0.7	0.9	0.3	0.5	0.7	0.5	0.7	0.9	0.5	0.7	0.9
C-3	0.7	0.9	1	0.7	0.9	1	0.3	0.5	0.7	0.5	0.7	0.9	0.3	0.5	0.7	0.5	0.7	0.9	0.5	0.7	0.9
C-4	0.3	0.5	0.7	0.3	0.5	0.7	0.1	0.3	0.5	0.5	0.7	0.9	0.1	0.3	0.5	0.3	0.5	0.7	0.3	0.5	0.7
C-5	0.3	0.5	0.7	0.5	0.7	0.9	0.7	0.9	1	0.1	0.3	0.5	0.3	0.5	0.7	0.3	0.5	0.7	0.1	0.3	0.5
C-6	0.5	0.7	0.9	0.5	0.7	0.9	0.7	0.9	1	0.5	0.7	0.9	0.7	0.9	1	0.5	0.7	0.9	0.7	0.9	1
C-7	0.7	0.9	1	0.5	0.7	0.9	0.7	0.9	1	0.5	0.7	0.9	0.3	0.5	0.7	0.5	0.7	0.9	0.3	0.5	0.7
C-8	0.5	0.7	0.9	0.5	0.7	0.9	0.5	0.7	0.9	0.3	0.5	0.7	0.1	0.3	0.5	0.7	0.9	1	0.3	0.5	0.7
C-9	0.7	0.9	1	0.7	0.9	1	0.5	0.7	0.9	0.3	0.5	0.7	0.7	0.9	1	0.5	0.7	0.9	0.5	0.7	0.9
C-10	0.5	0.7	0.9	0.7	0.9	1	0.3	0.5	0.7	0.3	0.5	0.7	0.1	0.3	0.5	0.5	0.7	0.9	0.5	0.7	0.9
C-11	0.5	0.7	0.9	0.7	0.9	1	0.1	0.3	0.5	0.5	0.7	0.9	0.3	0.5	0.7	0.5	0.7	0.9	0.3	0.5	0.7
C-12	0.3	0.5	0.7	0.5	0.7	0.9	0.5	0.7	0.9	0.3	0.5	0.7	0.3	0.5	0.7	0.3	0.5	0.7	0.3	0.5	0.7
C-13	0.5	0.7	0.9	0.3	0.5	0.7	0.5	0.7	0.9	0.5	0.7	0.9	0.5	0.7	0.9	0.3	0.5	0.7	0.3	0.5	0.7
C-14	0.5	0.7	0.9	0.3	0.5	0.7	0.3	0.5	0.7	0.3	0.5	0.7	0.3	0.5	0.7	0.3	0.5	0.7	0.3	0.5	0.7
C-15	0.1	0.3	0.5	0.3	0.5	0.7	0.1	0.3	0.5	0.3	0.5	0.7	0.1	0.3	0.5	0.1	0.3	0.5	0.3	0.5	0.7

Appendix 6

See Table 16.

Table 16 Distance of the ratings of each factor from A^+ with respect to each criterion

	1	2	3	4	5	6	7	D+
C-1	0.981	0.979	0.966	0.94	0.931	0.931	0.942	6.67
C-2	0.981	0.941	0.888	0.826	0.86	0.9	0.942	6.338
C-3	0.981	0.941	0.888	0.826	0.87	0.919	0.967	6.391
C-4	0.989	0.971	0.942	0.9	0.919	0.946	0.98	6.647
C-5	0.989	0.958	0.916	0.86	0.879	0.906	0.942	6.45
C-6	0.984	0.955	0.914	0.86	0.879	0.906	0.942	6.441
C-7	0.981	0.953	0.914	0.86	0.879	0.906	0.942	6.434
C-8	0.984	0.955	0.914	0.86	0.884	0.915	0.953	6.467
C-9	0.981	0.941	0.888	0.826	0.865	0.909	0.953	6.363
C-10	0.984	0.943	0.889	0.826	0.87	0.919	0.967	6.398
C-11	0.984	0.943	0.889	0.826	0.875	0.928	0.98	6.426
C-12	0.989	0.958	0.916	0.86	0.884	0.915	0.953	6.476
C-13	0.984	0.968	0.94	0.9	0.91	0.928	0.953	6.583
C-14	0.984	0.968	0.94	0.9	0.914	0.937	0.967	6.611
C-15	0.993	0.974	0.943	0.9	0.919	0.946	0.98	6.656

Appendix 7

See Table 17.

Table 17 Distance of the ratings of each factor from a with respect to each criterion

	1	2	3	4	5	6	7	**D-**
C-1	0.02	0.021	0.039	0.069	0.073	0.073	0.059	**0.352**
C-2	0.02	0.083	0.133	0.176	0.158	0.124	0.059	**0.752**
C-3	0.02	0.083	0.133	0.176	0.156	0.118	0.035	**0.721**
C-4	0.012	0.037	0.068	0.106	0.1	0.082	0.023	**0.427**
C-5	0.012	0.059	0.1	0.144	0.135	0.113	0.059	**0.622**
C-6	0.016	0.06	0.1	0.144	0.135	0.113	0.059	**0.628**
C-7	0.02	0.061	0.101	0.144	0.135	0.113	0.059	**0.632**
C-8	0.016	0.06	0.1	0.144	0.134	0.11	0.048	**0.612**
C-9	0.02	0.083	0.133	0.176	0.157	0.121	0.048	**0.737**
C-10	0.016	0.083	0.133	0.176	0.156	0.118	0.035	**0.717**
C-11	0.016	0.083	0.133	0.176	0.156	0.117	0.023	**0.703**
C-12	0.012	0.059	0.1	0.144	0.134	0.11	0.048	**0.607**
C-13	0.016	0.038	0.069	0.106	0.102	0.088	0.048	**0.466**
C-14	0.016	0.038	0.069	0.106	0.1	0.084	0.035	**0.448**
C-15	0.008	0.036	0.068	0.106	0.1	0.082	0.023	**0.422**

Appendix 8

See Table 18.

Table 18 Influence matrix

	C-1	C-2	C-3	C-4	C-5	C-6	C-7	C-8	C-9	C-10	C-11	C-12	C-13	C-14	C-15
X_1															
C-1	0	2	3	3	2	4	2	3	2	2	3	4	3	4	2
C-2	2	0	3	3	2	4	3	2	3	4	2	3	2	1	1
C-3	2	3	0	2	3	2	1	4	2	1	3	1	2	1	3
C-4	2	3	3	0	3	3	2	3	2	3	3	0	3	0	3
C-5	1	2	2	1	0	3	1	2	2	2	2	4	2	4	2
C-6	2	2	2	3	2	0	3	2	2	3	3	3	3	2	2
C-7	2	1	2	3	2	1	0	2	1	1	2	3	2	1	2
C-8	2	2	2	3	2	2	1	0	2	2	2	1	3	2	3
C-9	1	2	3	2	3	1	2	4	0	3	1	2	3	3	2
C-10	2	2	2	3	2	2	2	2	2	0	3	3	0	2	3
C-11	2	1	2	2	2	3	1	2	2	2	0	2	2	2	2
C-12	2	0	1	2	3	2	2	3	2	3	2	0	2	1	2
C-13	2	3	1	2	3	2	2	2	2	0	1	2	0	2	2
C-14	1	2	2	2	2	1	1	2	2	2	2	1	2	0	2
C-15	2	3	1	1	3	2	2	3	2	2	2	1	3	2	0
X_2															
C-1	0	2	3	3	2	4	2	3	2	2	3	4	3	4	2
C-2	2	0	3	3	2	4	3	2	3	4	2	3	2	1	1
C-3	2	3	0	2	3	2	1	4	2	1	3	1	2	1	3
C-4	2	3	3	0	3	3	2	3	2	3	3	0	3	0	3
C-5	1	2	2	1	0	3	1	2	2	2	2	4	2	4	2
C-6	2	2	2	3	2	0	3	2	2	3	3	3	3	2	2
C-7	2	1	2	3	2	1	0	2	1	1	2	3	2	1	2
C-8	2	2	2	3	2	2	1	0	2	2	2	1	3	2	3
C-9	1	2	3	2	3	1	2	4	0	3	1	2	3	3	2
C-10	2	2	2	3	2	2	2	2	2	0	3	3	0	2	3
C-11	2	1	2	2	2	3	1	2	2	2	0	2	2	2	2
C-12	2	0	1	2	3	2	2	3	2	3	2	0	2	1	2
C-13	2	3	1	2	3	2	2	2	2	0	1	2	0	2	2
C-14	1	2	2	2	2	1	1	2	2	2	2	1	2	0	2
C-15	2	3	1	1	3	2	2	3	2	2	2	1	3	2	0
X_3															
C-1	0	2	1	2	1	2	2	2	1	2	2	2	2	1	2
C-2	3	0	2	0	1	3	3	3	2	2	2	2	3	2	1
C-3	3	3	0	1	2	2	2	2	1	1	2	2	2	2	1
C-4	2	4	2	0	2	1	2	2	2	3	1	1	2	2	3
C-5	2	3	2	2	0	2	2	2	2	3	2	2	2	2	3
C-6	4	3	2	2	3	0	3	2	1	2	3	2	2	1	2
C-7	2	2	1	2	1	1	0	1	2	2	1	2	2	1	2
C-8	3	3	3	3	2	2	2	0	3	2	2	2	2	2	3
C-9	2	2	2	2	2	1	2	2	0	2	2	2	2	2	2
C-10	3	3	2	2	2	2	3	3	3	0	2	2	0	2	3
C-11	3	3	2	2	2	2	3	2	1	1	0	2	3	2	2
C-12	4	0	1	1	4	3	3	1	1	2	4	0	3	1	1
C-13	2	3	1	2	2	1	3	2	3	0	2	3	0	2	2
C-14	4	0	1	1	4	1	2	2	2	2	2	1	2	0	2
C-15	2	3	2	2	2	2	2	3	1	2	2	1	3	2	0
X_4															
C-1	0	2	1	2	1	2	2	2	1	2	2	2	2	1	2
C-2	2	0	2	3	2	3	2	4	2	3	1	3	2	2	3

Table 18 (continued)

	C-1	C-2	C-3	C-4	C-5	C-6	C-7	C-8	C-9	C-10	C-11	C-12	C-13	C-14	C-15
C-3	3	3	0	1	2	2	2	2	1	1	2	2	2	2	1
C-4	2	1	2	0	1	2	3	3	2	2	2	2	1	2	1
C-5	2	3	2	2	0	2	2	2	2	3	2	2	2	2	3
C-6	4	3	2	2	3	0	3	2	1	2	3	2	2	1	2
C-7	2	2	1	2	1	1	0	1	2	2	1	2	2	1	2
C-8	3	3	3	3	2	2	2	0	3	2	2	2	2	2	3
C-9	2	2	2	2	2	1	2	2	0	2	2	2	2	2	2
C-10	3	4	2	2	3	2	3	3	3	0	2	2	1	2	4
C-11	3	3	2	2	2	2	3	2	1	1	0	2	3	2	2
C-12	4	0	1	1	4	3	3	1	1	2	4	0	3	1	1
C-13	2	3	1	2	2	1	3	2	3	0	2	3	0	2	2
C-14	4	0	1	1	4	1	2	2	2	2	2	1	2	0	2
C-15	2	3	2	2	2	2	2	3	1	2	2	1	3	2	0
X_5															
C-1	0	1	1	2	1	2	2	2	1	2	2	2	2	1	2
C-2	3	0	3	1	3	2	2	3	4	3	2	3	4	1	3
C-3	2	3	0	1	2	1	2	3	1	2	2	2	1	1	2
C-4	2	0	2	0	2	3	1	1	2	2	1	2	3	2	1
C-5	2	2	2	2	0	3	2	3	2	2	2	2	3	2	2
C-6	1	2	1	2	1	0	2	2	2	2	1	1	2	1	1
C-7	2	4	1	2	3	2	0	3	1	2	3	3	2	2	1
C-8	2	1	2	1	2	3	2	0	1	2	2	2	2	1	2
C-9	2	2	2	2	3	2	2	2	0	2	2	2	3	2	1
C-10	3	3	2	2	2	3	2	3	3	0	2	3	1	2	2
C-11	2	3	2	2	2	2	2	3	1	3	0	3	1	2	2
C-12	1	2	1	2	2	1	2	2	1	3	3	0	2	1	2
C-13	2	2	2	2	2	2	3	3	3	1	2	3	0	1	1
C-14	2	4	1	1	4	2	1	1	2	2	2	2	2	0	1
C-15	3	2	2	2	2	1	1	3	1	3	2	2	2	2	0
X_6															
C-1	0	3	2	3	2	3	2	4	3	1	2	3	4	3	3
C-2	2	0	3	3	4	2	3	4	4	2	1	2	3	4	3
C-3	2	1	0	3	2	3	2	2	1	2	1	1	3	2	3
C-4	2	1	2	0	3	1	3	2	3	2	2	1	2	1	3
C-5	2	1	1	2	0	4	2	1	2	3	2	3	3	2	4
C-6	0	3	2	3	2	0	2	4	3	1	2	3	4	3	3
C-7	1	2	3	2	3	3	0	3	1	2	1	2	3	2	3
C-8	2	2	2	3	1	2	1	0	2	1	2	1	3	2	3
C-9	3	2	3	2	2	1	2	3	0	3	2	1	3	2	1
C-10	1	3	2	1	3	2	3	4	3	0	3	2	2	2	3
C-11	2	2	3	4	3	4	3	2	3	2	0	1	2	1	4
C-12	1	2	1	3	3	3	2	3	2	3	2	0	3	2	2
C-13	2	3	1	2	3	2	2	2	2	2	1	2	0	2	2
C-14	1	3	2	3	2	2	1	1	2	3	3	2	1	0	3
C-15	2	4	2	1	3	3	2	2	3	2	1	2	2	3	0
X_7															
C-1	0	2	1	2	1	2	2	2	1	2	2	2	2	1	2
C-2	3	0	2	0	1	3	3	3	2	2	2	2	3	2	1
C-3	3	3	0	1	2	2	2	2	1	1	2	2	2	2	1
C-4	2	4	2	0	2	1	2	2	2	3	1	1	2	2	3
C-5	2	3	2	2	0	2	2	2	2	3	2	2	2	2	3

Table 18 (continued)

	C-1	C-2	C-3	C-4	C-5	C-6	C-7	C-8	C-9	C-10	C-11	C-12	C-13	C-14	C-15
C-6	4	3	2	2	3	0	3	2	1	2	3	2	2	1	2
C-7	2	2	1	2	1	1	0	1	2	2	1	2	2	1	2
C-8	3	3	3	3	2	2	2	0	3	2	2	2	2	2	3
C-9	2	2	2	2	2	1	2	2	0	2	2	2	2	2	2
C-10	3	3	2	2	2	2	3	3	3	0	2	2	0	2	3
C-11	3	3	2	2	2	2	3	2	1	1	0	2	3	2	2
C-12	4	0	1	1	4	3	3	1	1	2	4	0	3	1	1
C-13	2	3	1	2	2	1	3	2	3	0	2	3	0	2	2
C-14	4	0	1	1	4	1	2	2	2	2	2	1	2	0	2
C-15	2	3	2	2	2	2	2	3	1	2	2	1	3	2	0

Appendix 9

Variable used in the article.

AHP Methodology

P1 = Pairwise comparison matrix
P2 = Principal matrix
λ_{max} = Average of the elements of P_4.
CI = Consistency index.
CR = Consistency Ratio.
RCI = random consistency Index.
n = Number of elements.
PV = priority Vector.

Fuzzy TOPSIS Methodology

A^+ = Fuzzy ideal solution.
A^- = Fuzzy negative ideal solution.
$Y_{ij} = (d_{ij,}\ e_{ij,}$ and $f_{ij}) =$ Triangular fuzzy number for the linguistic term.
$c^*_{j^*}$ and d^-_j = Benefit criteria and cost criteria respectively.
W = Weight of criteria.
V^* and V^- = Values considered for the ideal and negative ideal solution.
g, h, and i = aThe real numbers.
D^+ = Distance between to fuzzy numbers.
D^- = Distance of rating.
C = Closeness the ideal solution.

DEMATEL Methodology

Z = Average matrix.
x_{ij} = Elements of average matrix.
N = Normalized initial direct-relation matrix.
n_{ij} = Elements of normalized initial direct-relation matrix.
Y = Total relation matrix.
I = Identity matrix.

References

Agrawal S, Singh RK, Murtaza Q (2016) Prioritizing critical success factors for reverse logistics implementation using fuzzy-TOPSIS methodology. J Ind Eng Int 12:15–27. https://doi.org/10.1007/s40092-015-0124-8

Altuntas S, Dereli T (2015) A novel approach based on DEMATEL method and patent citation analysis for prioritizing a portfolio of investment projects. Expert Syst Appl 42:1003–1012. https://doi.org/10.1016/j.eswa.2014.09.018

Amui LBL, Jabbour CJC, de Sousa Jabbour ABL, Kannan D (2017) Sustainability as a dynamic organizational capability: a systematic review and a future agenda toward a sustainable transition. J Clean Prod 142:308–322. https://doi.org/10.1016/j.jclepro.2016.07.103

Bag S, Wood LC, Xu L et al (2020) Big data analytics as an operational excellence approach to enhance sustainable supply chain performance. Resour Conserv Recycl 153:104559. https://doi.org/10.1016/j.resconrec.2019.104559

Barlow M (2013) Real-time big data analytics: emerging architecture. O'Reilly Media, Sebastopol, USA

Belhadi A, Zkik K, Cherrafi A et al (2019) Understanding Big Data Analytics for manufacturing processes: insights from literature review and multiple case studies. Comput Ind Eng 137:106099. https://doi.org/10.1016/j.cie.2019.106099

Bi Z, Cochran D (2014) Big data analytics with applications. J Manag Anal 1:249–265. https://doi.org/10.1080/23270012.2014.992985

Braganza A, Brooks L, Nepelski D et al (2017) Resource management in big data initiatives: Processes and dynamic capabilities. J Bus Res 70:328–337. https://doi.org/10.1016/j.jbusres.2016.08.006

Brinch M, Stentoft J, Jensen JK (2017) Big data and its applications in supply chain management: findings from a delphi study. Hawaii Int Conf Syst Sci. https://doi.org/10.24251/HICSS.2017.161

Chen M, Mao S, Liu Y (2014) Big data: a survey. Mob Networks Appl 19:171–209. https://doi.org/10.1007/s11036-013-0489-0

Cui Y, Kara S, Chan KC (2020) Manufacturing big data ecosystem: a systematic literature review. Robot Comput Integr Manuf 62:101861. https://doi.org/10.1016/j.rcim.2019.101861

Das R, Shaw K, Irfan M (2020) Supply chain network design considering carbon footprint, water footprint, supplier's social risk, solid waste, and service level under the uncertain condition. Clean Technol Environ Policy 22:337–370. https://doi.org/10.1007/s10098-019-01785-y

de Camargo FP, Roman Pais Seles BM, Chiappetta Jabbour CJ et al (2018) Management theory and big data literature: from a review to a research agenda. Int J Inf Manage 43:112–129. https://doi.org/10.1016/j.ijinfomgt.2018.07.005

Del Fabbro E, Santarossa D (2016) Ergonomic Analysis in Manufacturing Process A real time approach. Procedia CIRP 41:957–962. https://doi.org/10.1016/j.procir.2015.12.056

Demirkan H, Delen D (2013) Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud. Decis Support Syst 55:412–421. https://doi.org/10.1016/j.dss.2012.05.048

Doolun IS, Ponnambalam SG, Subramanian N, G K, (2018) Data driven hybrid evolutionary analytical approach for multi objective location allocation decisions: automotive green supply chain empirical evidence. Comput Oper Res 98:265–283. https://doi.org/10.1016/j.cor.2018.01.008

Duan Y, Edwards JS, Dwivedi YK (2019) Artificial intelligence for decision making in the era of big data–evolution, challenges and research agenda. Int J Inf Manage 48:63–71. https://doi.org/10.1016/j.ijinfomgt.2019.01.021

Dubey R, Gunasekaran A, Childe SJ et al (2016) The impact of big data on world-class sustainable manufacturing. Int J Adv Manuf Technol 84:631–645. https://doi.org/10.1007/s00170-015-7674-1

Dubey R, Gunasekaran A, Childe SJ et al (2019a) Big data analytics and organizational culture as complements to swift trust and collaborative performance in the humanitarian supply chain. Int J Prod Econ 210:120–136. https://doi.org/10.1016/j.ijpe.2019.01.023

Dubey R, Gunasekaran A, Childe SJ et al (2019b) Empirical investigation of data analytics capability and organizational flexibility as complements to supply chain resilience. Int J Prod Res. https://doi.org/10.1080/00207543.2019.1582820

Elgendy N, Elragal A (2014) Big data analytics: a literature review paper. pp 214–227

ElMaraghy HA, Youssef AMA, Marzouk AM, ElMaraghy WH (2017) Energy use analysis and local benchmarking of manufacturing lines. J Clean Prod 163:36–48. https://doi.org/10.1016/j.jclepro.2015.12.026

Fercoq A, Lamouri S, Carbone V (2016) Lean/green integration focused on waste reduction techniques. J Clean Prod 137:567–578. https://doi.org/10.1016/j.jclepro.2016.07.107

Gandhi S, Mangla SK, Kumar P, Kumar D (2015) Evaluating factors in implementation of successful green supply chain management using DEMATEL: a case study. Int Strateg Manag Rev 3:96–109. https://doi.org/10.1016/j.ism.2015.05.001

Gandomi A, Haider M (2015) Beyond the hype: big data concepts, methods, and analytics. Int J Inf Manage 35:137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Gawankar SA, Gunasekaran A, 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. Int J Prod Res 58:1574–1593. https://doi.org/10.1080/00207543.2019.1668070

Gong M, Simpson A, Koh L, Tan KH (2018) Inside out: The interrelationships of sustainable performance metrics and its effect on business decision making: Theory and practice. Resour Conserv Recycl 128:155–166. https://doi.org/10.1016/j.resconrec.2016.11.001

Gunasekaran A, Papadopoulos T, Dubey R et al (2017) Big data and predictive analytics for supply chain and organizational performance. J Bus Res 70:308–317. https://doi.org/10.1016/j.jbusres.2016.08.004

Gupta A, Singh RK, Suri PK (2018) Prioritizing critical success factors for sustainable service quality management by logistics service providers. Vis J Bus Perspect 22:295–305. https://doi.org/10.1177/0972262918786102

Gupta S, Modgil S, Gunasekaran A (2020) Big data in lean six sigma: a review and further research directions. Int J Prod Res 58:947–969. https://doi.org/10.1080/00207543.2019.1598599

Hazen BT, Skipper JB, Ezell JD, Boone CA (2016) Big data and predictive analytics for supply chain sustainability: a theory-driven research agenda. Comput Ind Eng 101:592–598. https://doi.org/10.1016/j.cie.2016.06.030

Hidayanto AN, Abednego N, Aminah S, Sucahyo YG (2015) Analysis of cloud adoption determinants by using BOCR analysis and DEMATEL. Int J Bus Inf Syst 18:221. https://doi.org/10.1504/IJBIS.2015.067265

Ivanov D, Dolgui A, Sokolov B (2019) The impact of digital technology and Industry 4.0 on the ripple effect and supply chain risk analytics. Int J Prod Res 57:829–846. https://doi.org/10.1080/00207543.2018.1488086

Janssen M, van der Voort H, Wahyudi A (2017) Factors influencing big data decision-making quality. J Bus Res 70:338–345. https://doi.org/10.1016/j.jbusres.2016.08.007

Ji-fan Ren S, Fosso Wamba S, Akter S et al (2017) Modelling quality dynamics, business value and firm performance in a big data analytics environment. Int J Prod Res 55:5011–5026. https://doi.org/10.1080/00207543.2016.1154209

Kalema BM, Mokgadi M (2017) Developing countries organizations readiness for big data analytics. Probl Perspect Manag 15:260–270. https://doi.org/10.21511/ppm.15(1-1).2017.13

Kim G, Park CS, Yoon KP (1997) Identifying investment opportunities for advanced manufacturing systems with comparative-integrated performance measurement. Int J Prod Econ 50:23–33. https://doi.org/10.1016/S0925-5273(97)00014-5

Kumar P, Singh RK (2012) A fuzzy AHP and TOPSIS methodology to evaluate 3PL in a supply chain. J Model Manag 7:287–303. https://doi.org/10.1108/17465661211283287

Lavalle S, Lesser E, Shockley R, Hopkins MS, Kruschwitz N (2011) Big data, analytics and the path from insights to value. MIT Sloan Manag Review 52:21–32

Lee I (2017) Big data: dimensions, evolution, impacts, and challenges. Bus Horiz 60:293–303. https://doi.org/10.1016/j.bushor.2017.01.004

Lee J, Lapira E, Bagheri B, Kao H (2013) Recent advances and trends in predictive manufacturing systems in big data environment. Manuf Lett 1:38–41. https://doi.org/10.1016/j.mfglet.2013.09.005

Lee J, Ardakani HD, Yang S, Bagheri B (2015) Industrial big data analytics and cyber-physical systems for future maintenance and service innovation. Procedia CIRP 38:3–7. https://doi.org/10.1016/j.procir.2015.08.026

Li J, Tao F, Cheng Y, Zhao L (2015) Big data in product lifecycle management. Int J Adv Manuf Technol 81:667–684. https://doi.org/10.1007/s00170-015-7151-x

Lima Junior FR, Osiro L, Carpinetti LCR (2014) A comparison between fuzzy AHP and fuzzy TOPSIS methods to supplier selection. Appl Soft Comput 21:194–209. https://doi.org/10.1016/j.asoc.2014.03.014

Lin R-J (2013) Using fuzzy DEMATEL to evaluate the green supply chain management practices. J Clean Prod 40:32–39. https://doi.org/10.1016/j.jclepro.2011.06.010

Lin K-P, Tseng M-L, Pai P-F (2018) Sustainable supply chain management using approximate fuzzy DEMATEL method. Resour Conserv Recycl 128:134–142. https://doi.org/10.1016/j.resconrec.2016.11.017

Luthra S, Govindan K, Kharb RK, Mangla SK (2016) Evaluating the enablers in solar power developments in the current scenario using fuzzy DEMATEL: an Indian perspective. Renew Sustain Energy Rev 63:379–397. https://doi.org/10.1016/j.rser.2016.04.041

Manavalan E, Jayakrishna K (2019) A review of Internet of Things (IoT) embedded sustainable supply chain for industry 4.0 requirements. Comput Ind Eng 127:925–953. https://doi.org/10.1016/j.cie.2018.11.030

Mangla SK, Luthra S, Rich N et al (2018) Enablers to implement sustainable initiatives in agri-food supply chains. Int J Prod Econ 203:379–393. https://doi.org/10.1016/j.ijpe.2018.07.012

Munirathinam S, Ramadoss B (2014) Big data predictive analytics for proactive semiconductor equipment maintenance. In: 2014 IEEE international conference on big data (big data). IEEE, pp 893–902

Mustapha MA, Manan ZA, Wan Alwi SR (2017) Sustainable Green Management System (SGMS)–an integrated approach towards organisational sustainability. J Clean Prod 146:158–172. https://doi.org/10.1016/j.jclepro.2016.06.033

Piyathanavong V, Garza-Reyes JA, Kumar V et al (2019) The adoption of operational environmental sustainability approaches in the Thai manufacturing sector. J Clean Prod 220:507–528. https://doi.org/10.1016/j.jclepro.2019.02.093

Popovič A, Hackney R, Tassabehji R, Castelli M (2018) The impact of big data analytics on firms' high value business performance. Inf Syst Front 20:209–222. https://doi.org/10.1007/s10796-016-9720-4

Raut RD, Mangla SK, Narwane VS et al (2019) Linking big data analytics and operational sustainability practices for sustainable business management. J Clean Prod 224:10–24. https://doi.org/10.1016/j.jclepro.2019.03.181

Saaty TL (1980) The analytic hierarchy process. McGraw-Hill, New York, p 324

Saaty TL (2000) Fundamentals of decision making and priority theory with the analytic hierarchy process. Vol. 6 RWS publications

Samaie F, Javadi S, Meyar-Naimi H, Feshki-Farahani H (2020) Environmental sustainability policy on plug-in hybrid electric vehicle penetration utilizing fuzzy TOPSIS and game theory. Clean Technol Environ Policy. https://doi.org/10.1007/s10098-020-01821-2

Sanders NR (2016) How to use big data to drive your supply chain. Calif Manage Rev 58:26–48. https://doi.org/10.1525/cmr.2016.58.3.26

Singh RK (2012) Justification of coordinated supply chain in small and medium enterprises using analytic hierarchy process. Int J Serv Sci 4:277. https://doi.org/10.1504/IJSSCI.2012.051062

Singh RK (2013) Prioritizing the factors for coordinated supply chain using analytic hierarchy process (AHP). Meas Bus Excell 17:80–97. https://doi.org/10.1108/13683041311311383

Singh RK, Agrawal S (2018) Analyzing disposition strategies in reverse supply chains: fuzzy TOPSIS approach. Manag Environ Qual An Int J 29:427–443. https://doi.org/10.1108/MEQ-12-2017-0177

Singh SK, El-Kassar A-N (2019) Role of big data analytics in developing sustainable capabilities. J Clean Prod 213:1264–1273. https://doi.org/10.1016/j.jclepro.2018.12.199

Singh RK, Luthra S, Mangla SK, Uniyal S (2019) Applications of information and communication technology for sustainable growth of SMEs in India food industry. Resour Conserv Recycl 147:10–18. https://doi.org/10.1016/j.resconrec.2019.04.014

Song M, Fisher R, Kwoh Y (2019) Technological challenges of green innovation and sustainable resource management with large scale data. Technol Forecast Soc Change 144:361–368. https://doi.org/10.1016/j.techfore.2018.07.055

Tan J, Tan RR, Aviso KB et al (2017) Study of microalgae cultivation systems based on integrated analytic hierarchy process–life cycle optimization. Clean Technol Environ Policy 19:2075–2088. https://doi.org/10.1007/s10098-017-1390-5

Tao F, Cheng J, Qi Q et al (2018) Digital twin-driven product design, manufacturing and service with big data. Int J Adv Manuf Technol 94:3563–3576. https://doi.org/10.1007/s00170-017-0233-1

Telukdarie A, Buhulaiga E, Bag S et al (2018) Industry 4.0 implementation for multinationals. Process Saf Environ Prot 118:316–329. https://doi.org/10.1016/j.psep.2018.06.030

Tseng M-L, Lim MK, Wu K-J (2019) Improving the benefits and costs on sustainable supply chain finance under uncertainty. Int J Prod Econ 218:308–321. https://doi.org/10.1016/j.ijpe.2019.06.017

Wamba SF, Dubey R, Gunasekaran A, Akter S (2020) The performance effects of big data analytics and supply chain ambidexterity: the moderating effect of environmental dynamism. Int J Prod Econ 222:107498. https://doi.org/10.1016/j.ijpe.2019.09.019

Wang Z, Xue M, Wang Y et al (2019) Big data: new tend to sustainable consumption research. J Clean Prod 236:117499. https://doi.org/10.1016/j.jclepro.2019.06.330

Weerakkody V, Kapoor K, Balta ME et al (2017) Factors influencing user acceptance of public sector big open data. Prod Plan Control 28:891–905. https://doi.org/10.1080/09537287.2017.1336802

Wilcox T, Jin N, Flach P, Thumim J (2019) A big data platform for smart meter data analytics. Comput Ind 105:250–259. https://doi.org/10.1016/j.compind.2018.12.010

Wood LC, Wang C, Abdul-Rahman H, Jamal Abdul-Nasir NS (2016) Green hospital design: integrating quality function deployment and end-user demands. J Clean Prod 112:903–913. https://doi.org/10.1016/j.jclepro.2015.08.101

Yaqoob I, Hashem IAT, Gani A et al (2016) Big data: from beginning to future. Int J Inf Manage 36:1231–1247. https://doi.org/10.1016/j.ijinfomgt.2016.07.009

Zhong RY, Huang GQ, Lan S et al (2015) A big data approach for logistics trajectory discovery from RFID-enabled production data. Int J Prod Econ 165:260–272. https://doi.org/10.1016/j.ijpe.2015.02.014

Zhong RY, Newman ST, Huang GQ, Lan S (2016) Big Data for supply chain management in the service and manufacturing sectors: Challenges, opportunities, and future perspectives. Comput Ind Eng 101:572–591. https://doi.org/10.1016/j.cie.2016.07.013

Zimmermann H-J (1985) Applications of fuzzy set theory to mathematical programming. Inf Sci (Ny) 36:29–58. https://doi.org/10.1016/0020-0255(85)90025-8

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

Received: 22 December 2019 | Revised: 30 September 2020 | Accepted: 8 October 2020

IET Renewable Power Generation

DOI: 10.1049/rpg2.12061

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH PAPER

Cascade-I$^{\lambda}$D$^{\mu}$N controller design for AGC of thermal and hydro-thermal power systems integrated with renewable energy sources

Yogendra Arya[1] | **Nishant Kumar**[2] | **Pankaj Dahiya**[3] | **Gulshan Sharma**[4] | **Emre Çelik**[5] | **Sandeep Dhundhara**[6] | **Mandeep Sharma**[7]

[1] Department of Electrical and Electronics Engineering, Maharaja Surajmal Institute of Technology, Janakpuri, New Delhi, India

[2] Electrical and Computer Engineering Department, Postdoctoral Research Fellow, National University of Singapore, Singapore

[3] Department of Electronics & Communication Engineering, Delhi Technological University, Shahbad Daulatpur Village, Rohini, Delhi, Delhi, India

[4] Department of Electrical Power Engineering, Durban University of Technology, Durban, South Africa

[5] Department of Electrical and Electronics Engineering, Engineering Faculty, Düzce University, Düzce, Turkey

[6] Department of Basic Engineering, College of Agricultural Engineering and Technology, CCS Haryana Agricultural University, Hisar, India

[7] Department of Electrical Engineering, Baba Hira Singh Bhattal Institute of Engineering and Technology, Lehragaga, Punjab, India

Correspondence
Yogendra Arya, Department of Electrical and Electronics Engineering, Maharaja Surajmal Institute of Technology, Janakpuri, New Delhi, India.
Email: mr.y.arya@gmail.com

Abstract

Expert, intelligent and robust automatic generation control (AGC) scheme is requisite for stable operation and control of power system (PS) integrated with renewable energy sources (RES) under sudden/random small load disturbances. Large frequency deviations appear if AGC capacity is inept to compensate for the imbalances of generation and demand. In this paper, a cascade-fractional order ID with filter (C-I$^{\lambda}$D$^{\mu}$N) controller is proposed as an expert supplementary controller to promote AGC recital adequately in electric power systems incorporated with RES like solar, wind and fuel cells. The imperialist competitive algorithm is fruitfully exploited for optimizing the controller parameters. First, a 2-area reheat thermal system is examined critically and then to authorize the worth of the proposed controller, the study is protracted to a 2-area multi-source hydro-thermal system. The prominent benefits of C-I$^{\lambda}$D$^{\mu}$N controller with/without renewable energy sources consist of its great indolence to large load disturbances and superiority over various optimized classical/fuzzy controllers published recently. The sensitivity study validates the robustness of the recommended controller against ±20% deviations in the system parameters and random step load perturbations.

1 | INTRODUCTION

The generation control is progressively turning challenging due to the gradual enhancement in the power demands and depletion of power resources. The expanding power demands are posing severe dangers to the stable conduct of power system. The growing power demand causes sluggishness in turbine speed and hence, frequency in output voltage of the alternator decline undesirably. In an interconnected large-scale power system (PS), all the synchronously connected generation units perform with the common frequency. It is desirable to sustain frequency/scheduled tie power flows irrespective of load variations in control areas of interconnected PS during normal operating and small abrupt disturbance conditions. To carry out this, an efficient and intelligent automatic generation control (AGC) strategy is vital to spontaneously manipulate the operation of hydro or steam valves which regulates real output power of generators [1].

The PS is turning complicated day-by-day, because of upsurge in its dimension and mixing of renewable energy sources (RES). Because of the ongoing depletion of fossil fuels and the environmental hazards caused by them, the

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2021 The Authors. *IET Renewable Power Generation* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology

infiltration of RES in modern power system is attracting electricity industry. The fascinating features of RES are that they are environmentally-friendly and ample in nature. Furthermore, due to the advancement in technology and bulk production, their energy conversion efficiency is swelling and the cost is shrinking year by year [2]. However, the RES such as solar and wind are extensively accessible in nature and dominate other RES. Hence, the penetration of wind and solar into the prevailing power systems has constantly been a prominent area of curiosity for researchers. Rahman et al. studied the impact of a dish-Stirling solar thermal system in AGC of a 2-area reheat thermal system (RTS) [2]. Kler et al. conducted an AGC study on a 2-area RTS having RES like wind turbine generators (WTGs), solar thermal generators (STGs) and aqua electrolyser (AE)-fuel cells (FCs) units [3]. An independent hybrid generating system comprising of WTG, STG, solar photovoltaic (PV), diesel engine generators, AE, FC, battery energy storage system (BESS), flywheel energy storage system and ultracapacitors is studied comprehensively in [4]. The recent literature presented a critical study on the impact of AE-FC units on AGC performance of interconnected 2-area non-reheat thermal system, PV-reheat thermal and multi-source hydro-thermal system (MSHTS) [5]. Further, the effects of STG and geo thermal (GT) plants are recently inspected in 2 and 3-area deregulated systems [6, 7]. An islanding smart distribution network heavily penetrated with renewable sources is demonstrated in [8]. Hubei power grid composed of thermal, pumped storage, hydro and distributed renewable energy sources (DREs) is inspected in [9]. Guangdong power grid having thermal, hydro, wind, PV and electric vehicles is studied in [10]. A 3-area micro grid system having various DREs is proposed in [11]. Consequently, various RES systems are studied in different AGC system. But, there still exists scope for AGC of interconnected power systems with integration of RES such as STGs, WTGs and AE-FC units. This necessities additional examination as tried here.

The competence of an AGC system is stated in terms of brisk and stress-free control action. With the integration of RES, power system requires robust and efficient AGC strategy to withstand its controllability and stability in the presence of stochastic nature of RES. To boost AGC outcome, different control and tuning approaches are available in the recent literature. Some of them implemented in various traditional and restructured systems are classical [1], minority-charge carrier inspired algorithm optimized I/PI/ID [1], biogeography based optimisation based 3-degree-of-freedom-PID (3DOF-PID) [2], genetic algorithm (GA) based integral minus proportional derivative (IPD)/PI/PID [3, 4], imperialist competitive algorithm (ICA) based fuzzy-tilt-I-D with filter-double integral (FTIDN-II) [5], sine–cosine algorithm (SCA) tuned cascade fractional order (FO) PI-FOPIDN [6], SCA tuned cascade FOPI-FOPDN [7], wolf pack hunting strategy based control [8], multi-agent double deep Q network-action discovery (AD) based control [9], deep policy dynamics based win or learn fast-policy hill climbing (PDWoLF-PHC) network based control [10], PDWoLF-PHC(λ) strategy [11, 12], deep-reinforcement-learning-based three-network double-delay actor-critic control strategy [13], SCA optimized proportional derivative-proportional integral derivative with double derivative filter (PDPID + DDF) [14], grey wolf optimization (GWO) optimized PI/PID [15], artificial bee colony algorithm (ABCA) optimized PI/PID [16], hybrid firefly algorithm-pattern search (hFA-PS) technique optimized PI/PID [17], ICA optimized PID [18], jaya algorithm-invasive weed optimization (JA-IWO) optimized PID [19], bacterial swarm optimization (BSO) optimized PID/FOPID [20], optics inspired optimization (OIO) optimized PID [21], symbiotic organisms search (SOS) algorithm optimized PID/PIDN [22, 23], quasi-oppositional differential search algorithm (QODSA) optimized PI/PID [24], hybrid IWO-PS (hIWO-PS) tuned 2DOF-PID [25] and whale optimization algorithm (WOA) optimized cascade PIDN-FOPD [26]. Since, the operating conditions of power system are liable to vary widely over the time due to wearing out of the components, the conventional controllers optimized for fixed operating condition might work inappropriately in changed operating condition. To overcome this difficulty and to get excellent control of generation/frequency/tie power in any interconnected PS, intelligent control system like fuzzy logic controller (FLC) is recommended over conventional one [5].

FLC can handle small deviation in system operating condition efficiently by automatically adjusting the controller parameters online. With the aid of FLC, diverse intelligent conventional controllers are proposed recently [5, 27–41]. The choice of FLC parameters like input/output scaling factors (SFs), control rules (CRs), membership functions (MFs) etc. is an intricate assignment. Classically, their selection depends on operator's experiential rules and usually are non-optimal. The inappropriate assortment of FLC parameters might impact the system results adversely. Hence, to overcome this problem, various intelligent optimization methods are applied fruitfully in different types of PSs for tuning the parameters of FLC having different configurations. The procedure of concurrent optimization of all FLC parameters might be a difficult, costly and time consuming task. Hence, prevalent trend of FLC optimization is the tuning/setting of input/output SFs only and MFs/CRs are kept fixed for dissimilar models as getting their values is simpler; and also they impact the system results to a great extent [5, 27–41]. Recently, hybrid particle swarm optimization-pattern search (hPSO-PS) [27] and bacterial foraging optimization algorithm (BFOA) [28] techniques are adopted to tune input/output SFs and ICA [29] to tune output SFs of fuzzy PI (FPI) controller. But, FPI structure may probably show substandard performance due to its inherent integral action. To conquer this drawback, a derivative mode is attached in PI configuration to get FPID structured controller. To improve FPI performance, some researchers have implemented different optimization methods to tune FPID controller. BFOA [28]/hybrid PSO-levy flight algorithm (hPSO-LFA) [30]/hybrid harmony search-cuckoo optimization algorithm (hHS-COA) [31]/hybrid improved FA-PS (hIFA-PS) [32] and hybrid differential evolution-PSO (hDE-PSO) [33] are employed efficiently to tune input/output SFs of FPID controller for different types of systems. In FPID controller, since derivative block is affected by higher frequency noise in area control error (ACE) fed to FPID, a first order filter is employed in derivative block to

attenuate higher frequency noise to get FPID with filter (FPIDN) controller. In the literature, some researchers have employed FPIDN controller with input/output SFs selected using hybrid local unimodal sampling-teaching learning based optimization (hLUS-TLBO) [34]/jaya algorithm (JA) [35] algorithm and it shows better performance than FPID controller.

Fractional order (FO) controllers are earning fame in the current years due to their extra flexibility and skill to handle control design specifications. FO controllers are verified to be superior than their integer order (IO) equivalents [36–41]. Recently, FO controllers are implemented in PS models and desirable results are observed. Further, the literature review indicates that the hybridization of fuzzy controller with FO controller shows enhanced performance compared to fuzzy hybridized with conventional controller. A FO fuzzy PID (FOFPID) controller tuned recently using ICA [36], BFOA [37, 38] and quasi-oppositional harmony search algorithm (QOHSA) [39] displays healthier results compared to their conventional counterparts for different single/multi-area PSs. Further, some intelligent cascade controllers like SCA based FOPI-FOPID [6]/FOPI-FOPD [7], WOA optimized PIDN-FOPD [26] and ICA based FFOPI-FOPD [40]/FOPID [41] are appeared in the literature. However, the performance assessment of cascade fuzzy-FOIDN (C-I$^{\lambda}$D$^{\mu}$N) structured controller has not been explored so far. Considering these facts, a C-I$^{\lambda}$D$^{\mu}$N controller is recommended for 2-area PS models. Besides the controller construction, system performance also relies on the suitable choice of optimization technique to tune the controller parameters. Due to better convergence rate, scarcity in AGC applications, robust control performance of ICA against nonlinearities, constraints and parameters uncertainties, the optimal adjustment of C-I$^{\lambda}$D$^{\mu}$N controller parameters is attained by adopting ICA.

Considering the above analysis, the novel contributions of this work are stated as follows:

(i) To propose a novel ICA tuned C-I$^{\lambda}$D$^{\mu}$N structured controller for AGC of interconnected 2-area RES penetrated RTS and MSHTS.
(ii) To authenticate the strength of the proposed method compared to the various intelligent control strategies existing in the state-of-the-art literature.
(iii) To exemplify the returns of RES penetration in RTS and MSHTS.

1.1 | Power system models

The AGC models investigated are 2-area RTS displayed in Figure 5(a) [3, 16, 20, 22, 28, 33] and 2-area MSHTS displayed in Figure 8 [15, 17, 28, 29, 34]. The systems consist of generating areas of equal sizes. The RTS consists of one single reheat type of turbine and MSHTS comprises of one non-reheat and one hydro turbines in each area. Both systems in area-1 embrace renewable energy sources (RES) based systems involving of 2 solar thermal generators (STGs), 3 wind turbine generators (WTGs), 1 aqua-electrolyser (AE) and 2 fuel cells (FCs). As this study is conducted in low frequency domain, the transfer functions (TFs) for these RES are denoted by first order lag [2–5]. The STG entails solar and thermal parts. The solar part comprises of parabolic collectors which targets the sunlight on to the thermal part to generate steam to run the turbine [3]. The AE is exploited to engage portion of the inconstant power produced by STGs and WTGs to generate hydrogen [3, 5]. Analogous to STGs and WTGs, the FC is also characterized via first order lag-based TF having time constant obtained experimentally [3]. The detailed RES system is demonstrated in Figure 5(b) [3].

Every control area comprises its own speed governing system, turbine and generator. There are three inputs and two outputs for an area. The inputs are control input that is, ΔP_C, change in load demand that is, ΔP_D, and change in tie power that is, ΔPtie. One extra input in area-1 exists due to RES output that is, ΔP_{gRES}. The outputs consist of change in area frequency (ΔF) and area control error (ACE). The test models are picked up from the literature stated above. The data used for the systems have been taken from the stated references and given in Appendix. The Nomenclature describes various symbols used.

2 | CONTROLLER STRUCTURE

The major intentions of a controller designed for an AGC system are to force frequency, generation and tie power variations to zero with least over/under shoots as swiftly as possible under sudden/step load perturbation (SLP). Under SLP, the resultant area control error (ACE) is employed to call upon the controller action to curtail ACE to zero in possible minimum time in each area. Conventional PID controller is favoured in the majority of PS industry due its simple design, robust performance, inexpensive and effective potential for use in the engineering disciplines. Though, because of the inherent intricacy of the extended PS, presence of nonlinearities, uncertainties and their combined impacts, PID controllers become comparably ineffectual to provide a desired level of resilience and outcome in changing operating situations of large scale PS. On the other hand, PID with fuzzy logic (FL) has demonstrated better flexibility and performance to achieve the challenges in PS control. The FL combined with PI/PID/PIDN/FOPID controller is recommended in majority of the recently appeared AGC literature [5, 27–41]. But, cascade type of fuzzy controller has not been discussed more. Hence, in this paper, a cascade fuzzy FOIDN (C-I$^{\lambda}$D$^{\mu}$N) controller is recommended for AGC of multi-area PSs as an expert, robust and intelligent controller. The controller construction is exposed in Figure 1. It consists of a FPIDN and I$^{\lambda}$D$^{\mu}$N controllers connected in cascade. The FPIDN has area control area (ACE) and derivative of ACE as two inputs. The area-1 (ACE$_1$) and area-2 (ACE$_2$) ACEs are given by

$$\left.\begin{aligned} \mathrm{ACE}_1 &= B_1\Delta F_1 + \Delta P\mathrm{tie}_{12} \\ \mathrm{ACE}_2 &= B_2\Delta F_2 + \alpha_{12}\Delta P\mathrm{tie}_{12} \end{aligned}\right\} \quad (1)$$

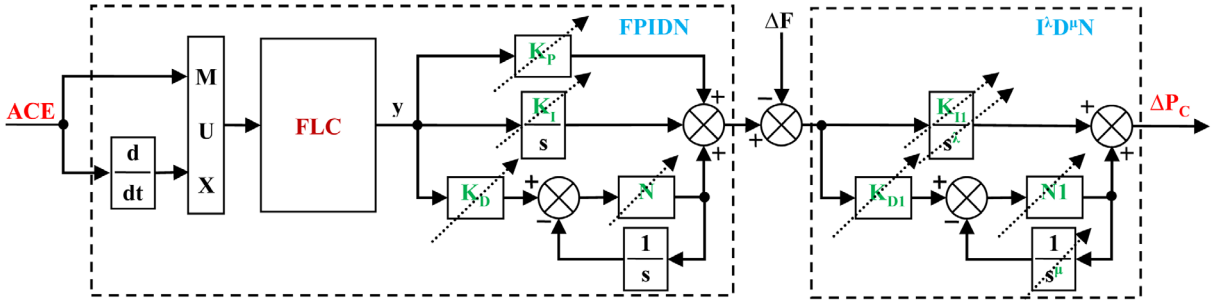


FIGURE 1 Configuration of C-$I^\lambda D^\mu N$ controller

where, B_i is the frequency bias, ΔF_i is area frequency deviation, $\Delta P\text{tie}_{ij}$ is tie tie flow deviation, α_{12} is the control area rating ratio. The FLC output y performances as the input to PIDN controller. The output of C-$I^\lambda D^\mu N$ controller is the control input ΔP_C for the PS as represented by Equation (2).

$$\Delta P_C = \left[\left\{\left(yK_P + \frac{yK_I}{s} + \frac{ysNK_D}{s+N}\right) - \Delta F\right\} \times \left(\frac{K_{I2}}{s^\lambda} + \frac{s^\mu N1 K_{D1}}{s^\mu + N1}\right)\right] \tag{2}$$

Fractional calculus (FC) is expansion of the nth order continuous differentiation/integration of an arbitrary function with a real value order. The notions of FO controller connect with differential equations employing FC. The FOPID/FOPIDN controller is the expansion of the integer order (IO) PID/PIDN controller with fractional choice of integral-differential orders along with the IO PID/PIDN controller gains. The transfer function (TF) portrayal of FOPID ($PI^\lambda D^\mu$)/FOIDN ($I^\lambda D^\mu N$) controller is given as:

$$C(s) = K_P + \frac{K_I}{s^\lambda} + K_D s^\mu \ (PI^\lambda D^\mu) \tag{3}$$

$$C(s) = \frac{K_{I1}}{s^\lambda} + \frac{K_{D1} s^\mu N1}{s^\mu + N1} \ (I^\lambda D^\mu N) \tag{4}$$

In Equations (2) and (4), N and $N1$ denote coefficients of filter. The $N/N1$ curtails the impact of high frequency noise that appears because of regular switching of loads. In this work a well-known approximation termed as CRONE approximation, proposed by Oustaloup is utilized [6]. CRONE practices a recursive dispersal of A poles and A zeros leading to a TF inside a pre-specified frequency limit [ω_L, ω_H]. For computer simulations, the considered ω_L, ω_H and A are 0.01 rad/s, 100 rad/s and 6, in order. For further details of FO controllers, [6, 7, 20, 26, 36–41] can be referred.

The controller gains of C-$I^\lambda D^\mu N$ controller are to be selected via an appropriate method. Here, seven membership functions (MFs) are adopted for each inputs/output. The innermost five MFs have triangular shapes and the outer two have trapezoidal shapes as showcased in Figure 2 [29]. The rule base is utilized from [29]. In Figure 2, the fuzzy linguistic variables L, N, M, S, Z and P signify large, negative, medium, small, zero and positive, in order. The unfuzzified output of FLC is decided via center of

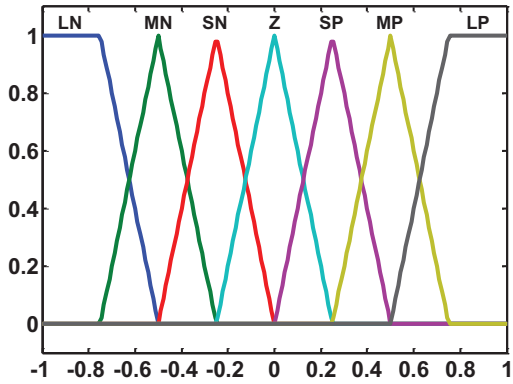


FIGURE 2 MFs for FLC inputs/output [29]

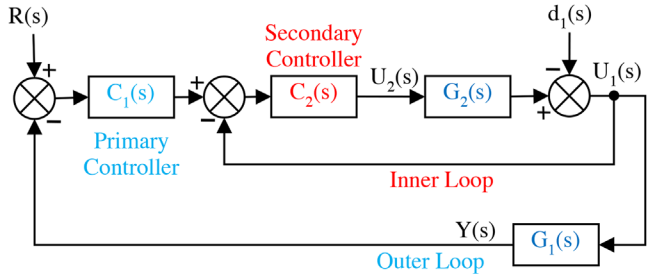


FIGURE 3 Cascade controller model

gravity defuzzification method and rule base is fired adopting Mamdani fuzzy inference system (FIS).

Handling two subsequent processes gives the notion of cascade control (CC). CC can advance control system outcome compared to single-loop control. CC is effective in fast disturbance rejection and provides fast/enhanced system response [6, 41]. The basic CC system includes two control loops as revealed in Figure 3. In Figure 3, primary/master/outer controller $C_1(s)$ is FPIDN and secondary/slave/inner controller $C_2(s)$ is FOIDN. The primary process is termed as $G_1(s)$ and the complete process is exposed to load disturbance $d_1(s)$ creating $Y(s)$ output. The input to $G_1(s)$ is $U_1(s)$, which is the inner process output. $Y(s)$ is organized in reference to $R(s)$. $U_2(s)$ is the input to the secondary loop process $G_2(s)$. The

major purpose of the secondary loop is to bound the result of secondary process gain deviations on the control system outcome [6, 7]. These gain changes might originate from variations in operating conditions.

3 | IMPERIALIST COMPETITIVE ALGORITHM (ICA)

ICA, a meta-heuristic evolutionary algorithm stimulated by socio-political practices of imperialistic competition of people in the physical world was established by Atashpaz–Gargari and Lucas [42]. It establishes abundant recital both in convergence rate and global optimal solution achievement. It commences with early population of countries segregated into two groups; imperialists having the best/lowest values of fitness/cost function and other countries with less power named as colonies of the imperialists and all together constitute few initial empires. In the initial population, the colonies attached with the imperialists based on imperialists' powers so that the imperialist with extra power has extra colonies. The empire's power depends on the powers of the imperialist and its colonies. The colonies begin to vary their societies as per relevant imperialist. Here, colonies start moving regarding their relevant imperialist and this process is called assimilation. In it, the colony having more power may swap the imperialist with less power and vice versa.

During the imperialistic competition among the empires; the powerful empires try to boost their powers. The empires which are not powerful enough to compete and not able to raise their power will crumble. The competition will slightly enhance the power of powerful empires and decrease the power of less-powerful empires. Due to this, during imperialistic competition, feeble empires will gradually lose their powers and breakdown at last. At the completion of competition, countries converge to a condition where there exists only one empire, all countries act as colonies of this empire and these colonies occupy the equivalent position/power as the imperialist.

The main steps concerned with the decision regarding the optimal controller parameters via ICA are specified as follows.

3.1 | Initialize the empires

In this step, an array of variables (country) to be optimized is decided as:

$$\text{Country} = \left[p_1, p_2, p_3, \ldots, p_n\right], n = 1, 2, \ldots, N_{\text{var}} \tag{5}$$

where, p_i are variables to be tuned. The country's cost is determined as:

$$\text{cost} = \text{f}\left(\text{country}\right) = \text{f}(p_1, p_2, p_3, \ldots, p_n) \tag{6}$$

To shape the early empires, the colonies are distributed among imperialist evenly using the imperialist's normalized cost

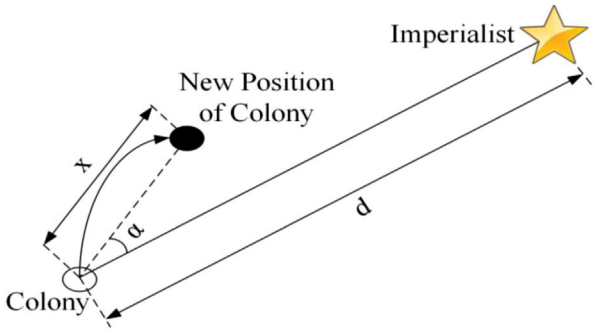


FIGURE 4 Shifting of a colony to its related imperialist [37]

described as:

$$\text{C}_n = \text{C}_n - \max_i (\text{C}_i) \tag{7}$$

here, C_n is the nth imperialist's cost and C_n denotes its normalized cost. The normalized power of each imperialist is obtained as:

$$P_n = \frac{|\text{C}_n|}{\left|\sum_{i=1}^{N_{\text{imp}}} \text{C}_i\right|} \tag{8}$$

here, N_{imp} denotes the powerful countries to be imperialists to constitute empires. The nth empire's colonies can be given as:

$$N \times \text{C}_n = \text{round}(P_n \times N_{\text{col}}) \tag{9}$$

here, $N \times \text{C}_\text{n}$ denotes nth empire's preliminary colonies and N_{col} is the total number of colonies.

3.2 | Assimilation

The colonies of every imperialist are assimilated regarding their particular imperialist as:

$$x_{\text{col}}^{\text{new}} = x_{\text{col}}^{\text{old}} + \beta \times x \otimes \left(x_{\text{imp}} - x_{\text{col}}^{\text{old}}\right) \tag{10}$$

here, $x_{\text{col}}^{\text{old}}$ depicts colony's old position, β denotes assimilation factor and x is a random variable with even distribution. X_{imp}, $X_{\text{col}}^{\text{new}}$ are the imperialist's position and colony's new position, in order. Figure 4 displays the drive of a colony near its related imperialist. The progress is explained via x (distance) and α (angle) chosen randomly with even distribution (11-12).

$$x = \text{u}(0, \beta \times \text{d}) \tag{11}$$

$$\alpha = u\,(-\eta, \eta) \tag{12}$$

where $\beta > 1$, d denotes the distance between colony-imperialist and η is a limit angle for adjusting departure from the original direction [42].

TABLE 1 Optimized gains of controllers

	2-area reheat thermal system		2-area multi-source hydro-thermal system		
Controller gains	**C-I$^{\lambda}$D$^{\mu}$N**	**C-I$^{\lambda}$D$^{\mu}$N (RES)**	**FPIDN**	**C-I$^{\lambda}$D$^{\mu}$N**	**C-I$^{\lambda}$D$^{\mu}$N (RES)**
K_P	4.3654	4.8625	2.3021	2.3645	2.8599
K_I	9.5127	9.3283	7.8988	4.6854	5.9901
K_D	0.1354	0.2531	0.0894	0.1354	0.1845
N	482.3548	397.2634	250.2364	409.6856	395.8695
K_{I1}	7.3524	9.9289	–	4.3655	5.9649
λ	0.3395	0.2595	–	0.5528	0.5027
K_{D1}	0.5920	0.8121	–	0.8369	0.8082
μ	0.3843	0.3495	–	0.0525	0.0687
$N1$	468.9861	406.4528	–	463.6857	426.5469

3.3 | Switching positions

During drive in the direction of the imperialist, a colony may arrive at a station having lesser cost value compared to the imperialist. In this case, the position of imperialist and colony are interchanged and then colonies commence moving near this new location.

3.4 | Total cost of empires

An empire comprises of the imperialists and colonies. Hence, the cost of an empire is given as:

$$T \times \mathrm{C}_n = \mathrm{f}(\text{imperialist}_n) + \xi \times \text{mean}(\{\mathrm{f}(\text{colonies of empires}_n)\} \tag{13}$$

here, $T \times C_n$ is the nth empire's total cost and ξ is a positive number < 1. A small value of ξ implies that the empire's power is due to only the imperialist. However, a high value of ξ implies that the empire's total power is due to mainly the colonies.

3.5 | Imperialist competition

In imperialist competition (IC), the power of frailer empires gradually decreases and that of powerful ones increases. This practice embraces the competition between imperialists to snatch colonies associated with other imperialists and this is performed by choosing the weakest colony from the weakest empire and creating a competition among all empires to own this colony. The most dominant empire conquers the selected colony. To initialize the IC, primary, the possession probability

(PP) of every empire is calculated on the basis of its total power. The normalized cost of nth empire is computed by

$$NT \times \mathrm{C}_n = T \times \mathrm{C}_n - \max_i (T \times \mathrm{C}_i) \tag{14}$$

here $NT \times C_n$ is the nth empire's normalized cost. The PP of every empire is calculated as:

$$P_{\mathrm{pn}} = \frac{|NT \times \mathrm{C}_n|}{\left|\sum_{i=1}^{N_{\mathrm{imp}}} NT\mathrm{C}_i\right|} \tag{15}$$

After computing the PP vector, additional vector termed D having the equal size is expressed by subtracting R (a random vector) from the PP vector ($D = P - R$), the possessed colony owns the highest value of D.

3.6 | Revolution

In each iteration, in addition to assimilation, the revolution process is utilized to shift the colony positions. Few weakest colonies are selected and replaced with new colonies, haphazardly. Revolution is a sudden change, which improves the exploitation and exploration processes. It creates variety during searching of solution space. Comparable to the mutation process in GA, revolution advances the global convergence and thwart ICA from being trapped in local minimum. The number of colonies presumed to be interchanged with the equal number of new colonies is:

$$N_{\mathrm{RC}} = \text{round}\{\mathrm{P}_{\text{revolution}} \times \text{No.(colonies of empire}_n)\} \tag{16}$$

here N_{RC} denotes revolutionary colonies and $P_{\text{revolution}}$ is the revolution rate. Prior the movement of each colony, a random number having constant distribution [0,1] is originated and matched with $P_{\text{revolution}}$. If this number is less than $P_{\text{revolution}}$, revolution process is adopted and the recent colony will substitute with the previous one.

3.7 | Eliminating powerless empires

In the imperialist competition, feeble empires will breakdown and colonies associated with them will be allocated among remaining empires. An empire is expected to downfall when it drops all of its colonies.

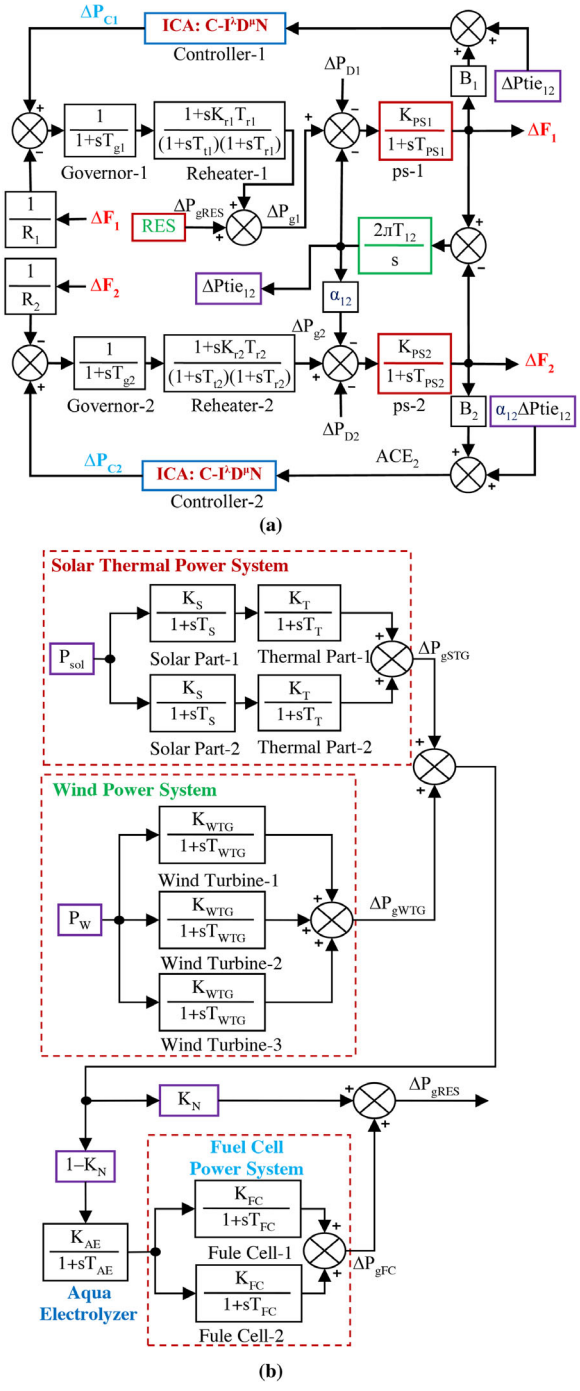


FIGURE 5 (a) 2-area reheat thermal system model [28] and (b) renewable energy sources model [3]

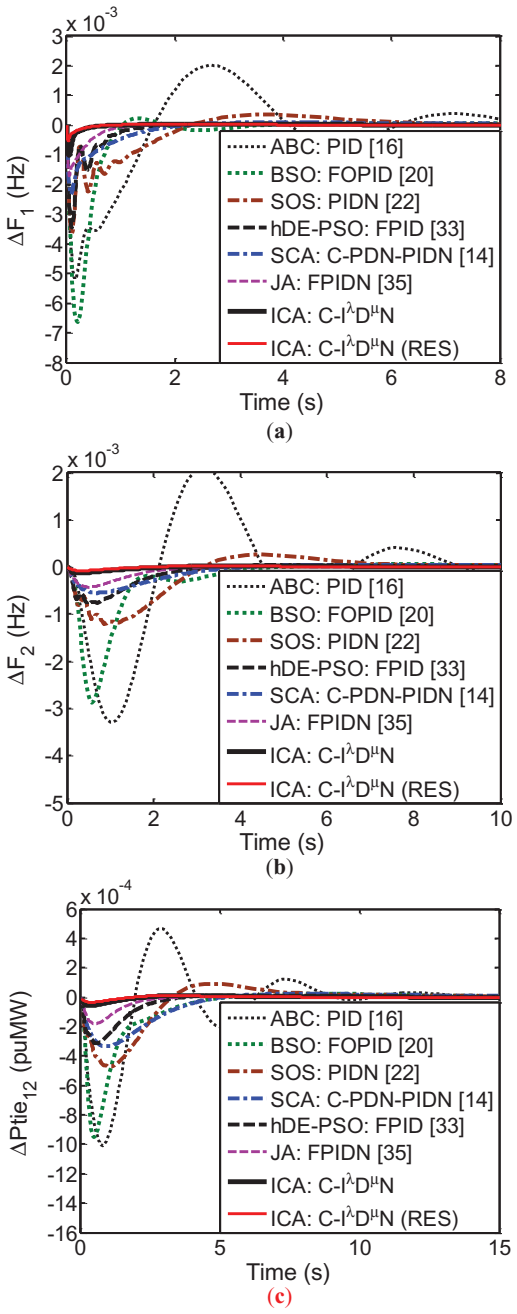


FIGURE 6 2-area reheat thermal system responses (a) ΔF_1, (b) ΔF_2 and (c) $\Delta Ptie_{12}$

3.8 | Convergence

The preceding steps are replicated till the criterion decided for convergence is attained. At the convergence point all empires breakdown except the utmost powerful one. In this condition, all colonies occupy identical positions and costs; competition ends and ICA stops. For more detail please refer [18, 42]. The parameters used in ICA are taken from [29].

4 | OPTIMIZATION CONSTRICTIONS AND OBJECTIVE FUNCTION

To tune the parameters of a controller utilizing an optimization algorithm, an objective/cost function is needed. A cost function (J) is designed based on the desired provisions and limitations. Some of the industry-linked cost functions are integral time absolute error (ITAE), integral square error (ISE), integral time square error (ITSE) and integral absolute error (IAE) [3, 6, 14, 28, 29, 36, 38, 40, 41]. However, in this paper, ISE is utilized because of its better performance [28, 29, 36–38, 40, 41].

TABLE 2 $T_S/U_S/O_S/J_S$ for 2-area reheat thermal system results shown in Figure 6(a–c)

	T_S (s)			U_S (−ve) (Hz)		U_S (−ve) (puMW)	O_S (Hz)		O_S (puMW)	J_S (puMW)			
Controllers	ΔF_1	ΔF_2	$\Delta P\text{tie}_{12}$	ΔF_1	ΔF_2	$\Delta P\text{tie}_{12}$	ΔF_1	ΔF_2	$\Delta P\text{tie}_{12}$	**ISE**	**ITSE**	**IAE**	**ITAE**
ABC: PID [16]	5.39	6.01	1.48	0.0051	0.00329	0.001008	2.01×10^{-3}	2.09×10^{-3}	4.66×10^{-4}	3.62×10^{-5}	5.73×10^{-5}	0.0203	0.0572
SOS: PIDN [22]	1.81	2.29	1.01	0.0036	0.00120	0.000471	3.56×10^{-4}	2.65×10^{-4}	9.15×10^{-5}	6.57×10^{-6}	6.62×10^{-6}	0.0079	0.0203
BSO: FOPID [20]	0.83	1.41	1.05	0.0066	0.00287	0.000949	2.37×10^{-4}	0	0	1.76×10^{-5}	7.36×10^{-6}	0.0078	0.0143
SCA: C-PDN-PIDN [14]	0.95	1.10	0.92	0.0023	0.00054	0.000334	8.92×10^{-5}	4.85×10^{-5}	2.30×10^{-5}	2.05×10^{-6}	1.47×10^{-6}	0.0041	0.0115
hDE-PSO: FPID [33]	0.73	1.24	0.53	0.0034	0.00078	0.000313	0	0	0	2.57×10^{-6}	1.05×10^{-6}	0.0032	0.0062
JA: FPIDN [35]	0.45	0.53	0.52	0.0016	0.00042	0.000180	0	0	0	7.44×10^{-7}	2.62×10^{-7}	0.0015	0.0022
ICA: C-I$^{\lambda}$D$^{\mu}$N	**0.11**	**0.35**	**0.36**	**0.0010**	**0.00013**	**0.000130**	0	0	0	$\mathbf{1.07 \times 10^{-7}}$	$\mathbf{3.32 \times 10^{-8}}$	**0.0006**	**0.0015**
ICA: C-I$^{\lambda}$D$^{\mu}$N (RES)	**0.04**	**0.21**	**0.22**	**0.0005**	**0.00007**	**0.000034**	0	0	0	$\mathbf{3.31 \times 10^{-8}}$	$\mathbf{1.38 \times 10^{-8}}$	**0.0004**	**0.0007**

Bold faces values designate the best values.

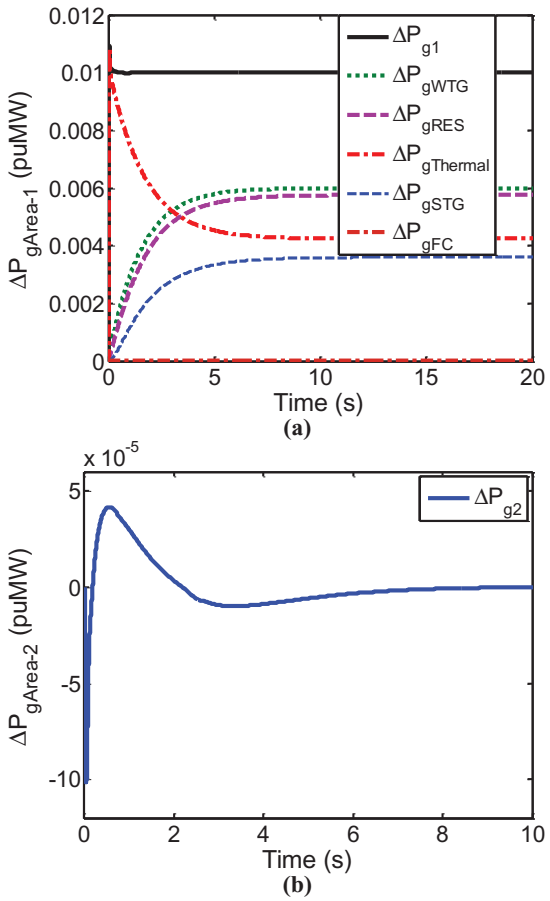


FIGURE 7 2-area reheat thermal system responses (a) $\Delta P_{gArea\text{-}1}$ and (b) $\Delta P_{gArea\text{-}2}$

ISE incorporates the square of the errors over the time, therefore, big errors are penalized more compared to smaller ones. Other performance criteria are utilized to display an effective comparison of the system results. The structure for ISE is given in Equation (17) and other cost functions (J_S) are expressed by Equations (18)–(20).

$$\text{ISE} = \int_0^\infty \left\{\Delta F_1^2 + \Delta F_2^2 + \Delta P\text{tie}_{12}^2\right\} \mathrm{d}t \tag{17}$$

$$\text{ITSE} = \int_0^\infty \mathrm{t}\left\{\Delta F_1^2 + \Delta F_2^2 + \Delta P\text{tie}_{12}^2\right\} \mathrm{d}t \tag{18}$$

$$\text{IAE} = \int_0^\infty \left\{|\Delta F_1| + \Delta F_2| + |\Delta P\text{tie}_{12}|\right\} \mathrm{d}t \tag{19}$$

$$\text{ITAE} = \int_0^\infty t\left\{|\Delta F_1| + |\Delta F_2| + |\Delta P\text{tie}_{12}\right\} \mathrm{d}t \tag{20}$$

The results of the systems are attained with the controller parameters acquired for smallest value of Equation (17) using ICA. J_S in Equations (18)–(20) are also computed for comparison purpose. The parameters to be tuned for the proposed C-I$^\lambda$D$^\mu$N controller are 9 ($K_P/K_I/K_D/N/K_{I1}/\lambda/K_{D1}/\mu/N1$) and

TABLE 3 Comparative statistical analysis

Controllers	Minimum	Average	Maximum	SD
C-I$^\lambda$D$^\mu$N	1.07×10^{-7}	4.87×10^{-7}	9.85×10^{-7}	0.17×10^{-7}
C-I$^\lambda$D$^\mu$N (RES)	3.31×10^{-8}	6.43×10^{-8}	9.98×10^{-8}	0.62×10^{-8}

for outer loop FPIDN are 4 ($K_P/K_I/K_D/N$) parameters. Both controllers are tuned using ICA to get smallest value of (17) in the presence of the following restraints:

$$\left.\begin{aligned} K_P^{\min} \le K_P \le K_P^{\max} \\ K_I^{\min} \le K_I \le K_I^{\max} \\ K_D^{\min} \le K_D \le K_D^{\max} \\ N^{\min} \le N \le N^{\max} \\ K_{I1}^{\min} \le K_{I1} \le K_{I1}^{\max} \\ \lambda^{\min} \le \lambda \le \lambda^{\max} \\ K_{D1}^{\min} \le K_{D1} \le K_{D1}^{\max} \\ \mu^{\min} \le \mu \le \mu^{\max} \\ N1^{\min} < N1 < \mathrm{N1}^{\max} \end{aligned}\right\} \tag{21}$$

The min/max in Equation (21) represents minimum/maximum values of the concerned parameter.

5 | SIMULATION RESULTS

5.1 | Implementation of ICA based control strategy

The models of the systems under examination shown in Figures 5(a) and 8 are designed on MATLAB/SIMULINK platform and ICA programs are prepared in .mfile. The applicable parameters are listed in Appendix. The designed models are simulated in separate programs with/without considering RES for various disturbances in control areas. The objective function is computed in the .m file and utilized in the optimization algorithm. For the execution of ICA, the parameters should be selected carefully as indicated in Section 4. In the present study, no of countries = 500 and number of iteration are taken equal to 100. MATLAB/SIMULINK 7.5.0 (R2007b) is utilized for coding, models and results of the systems. Simulations are performed on an Intel(R), core(TM) i-5-6200U CPU of 2.4 GHz and 8 GB RAM computer. The optimization procedure is reiterated 50 times and the optimum parameters of FPIDN, C-I$^\lambda$D$^\mu$N and C-I$^\lambda$D$^\mu$N (RES) controllers obtained at minimum values of J are provided in Table 1 for 2-area RTS and 2-area MSHTS PS models under study.

5.2 | 2-area reheat thermal system

Primarily, the study is conducted on a more pragmatic 2-area reheat thermal system (RTS) integrated with renewable energy sources (RES) in area-1 (Figure 5(a)). The details of RES are

TABLE 4 T_S, $U_S/O_S/J_S$ for 2-area multi-source hydro-thermal system results shown in Figure 9(a–c)

	T_S (s)			U_S (−ve) (Hz)		U_S (−ve) (puMW)	O_S (Hz)		O_S (puMW)	J_S(puMW)			
Controller type	ΔF_1	ΔF_2	$\Delta Ptie_{12}$	ΔF_1	ΔF_2	$\Delta Ptie_{12}$	ΔF_1	ΔF_2	$\Delta Ptie_{12}$	**ISE**	**ITSE**	**IAE**	**ITAE**
JA-IWO: PID [19]	3.93	4.44	2.25	0.0146	0.0066	0.0027	2.16×10^{-3}	8.00×10^{-4}	2.77×10^{-4}	1.71×10^{-4}	1.26×10^{-4}	0.0311	0.0598
GWO: PID [15]	1.21	3.50	2.64	0.0149	0.0069	0.0026	0	0	0	1.47×10^{-4}	9.03×10^{-5}	0.0233	0.0259
hFA-PS: PID [17]	2.94	4.53	3.30	0.0134	0.0066	0.00220	0	0	0	1.17×10^{-4}	8.50×10^{-5}	0.0255	0.0395
OIO: PID [21]	1.83	3.32	2.06	0.0088	0.0030	0.001000	0	0	0	3.07×10^{-5}	1.80×10^{-5}	0.0132	0.0289
QODSA: PID [24]	1.10	3.36	2.57	0.0154	0.0078	0.002737	2.50×10^{-4}	0	0	1.50×10^{-4}	8.81×10^{-5}	0.0228	0.0255
BFOA: FOPID [38]	1.20	2.15	1.67	0.0115	0.0046	0.0015	4.51×10^{-4}	0	0	5.22×10^{-5}	2.14×10^{-5}	0.0116	0.0121
BFOA: FPI [28]	1.41	2.86	1.92	0.0093	0.0029	0.00108	0	0	0	3.47×10^{-5}	1.76×10^{-5}	0.0122	0.0197
BFOA: FPID [28,38]	0.82	1.79	1.17	0.0065	0.0019	0.00079	2.59×10^{-4}	0	0	1.60×10^{-5}	6.07×10^{-6}	0.0068	0.0083
ICA: FPIDN	0.78	1.59	0.60	0.0047	0.0011	0.00043	0	0	0	6.01×10^{-6}	2.20×10^{-6}	0.0046	0.0082
ICA: C-I$^{\lambda}$D$^{\mu}$N	**0.24**	**0.37**	**0.38**	**0.0027**	**0.0004**	**0.00015**	0	0	0	$\mathbf{9.41 \times 10^{-7}}$	$\mathbf{1.97 \times 10^{-7}}$	**0.0015**	**0.0033**
ICA: C-I$^{\lambda}$D$^{\mu}$N (RES)	**0.18**	**0.21**	**0.27**	**0.0020**	**0.0002**	**0.00009**	0	0	0	$\mathbf{3.68 \times 10^{-7}}$	$\mathbf{5.86 \times 10^{-8}}$	**0.0008**	**0.0013**

Bold faces values designate the best values.

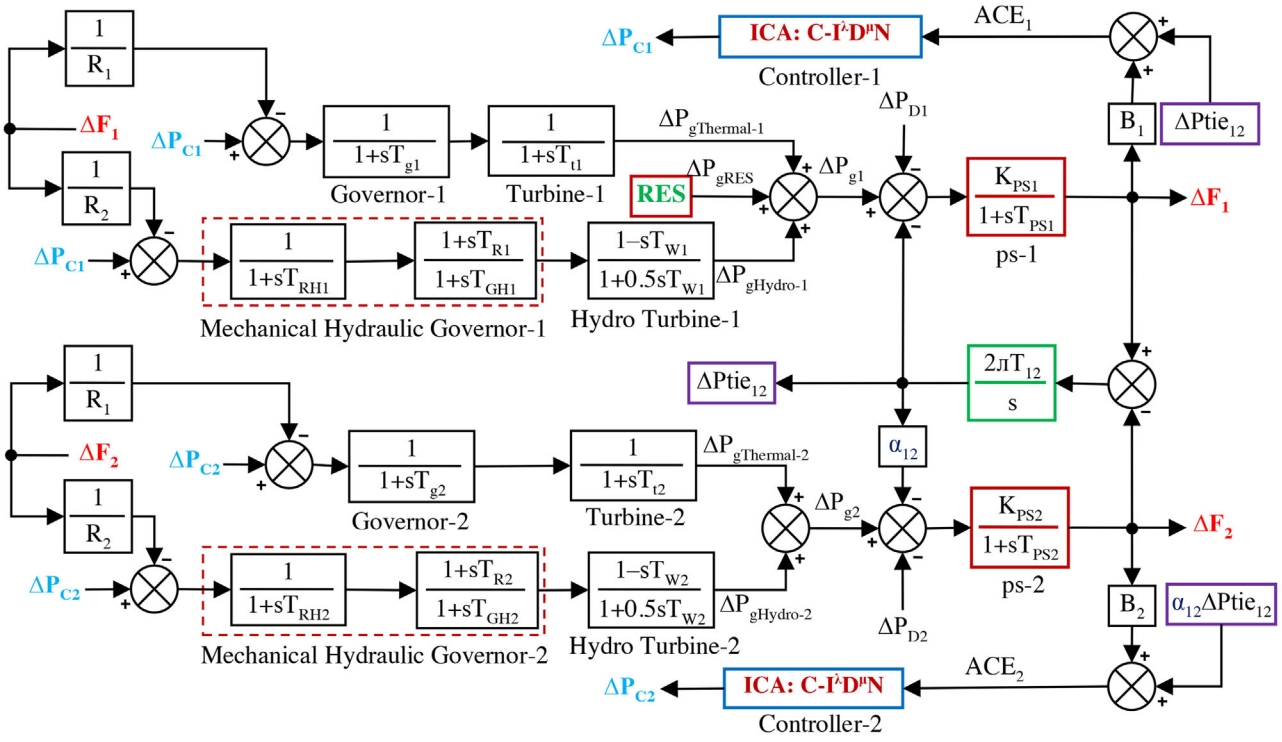


FIGURE 8 2-area multi-source hydro-thermal system model [29]

represented in Figure 5(b). The system dynamic results of change in area-1 frequency (ΔF_1), change in area-2 frequency (ΔF_2) and change in tie power ($\Delta P\text{tie}_{12}$) with the proposed method are obtained by directing a step load perturbation (SLP) of 0.01 puMW in area-1 at time $t = 0$ s. Additionally, a SLP of 0.001 puMW for each STG and 0.002 puMW for each WTG at time $t = 0$ s are considered. The results are exposed in Figure 6(a–c). To reveal the contribution of C-I$^{\lambda}$D$^{\mu}$N controller, its outcomes are matched with various existing and the best claimed intelligent controllers like sine-cosine algorithm (SCA) tuned cascade PDN-PIDN (C-PDN-PIDN) [14], artificial bee colony (ABC) tuned PID [16], bacterial swarm optimization (BSO) tuned fractional order PID (FOPID) [20], SOS algorithm tuned PIDN [22], hybrid differential evolution-PSO (hDE-PSO) tuned FPID [33] and jaya algorithm (JA) tuned FPIDN [35] controllers. The inspection of Figure 6(a–c) divulges that the C-I$^{\lambda}$D$^{\mu}$N controller significantly displays superiority over other controllers in concerning swift responses, no oscillations and negligible settling times (T_S)/peak undershoots (U_S)/overshoots (O_S) in $\Delta F_1/\Delta F_2/\Delta P\text{tie}_{12}$ results.

To inspect the outcome of the C-I$^{\lambda}$D$^{\mu}$N controller further, the numerical values of ISE/ITSE/IAE/ITAE defined via Equations (17)–(20), T_S, U_S and O_S are calculated through simulations and shown in Table 2. Error criteria ISE/ITSE/IAE/ITAE are obtained via simulating the models for 15 s and T_S are obtained for a tolerance band of ±0.0005. Table 2 also demonstrates the advantage of ICA tuned C-I$^{\lambda}$D$^{\mu}$N controller over SCA C-PDN-PIDN [14], ABC PID [16], BSO FOPID [20], SOS PIDN [22], hDE-PSO FPID [33] and JA FPIDN [35] control schemes from least ISE (1.07×10^{-7}), ITSE (3.32×10^{-8}), IAE (0.0006), ITAE (0.0015), T_S ($\Delta F_1 = 0.11$ s, $\Delta F_2 = 0.35$ s, $\Delta P\text{tie}_{12} = 0.36$ s) and U_S ($\Delta F_1 = -0.0010$ Hz, $\Delta F_2 = -0.00013$ Hz, $\Delta P\text{tie}_{12} = -0.000130$ puMW) point of view. Fascinatingly, this observation is not being raised in the corresponding literature that is, in [3] and hence, it establishes a literature gap. It is plainly perceived that C-I$^{\lambda}$D$^{\mu}$N controller is competent in alleviating the oscillations in the system responses. Hence, the disturbance rejection characteristic of C-I$^{\lambda}$D$^{\mu}$N is reasonably appropriate for the AGC compared to other methods.

Further, it is observed that RES comprising of solar thermal generators (STGs), wind turbine generators (WTGs) and aqua electrolyser (AE)-fuel cells (FCs) assist frequency regulation via generating powers in dynamic load condition as indicated in Figure 6(a–c). With C-I$^{\lambda}$D$^{\mu}$N controller in the presence of RES, system performance is observed enriched in terms of further reduction in ISE (3.31×10^{-8}), ITSE (1.38×10^{-8}), IAE (0.0004), ITAE (0.0007), T_S ($\Delta F_1 = 0.04$ s, $\Delta F_2 = 0.21$ s, $\Delta P\text{tie}_{12} = 0.22$ s) and U_S ($\Delta F_1 = -0.0005$ Hz, $\Delta F_2 = -0.00007$ Hz, $\Delta P\text{tie}_{12} = -0.000034$ puMW (Table 2). Next, the generated powers in area-1 ($\Delta P_{\text{gArea-1}}$) and area-2 ($\Delta P_{\text{gArea-2}}$) are shown in Figure 7(a,b). The generated powers by STGs, WTGs, fuel cells (FCs) RES and area-1 thermal are $\Delta P_{\text{gSTG}} = 0.0036$ puMW, $\Delta P_{\text{gWTG}} = 0.006$ puMW, $\Delta P_{\text{gFC}} = 1.536 \times 10^{-7}$puMW, $\Delta P_{\text{gRES}} = 0.00576$ puMW, $\Delta P_{\text{gThermal-1}} = 0.00424$ puMW, respectively. The total power generation in area-1 that is, $\Delta P_{g1} = 0.01$ puMW. Here, ΔP_{g1} ($\Delta P_{\text{gRES}} + \Delta P_{\text{gThermal-1}}$) = ΔP_{D1} that is, generation is equal to

the power demand. However, in area-2, as $\Delta P_{D2} = 0$ puMW, $\Delta P_{gThermal\text{-}2} = \Delta P_{g2} = 0$ puMW.

To check the stability of the proposed ICA based controller, a comparative statistical investigation is performed. The statistical data of minimum, average, maximum and standard deviation (SD) of ISE objective function values achieved in 50 independent runs with the proposed controller (with/without RES) are summarized in Table 3. From the analysis Table 3, it is vibrant that minimum objective function value is attained with the proposed C-I$^{\lambda}$D$^{\mu}$N controller in the presence of RES (ISE = 3.31 $\times 10^{-8}$) compared to C-I$^{\lambda}$D$^{\mu}$N controller (ISE = 1.07 $\times 10^{-7}$). It is also clear from Table 3 that C-I$^{\lambda}$D$^{\mu}$N (RES) outperforms C-I$^{\lambda}$D$^{\mu}$N technique in terms of average, maximum and standard deviation values obtained. It can be stated that the proposed controller is stable and superior specifically with RES for an AGC problem.

5.3 | 2-area multi-source hydro-thermal system

To examine the competence of the C-I$^{\lambda}$D$^{\mu}$N controller to deal with different AGC systems, the study is also widened to a 2-area multi-source hydro-thermal system (MSHTS) whose TF model is displayed in Figure 8. In an area, two units that is, one non-reheat thermal and one mechanical governor outfitted hydro are available. Additionally, RES are integrated in area-1 [3]. The symbols used in the figure are explained in Nomenclature and their values are specified in Appendix. The simulation outcomes of the system at SLP = 0.015 puMW due to the C-I$^{\lambda}$D$^{\mu}$N/fuzzy-PIDN (FPIDN) controllers developed in the study and various other prevalent controllers like grey wolf optimization (GWO) [15]/hFA-PS [17]/JA-IWO [19]/optics inspired optimization (OIO) [21]/quasi oppositional differential search algorithm (QODSA) [24] optimized PID and BFOA optimized FPI/FPID/FOPID [28, 38] are displayed in Figure 9(a–c) and Table 4. It is obvious from Figure 9(a–c) and Table 4 that the offered controller parades improved dynamic outcome in contrast to the published controllers and ICA tuned FPIDN controller. It is perceived that lowest values of ISE (9.41 $\times 10^{-7}$), ITSE (1.97 $\times 10^{-7}$), IAE (0.0015), ITAE (0.0033), T_S (ΔF_1 = 0.24 s, ΔF_2 = 0.37 s, $\Delta P\text{tie}_{12}$ = 0.38 s) and U_S (ΔF_1 = −0.0027 Hz, ΔF_2 = −0.0004 Hz, $\Delta P\text{tie}_{12}$ = −0.00015 puMW) are obtained with C-I$^{\lambda}$D$^{\mu}$N controller compared to ICA based FPIDN controller which provides somewhat greater values of ISE (= 6.01 $\times 10^{-6}$) ITSE (= 2.20 $\times 10^{-6}$), IAE (= 0.0046), ITAE (= 0.0082), T_S (ΔF_1 = 0.78 s, ΔF_2 = 1.59 s, $\Delta P\text{tie}_{12}$ = 0.60 s) and U_S (ΔF_1 = −0.0047 Hz, ΔF_2 = −0.0011 Hz, $\Delta P\text{tie}_{12}$ = −0.00043 puMW). It should be noted that ICA based FPIDN controller offers better results than JA-IWO [19]/QODSA [24]/GWO [15]/hFA-PS [17]/OIO [21] based PID and BFOA based FPI/FPID/FOPID [28, 38] controllers. However, performance of the recommended C-I$^{\lambda}$D$^{\mu}$N controller is additionally advanced in the company of RES as clearly indicated in Figure 9(a–c) and Table 4.

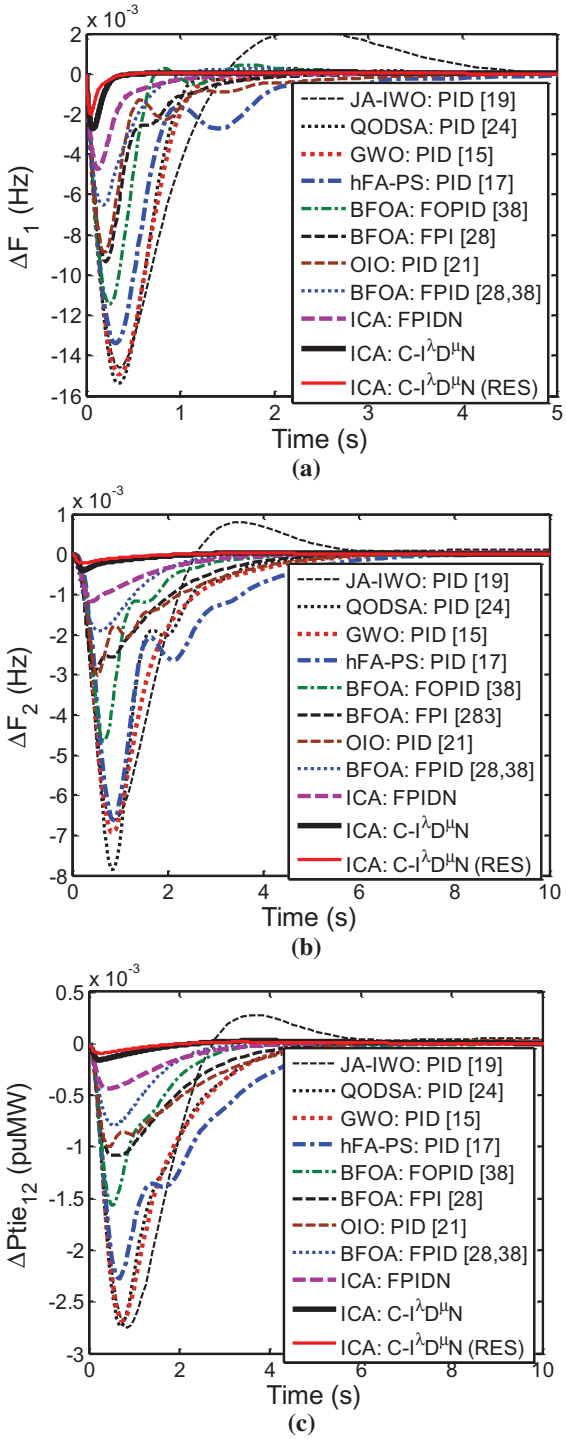


FIGURE 9 2-area multi-source hydro-thermal system responses: (a) ΔF_1, (b) ΔF_2 and (c) $\Delta P\text{tie}_{12}$

The generation results in area-1 ($\Delta P_{gArea\text{-}1}$) and area-2 ($\Delta P_{gArea\text{-}2}$) are exposed in Figure 10(a,b). The generated powers by STGs, WTGs, FCs, RES, area-1 thermal and area-1 hydro are ΔP_{gSTG} = 0.0036 puMW, ΔP_{gWTG} = 0.006 puMW, ΔP_{gFC} = 1.536 $\times 10^{-7}$ puMW, ΔP_{gRES} = 0.00576 puMW, $\Delta P_{gThermal\text{-}1}$ = 0.004622 puMW, $\Delta P_{gHydro\text{-}1}$ = 0.004618 puMW,

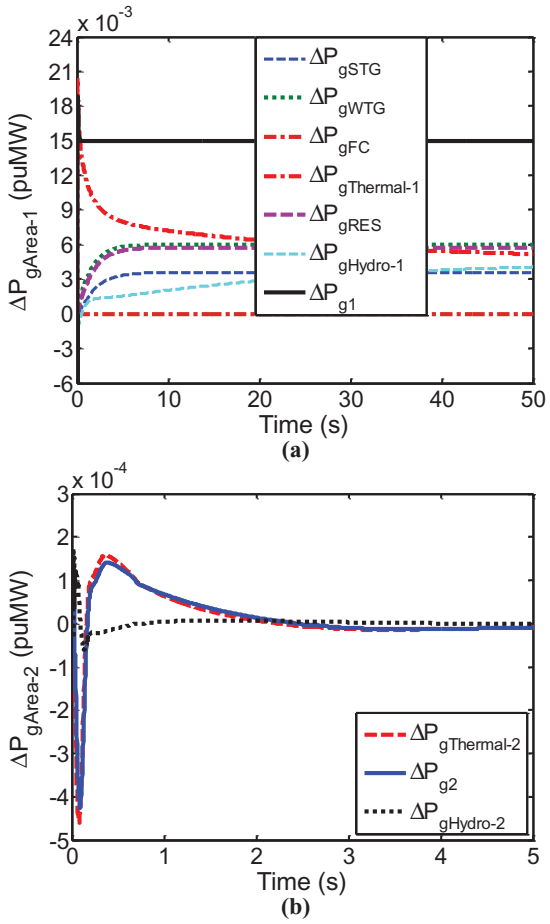


FIGURE 10 2-area multi-source hydro-thermal system responses: (a) $\Delta P_{\text{gArea-1}}$ and (b) $\Delta P_{\text{gArea-2}}$

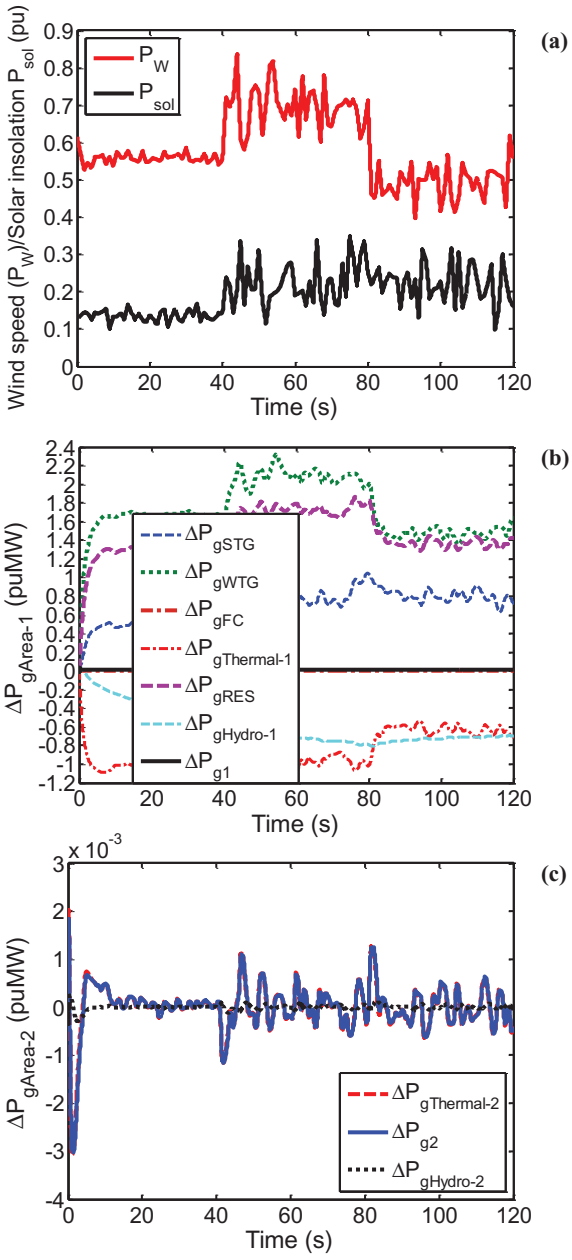


FIGURE 11 2-area multi-source hydro-thermal system responses under abrupt $P_{\text{sol}}/P_{\text{W}}$: (a) $P_{\text{sol}}/P_{\text{W}}$, (b) $\Delta P_{\text{gArea-1}}$ and (c) $\Delta P_{\text{gArea-2}}$

respectively. The total power generation in area-1 that is, ΔP_{g1} $(= \Delta P_{\text{gRES}} + \Delta P_{\text{gThermal-1}} + \Delta P_{\text{gHydro-1}}) = 0.015$ puMW, which is equal to the demand in area-1 that is, ΔP_{D1}. However, in area-2, as $\Delta P_{\text{D2}} = 0$ puMW, $\Delta P_{\text{gThermal-2}} = 0$ puMW, $\Delta P_{\text{gHydro-2}} = 0$ puMW and $\Delta P_{\text{g2}} = (\Delta P_{\text{gThermal-2}} + \Delta P_{\text{gHydro-2}}) = 0$ puMW.

Next, in practical scenario, the power outputs of STG and WTG renewable sources are highly unpredictable and variable due to their dependence on environmental factors like solar irradiation and wind speed. Consequently, the deviations in their power outputs are inspected in the existence of random powers (P_{sol}) and (P_{W}) caused due to solar insolation and wind speed respectively varying abruptly with time as shown in Figure 11(a). However, SLP in area-1 remains the same. With the proposed controller, the simulation results of generations in area-1 that is ΔP_{gSTG}, ΔP_{gWTG}, ΔP_{gFC}, ΔP_{gRES}, $\Delta P_{\text{gThermal-1}}$, $\Delta P_{\text{gHydro-1}}$ and ΔP_{g1} are shown in Figure 11(b). It is observed from Figure 11(b) that the generations are changing randomly in contrast to Figure 10(a). However, in steady state, ΔP_{g1} fluctuates around the required 0.015 puMW ($= \Delta P_{\text{D1}}$) power. The simulation results of generations in area-2 shown in Figure 11(c) fluctuates around zero as $\Delta P_{\text{D2}} = 0$ puMW, which satisfies the AGC requirement effectively. It is observed that the proposed controller works efficiently in this realistic circumstance. However, J_{S} values increases from the values shown in Table 4 (ISE $= 3.68 \times 10^{-7}$, ITSE $= 5.86 \times 10^{-8}$, IAE $= 0.0008$ and ITAE $= 0.0013$) due to longer simulation time and abrupt/higher disturbances taking place in the form of $P_{\text{sol}}/P_{\text{W}}$ power as: ISE $= 3.93 \times 10^{-4}$, ITSE $= 1.27 \times 10^{-2}$, IAE $= 0.2072$ and ITAE $= 11.32$.

6 | SENSITIVITY STUDY

A sensitivity study is performed to check the robustness of the recommended C-I$^{\lambda}$D$^{\mu}$N controller for extensive deviations in the system operating conditions. To illustrate the robustness,

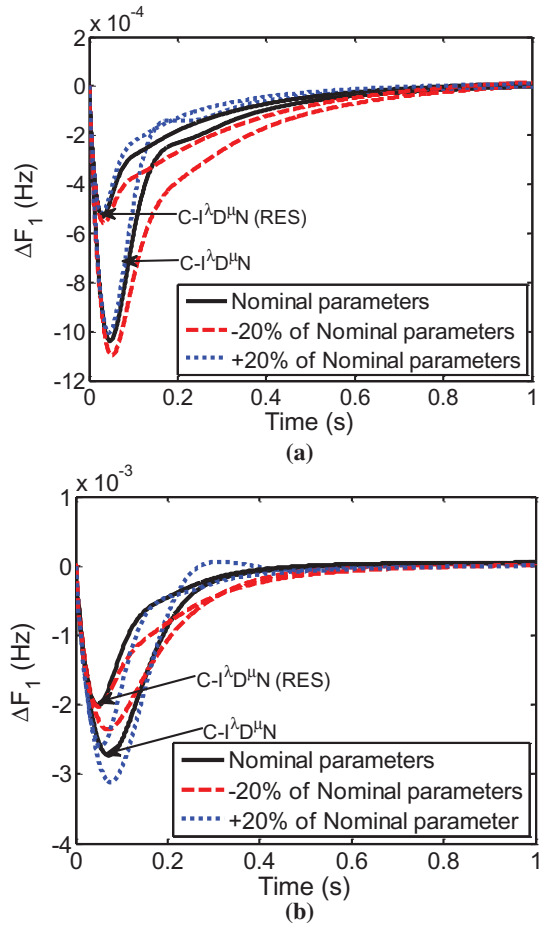


FIGURE 12 Sensitivity study responses: (a) ΔF_1 of reheat thermal system and (b) ΔF_1 of multi-source hydro-thermal system

system's nominal parameters stated in Appendix are altered for ±20%. The C-I$^{\lambda}$D$^{\mu}$N controller should behave healthy devoid of considerable deterioration in the system's results under parameter deviations. Figure 12(a,b) demonstrates ΔF_1 results under nominal and ±20% variations in parameters for 2-area reheat thermal system and multi-source hydro-thermal system, in order with/without RES. Here, ΔF_2 and $\Delta P\text{tie}_{12}$ responses are not provided to save the space. The inspection of Figure 12(a,b) evidently reveals that all results with/without RES are stable and they do not deviate much from the nominal responses. However, controller outperforms in the continuation of RES. The several performance indices that is, T_S and J_S for nominal parameters and ±20% of nominal parameter variation cases for reheat thermal system (with/without RES) and multi-source hydro-thermal system (with/without RES) are given in Table 5 for ΔF_1, ΔF_2 and $\Delta P\text{tie}_{12}$ responses. It can be perceived from Table 5 that settling time, ISE, ITSE, IAE and ITAE values diverge within tolerable ranges and are more or less equal to the corresponding values attained with nominal system parameters. Therefore, it can be established that the proposed controller is robust and executes suitably when system nominal parameters deviate in the range ±20%.

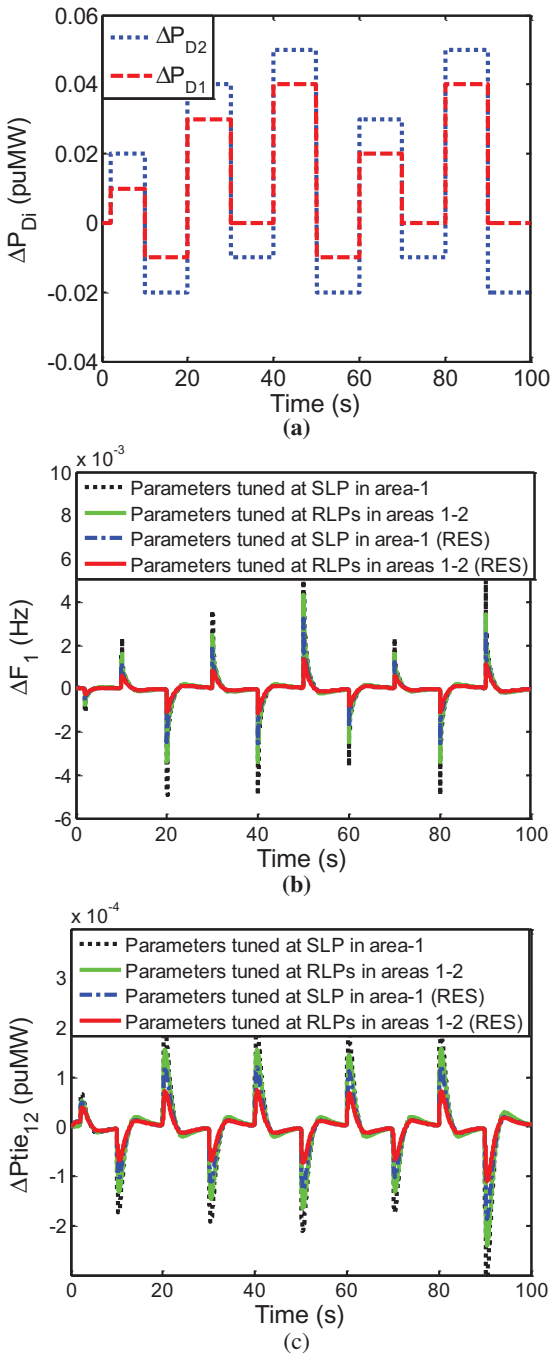


FIGURE 13 Responses of reheat thermal system during sensitivity study (a) RLPs in areas 1–2, (b) ΔF_1 and (c) $\Delta P\text{tie}_{12}$

Additionally, the robustness study of the C-I$^{\lambda}$D$^{\mu}$N controller is performed out under the existence of random load perturbations (RLPs) of high magnitudes in both control areas of 2-area reheat thermal system. Dissimilar RLPs in all control areas represent a much more realistic power system environment. At higher SLP/RLP scenario, the power system will become unbalanced if the designed controller is non-robust. Figure 13(a) shows RLPs of different magnitudes applied in areas 1 and 2 at $t = 2$ s of reheat thermal system. Figure 13(b,c) show ΔF_1 and $\Delta P\text{tie}_{12}$ responses with parameters of the C-I$^{\lambda}$D$^{\mu}$N

TABLE 5 Sensitivity study data for T_S and J_S

Parameter variations	ΔF_1 (Hz)	ΔF_2 (Hz)	$\Delta Ptie_{12}$ (puMW)	ISE	ITSE	IAE	ITAE
Reheat thermal system (C-I$^{\lambda}$D$^{\mu}$N) (Figure 12(a))							
Nominal parameters	0.11	0.35	0.36	1.07×10^{-7}	3.32×10^{-8}	0.0006	0.0015
−20% of Nominal parameters	0.12	0.36	0.37	1.66×10^{-7}	5.97×10^{-8}	0.0008	0.0018
+20% of Nominal parameters	0.10	0.34	0.35	0.11×10^{-7}	2.64×10^{-8}	0.0005	0.0014
Reheat thermal system (C-I$^{\lambda}$D$^{\mu}$N (RES) (Figure 12(a))							
Nominal parameters	0.04	0.21	0.22	3.31×10^{-8}	1.38×10^{-8}	0.0004	0.0007
−20% of Nominal parameters	0.05	0.22	0.23	5.53×10^{-8}	3.08×10^{-8}	0.0005	0.0010
+20% of Nominal parameters	0.05	0.20	0.21	2.33×10^{-8}	7.16×10^{-9}	0.0003	0.0005
Multi-source hydro-thermal system (C-I$^{\lambda}$D$^{\mu}$N) (Figure 12(b))							
Nominal parameters	0.24	0.37	0.38	9.41×10^{-7}	1.97×10^{-7}	0.0015	0.0033
−20% of Nominal parameters	0.27	0.39	0.40	8.42×10^{-7}	2.45×10^{-7}	0.0017	0.0038
+20% of Nominal parameters	0.21	0.33	0.34	1.13×10^{-6}	1.90×10^{-7}	0.0014	0.0028
Multi-source hydro-thermal system (C-I$^{\lambda}$D$^{\mu}$N (RES) (Figure 12(b))							
Nominal parameters	0.18	0.21	0.27	3.68×10^{-7}	5.86×10^{-8}	0.0008	0.0013
−20% of Nominal parameters	0.20	0.23	0.28	5.30×10^{-7}	1.10×10^{-7}	0.0011	0.0015
+20% of Nominal parameters	0.18	0.20	0.26	6.14×10^{-7}	9.03×10^{-8}	0.0009	0.0011

controller optimized at area-1 SLP of 0.01 puMW (given in Table 1) and RLPs in areas 1–2. The optimized controller parameters obtained at this SLP produce ISE = 9.47×10^{-5}/ITSE = 5.10×10^{-3}/IAE = 0.0551/ITAE = 2.8890 and ISE = 3.26×10^{-5}/ITSE = 1.76×10^{-3}/IAE = 0.0372/ITAE = 1.9520 (with RES), which are greater than their corresponding values given in Table 2. This means that the controller with RES produces better performance regarding J_S and results shown in Figure 13(b,c). Next, the optimized controller parameters obtained at RLPs are K_P = 4.8589, K_I = 4.6874, K_D = 0.1499, N = 460.4499, K_{I1} = 4.5624, λ = 0.3401, K_{D1} = 0.8031, μ = 0.4089 and $N1$ = 489.0021 and K_P = 4.5682, K_I = 4.9984, K_D = 0.4681, N = 447.8967, K_{I1} = 4.9564, λ = 0.2499, K_{D1} = 0.9785, μ = 0.4934 and $N1$ = 467.7896 (with RES). The J_S results are ISE = 5.57×10^{-5}/ITSE = 3.01×10^{-3}/IAE = 0.0501/ITAE = 2.6360 and ISE = 1.46×10^{-5}/ITSE = 7.87×10^{-4}/IAE = 0.0295/ITAE = 1.5470 (RES). Serious inspection of Figure 13(c) evidently discloses that both the responses are stable. However, system responses and J_S obtained at parameters optimized with RLPs are superior. Thus, the stable performance of the system at dissimilar operating conditions authorizes the robustness of ICA optimized C-I$^{\lambda}$D$^{\mu}$N controller.

7 | CONCLUSION

A novel ICA tuned C-I$^{\lambda}$D$^{\mu}$N controller is suggested for AGC of 2-area electric PSs penetrated with RES. Initially, a 2-area RTS is scrutinized in the presence of 1% SLP in control area-1. The simulation results due to the proposed controller are compared with the existing ABC tuned PID, BSO tuned FOPID, SOS tuned PIDN, hDE-PSO tuned FPID and JA tuned FPIDN controllers. It is noted that significant development regarding oscillation free results with smallest values of $T_S/U_S/O_S/J_S$ are obtained due to the suggested ICA tuned C-I$^{\lambda}$D$^{\mu}$N controller over other controllers. However, further enhanced results are acquired in the presence of RES integrated in area-1. To further validate the capability, extensibility and advantage of the method for other systems, the work is also protracted to a 2-area MSHTS. The asset of C-I$^{\lambda}$D$^{\mu}$N controller is established over ICA based FPIDN and intelligent JA- IWO/QODSA/GWO/hFA-PS/OIO tuned PID and BFOA tuned FPI/FPID/FOPID controllers prevailing in the literature. The worth of RES addition is validated via enhanced system outcome with the C-I$^{\lambda}$D$^{\mu}$N controller. The sensitivity analysis validates the robustness of the C-I$^{\lambda}$D$^{\mu}$N controller with/without RES under ±20% deviations in the system's parameters and random load perturbations (RLPs) created in both areas of reheat thermal system. It is observed that the proposed controller designed at nominal condition is robust and shows stable performance at extensive changes in the parameters of the systems and RLPs. Additionally, RES integration is suggested in PSs to get reliable, better quality, and environmentally-friendly electrical power.

Hence, after analysis and discussion done above, it is assured that the suggested ICA optimized C-I$^{\lambda}$D$^{\mu}$N controller may be an appropriate controller for solving more intricate engineering problems in isolated/multi-area interconnected traditional as well as modern restructured AGC systems with/without RES integration.

Nomenclature

B	frequency bias constant, puMW/Hz
K_{AE}	aqua electrolyser gain
K_{FC}	fuel cell gain
K_{PS}	power system gain
K_r	reheater gain
K_S, K_T	solar thermal generator gains
K_{WTG}	wind turbine generator gain
R	speed governor regulation parameter, Hz/puMW
T_{12}	synchronizing coefficient of tie-line, puMW/rad
T_{AE}	aqua electrolyser time constant (TC), s
T_{FC}	fuel cell time constant, s
T_g	thermal governor TC, s
T_{GH}	hydro governor main servo TC, s
T_{PS}	power system time constant (TC), s
T_r	reheater TC, s
T_R	hydro governor reset time, s
T_{RH}	time constant of hydro governor transient droop, s
T_S, T_T	solar thermal generator TCs, s
T_t	thermal turbine TC, s
T_W	water starting time, s
T_{WTG}	wind turbine generator TC, s
α_{12}	area rating ratio

ORCID

Yogendra Arya https://orcid.org/0000-0003-4661-1950
Nishant Kumar https://orcid.org/0000-0003-4084-7436
Pankaj Dahiya https://orcid.org/0000-0002-2722-1554
Gulshan Sharma https://orcid.org/0000-0002-4726-0956
Emre Çelik https://orcid.org/0000-0002-2961-0035
Sandeep Dhundhara https://orcid.org/0000-0003-4403-4897
Mandeep Sharma https://orcid.org/0000-0002-7217-1410

REFERENCES

1. Nanda, J., et al.: A new technique in hydro thermal interconnected automatic generation control system by using minority charge carrier inspired algorithm. Int. J. Elect. Power Energy Syst. 68, 259–268 (2015)
2. Rahman, A., et al.: AGC of dish-Stirling solar thermal integrated thermal system with biogeography based optimised three degree of freedom PID controller. IET Renew. Power Gener. 10(8), 1161–1170 (2016)
3. Kler, D., et al.: Optimal integral minus proportional derivative controller design by evolutionary algorithm for thermal-renewable energy-hybrid power systems. IET Renewab. Power Gener. 13(11), 2000–2012 (2019)
4. Das, D.C., Roy, A.K., Sinha, N.: GA based frequency controller for solar thermal-diesel-wind hybrid energy generation/energy storage system. Int. J. Elect. Power Energy Syst. 43(1), 262–279 (2012)
5. Arya Y.: Impact of hydrogen aqua electrolyzer-fuel cell units on automatic generation control of power systems with a new optimal fuzzy TIDF-II controller. Renewab. Energy 139, 468–482 (2019)
6. Tasnin, W., et al.: Deregulated AGC of multi-area system incorporating dish-stirling solar thermal and geothermal power plants using fractional order cascade controller. Int. J. Elect. Power Energy Syst. 101, 60–74 (2018)
7. Tasnin, W., Saikia, L.C.: Maiden application of an sine-cosine algorithm optimised FO cascade controller in automatic generation control of multi-area thermal system incorporating dish-stirling solar and geothermal power plants. IET Renewab. Power Gener. 12(5), 585–597 (2018)
8. Xi, L., et al.: Wolf pack hunting strategy for automatic generation control of an islanding smart distribution network. Energy Convers. Manage. 122, 10–24 (2016)
9. Xi, L., et al.: A novel multi-agent DDQN-AD method-based distributed strategy for automatic generation control of integrated energy systems. IEEE Trans. Sustainab. Energy 11 (2020) https://doi.org/10.1109/TSTE.2019.2958361.
10. Xi, L. et al.: A deep reinforcement learning algorithm for the order optimization allocation of total power in the interconnected power grids. CSEE J. Power Energy Syst., 2020, https://doi.org/10.17775/CSEEJPES.2019.01840.
11. Xi, L., et al.: A virtual generation ecosystem control strategy for automatic generation control of interconnected microgrids. IEEE Access 8, 94165–94175 (2020)
12. Xi, L., et al.: Smart generation control based on multi-agent reinforcement learning with the idea of the time tunnel. Energy 153, 977–987 (2018)
13. Xi, L., et al.: Automatic generation control based on multiple neural networks with actor-critic strategy. IEEE Trans. Neural Net. Learning Syst. (2020), https://doi.org/10.1109/TNNLS.2020.3006080.
14. Satapathy, P., et al.: Design of PD-PID controller with double derivative filter for frequency regulation. In 2nd IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems, Delhi, India, 2018, pp. 1142–1147 (2018)
15. Guha, D., et al.: Load frequency control of interconnected power system using grey wolf optimization. Swarm Evol. Comput. 27, 97–115 (2016)
16. Gozde, H., et al.: Comparative performance analysis of artificial bee colony algorithm in automatic generation control for interconnected reheat thermal power system. Int. J. Elect. Power Energy Syst. 42(1), 167–178 (2012)
17. Sahu, R.K., et al.: A hybrid firefly algorithm and pattern search technique for automatic generation control of multi area power systems. Int. J. Elect. Power Energy Syst. 64, 9–23 (2015)
18. Shabani, H., et al.: A robust PID controller based on imperialist competitive algorithm for load-frequency control of power systems. ISA Trans. 52(1), 88–95 (2013)
19. Oozeer, M.Y., Ramjug-Ballgobin, R.: A jaya-based invasive weed optimization technique for load frequency control. In: Fleming P, Lacquet B, Sanei S, Deb K, Jakobsson A. (eds.) Smart and Sustainable Engineering for Next Generation Applications. ELECOM 2018, Lecture Notes Elect Eng, vol. 561, pp. 11–21 Springer, Singapore (2019)
20. Ozdemir, M.T., et al: Tuning of optimal classical and fractional order PID parameters for automatic generation control based on the bacterial swarm optimization. IFAC-PapersOnLine 48(30), 501–506 (2015)
21. Ozdemir, M.T., Ozturk, D.: Comparative performance analysis of optimal PID parameters tuning based on the optics inspired optimization methods for automatic generation control. Energies 10(12), 2134.
22. Hasanien, H.M., El-Fergany, A.: Symbiotic organisms search algorithm for automatic generation control of interconnected power systems including wind farms. IET Gener. Transm. Distrib. 11(7), 1692–1700 (2017)
23. Guha, D., et al.: Symbiotic organism search algorithm applied to load frequency control of multi-area power system. Energy Syst. 9(2), 439–468 (2018)
24. Guha, D., et al.: Quasi-oppositional differential search algorithm applied to load frequency control. Eng. Sc. Tech. an Int. J. 19(4), 1635–1654 (2016)
25. Manoharan, N., et al.: Automatic generation control by hybrid invasive weed optimization and pattern search tuned 2-DOF PID controller. Int. J. Comput. Communicat. Contr. 12(4), 533–549 (2017)
26. Saha, A., Saikia, L.C.: Utilisation of ultra-capacitor in load frequency control under restructured STPP-thermal power systems using WOA optimised PIDN-FOPD controller. IET Gener. Transm. Distrib. 11(13), 3318–3331 (2017)
27. Sahu, R.K., et al.: A novel hybrid PSO-PS optimized fuzzy PI controller for AGC in multi area interconnected power systems. Int. J. Elect. Power Energy Syst. 64, 880–893 (2015)
28. Arya, Y., Kumar, N.: Design and analysis of BFOA-optimized fuzzy PI/PID controller for AGC of multi-area traditional/restructured electrical power systems. Soft Comput. 21(21), 6435–6452 (2017)
29. Arya, Y.: Automatic generation control of two-area electrical power systems via optimal fuzzy classical controller. J. Franklin Inst. 355(5), 2662–2688 (2018)

30. Barisal, A.K., et al.: A hybrid PSO-levy flight algorithm based fuzzy PID controller for automatic generation control of multi area power systems: fuzzy based hybrid PSO for automatic generation control. Int. J. Energy Optimizat. Eng. 6(2), 42–63 (2017)
31. Gheisarnejad, M.: An effective hybrid harmony search and cuckoo optimization algorithm based fuzzy PID controller for load frequency control. Appl. Soft. Comput. 65, 121–138 (2018)
32. Rajesh, K.S., et al.: Hybrid improved firefly-pattern search optimized fuzzy aided PID controller for automatic generation control of power systems with multi-type generations. Swarm Evol. Comput. 44, 200–211 (2019)
33. Sahu, B.K., et al.: Hybrid differential evolution particle swarm optimisation optimised fuzzy proportional-integral derivative controller for automatic generation control of interconnected power system. IET Gener. Transm. Distrib. 8(11), 1789–1800 (2014)
34. Mohanty, P.K., et al: Design and analysis of fuzzy PID controller with derivative filter for AGC in multi-area interconnected power system. IET Gener. Transm. Distrib. 10(15), 3764–3776 (2016)
35. Debnath, M.K., et al.: Application of fuzzy-PIDF controller for automatic generation control using jaya algorithm. Int. J. Pure Appl. Math. 114(9), 51–61 (2017)
36. Arya, Y.: AGC performance enrichment of multi-source hydrothermal gas power systems using new optimized FOFPID controller and redox flow batteries. Energy 127, 704–715 (2017)
37. Arya, Y.: AGC of restructured multi-area multi-source hydrothermal power systems incorporating energy storage units via optimal fractional-order fuzzy PID controller. Neural Comput. Appl. 31(3), 851–872 (2019)
38. Arya, Y., Kumar N.: BFOA-scaled fractional order fuzzy PID controller applied to AGC of multi-area multi-source electric power generating systems. Swarm Evol. Comput. 32, 202–218 (2017)
39. Mahto, T., Mukherjee, V.: Fractional order fuzzy PID controller for wind energy-based hybrid power system using quasi-oppositional harmony search algorithm. IET Gener. Transm. Distrib. 11(13), 3299–3309 (2017)
40. Arya, Y.: A new optimized fuzzy FOPI-FOPD controller for automatic generation control of electric power systems. J. Franklin Inst. 356(11), 5611–5629 (2019)
41. Arya, Y.: A novel CFFOPI-FOPID controller for AGC performance enhancement of single and multi-area electric power systems. ISA Trans. 100, 126–135 (2019), https://doi.org/10.1016/j.isatra.2019.11.025.
42. Atashpaz-Gargari, E., Lucas, C.: Imperialist competitive algorithm: an algorithm for optimization inspires by imperialistic competition. In: IEEE Congress on Evolutionary Computation, Singapore, 2007, pp. 4661–4667 (2007).

How to cite this article: Arya Y, Kumar N, Dahiya P, et al. A new optimal robust cascade-IDN controller design for AGC of thermal/hydro-thermal power systems integrated with renewable energy sources. *IET Renewable Power Generation*. 2020;1–17. https://doi.org/10.1049/rpg2.12061

APPENDIX: SYSTEM DATA

2-area reheat thermal system (RTS) [16]:

$P_{ri} = 2000$ MW, $T_{PSi} = 20$ s, $K_{PSi} = 120$ Hz/puMW, $T_{gi} = 0.08$ s, $T_{ti} = 0.3$ s, $K_{ri} = 0.5$, $T_{ri} = 10$ s, $2\pi T_{12} = 0.545$ puMW/Hz, $B_i = 0.425$ puMW/Hz, $R_i = 2.4$ Hz/puMW, $\alpha_{12} = -1$, $F^{0} = 60$ Hz, initial loading = 50%.

2-area multi-source hydro-thermal system (MSHTS) [17]:

$P_{ri} = 2000$ MW, $T_{PSi} = 20$ s, $K_{PSi} = 100$ Hz/puMW, $T_{ti} = 0.3$ s, $T_{gi} = 0.08$ s, $R_2 = 2.4$ Hz/puMW, $T_{12} = 0.0707$ puMW/rad, $B_i = 0.425$ puMW Hz < sp > -1 < ,/sp > $T_{Ri} = 5$ s, $T_{RHi} = 48.7$ s, $T_{Wi} = 1$ s, $T_{GHi} = 0.513$ s, $R_1 = 2$ Hz/puMW, $\alpha_{12} = -1$, $F^{0} = 60$ Hz, initial loading = 50%.

Renewable energy sources (RES) [3]:

$K_S = 1.8$, $T_S = 1.8$ s, $K_T = 1$, $T_T = 0.3$ s, $K_{WTG} = 1$, $T_{WTG} = 1.5$ s, $K_N = 0.6$, $K_{AE} = 0.002$, $T_{AE} = 0.5$ s, $K_{FC} = 0.01$, $T_{FC} = 4$ s.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

COMPAC: Compressed Time-Domain, Pooling-Aware Convolution CNN Engine With Reduced Data Movement for Energy-Efficient AI Computing

Aseem Sayal, *Member, IEEE*, Shirin Fathima, SS Teja Nibhanupudi, *Student Member, IEEE*, and Jaydeep P. Kulkarni, *Senior Member, IEEE*

***Abstract*— In this work, we demonstrate a compressed time-domain, pooling-aware convolution (COMPAC) convolutional neural network (CNN) engine for energy-efficient edge AI computing by performing multi-bit input and multi-bit weight multiply-and-accumulate (MAC) operations in the time domain. The multi-bit inputs are compactly represented as a single pulsewidth encoded input. This translates into reduced switching capacitance (C_{DYN}), compared with the baseline digital implementation, and can enable low-power neural network computing in an edge device. COMPAC CNN engine employs a novel and an improved version of the memory delay line (MDL) supporting the time residue scaling to perform the signed accumulation of multi-bit input and multi-bit weight products in the time domain. The compressed time-domain (CTD) approach is proposed to improve the throughput in time encoding of the input activations. The simulation results of the proposed CTD approach on the AlexNet CNN over 1000 ImageNet images show that 14.71 and 7.15 input clock cycles are consumed to time-encode an 8-bit input activation in two different CTD modes, improving the throughput by 88.60% and 94.46%, respectively, compared with the conventional pulsewidth modulation-based time-domain encoding. Furthermore, a pooling-aware convolution (PAC) technique is proposed to reduce the number of redundant MAC computations for the convolution layers that are followed by the max-pooling layer. The simulation results on the AlexNet CNN over 1000 ImageNet images show up to 31.47% (21.79%) reduction in the number of non-zero input activations MACs with a top-five classification accuracy loss of 0.60% (0.90%) with an on-chip access overhead of 60.53% (8.03%) for the PAC modes 1 (2) respectively. Finally, energy-efficient data flow for optimal on-/off-chip memory accesses for the time-domain MAC computation is proposed. COMPAC data flow the results in 86.97% reduced on-chip accesses and 29.74% reduced off-chip accesses compared with the Eyeriss approach, at iso-bit precision. COMPAC CNN engine implemented in 65-nm CMOS test chip demonstrates an energy efficiency of 1.044 TOPS/W and the throughput of 0.1278 GOPS at 720 mV for the AlexNet. The top-five classification accuracy of 76.90% measured over 1000 ImageNet images and 77.15% by simulating over 50 000 ImageNet images is achieved. The simulation results comprehending MDL circuit non-idealities for the AlexNet over 50 000 ImageNet validation set images show a classification accuracy loss within 1% compared with the 8-bit fixed-point software implementation.**

***Index Terms*— AlexNet, compressed time-domain (CTD), convolutional neural network (CNN), energy-efficient edge computing, ImageNet, memory delay line (MDL), multiply-and-accumulate (MAC), pooling-aware convolution (PAC), time residue scaling (TRS), time-domain processing.**

I. Introduction

MACHINE learning (ML) approaches are being explored for a wide variety of applications requiring complex computations [1]–[3]. In particular, convolutional neural networks (CNNs, class of artificial neural networks) are extensively used for many such ML applications due to their state-of-the-art classification accuracy at a much lesser complexity compared with their fully connected network (FCN) counterpart [4], [5]. The CNN inference process requires intensive compute and memory resources to compute an enormous number of multiply-and-accumulate (MAC) operations, which consumes a significant fraction of a CNN accelerator's total power budget [6], [7]. The computation in the cloud suffers from issues, such as high latency, limited bandwidth, security, and privacy when transferring data from the edge to the cloud. There arises a critical need to develop energy-efficient computing capabilities to perform MAC operations locally on an edge device. In this article, a compressed time-domain, pooling-aware convolution (COMPAC) CNN engine is proposed for energy-efficient MAC operation by leveraging the concept of a memory delay line (MDL) (presented in our previous work [8], [9]). The key contributions of this work are as follows.

1) Time residue scaling (TRS) technique is proposed in the MDL to perform a multi-bit input and multi-bit weight MAC operation in the time-domain accurately.
2) The compressed time-domain (CTD) approach is proposed, in which the total time taken to encode an input is proportional to its magnitude only rather than the full-scale magnitude. This is achieved by eliminating the dead time in the pulsewidth modulated (PWM) input

Manuscript received June 23, 2020; revised October 19, 2020; accepted November 24, 2020. This work was supported by the Cockrell School of Engineering and by the NASCENT Center's Nanoengineering Online Education Program at the University of Texas at Austin. This article was approved by Guest Editor Jonathan Chang. *(Corresponding author: Aseem Sayal.)*

The authors are with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA (e-mail: aseem.sayal@utexas.edu; shirin15@utexas.edu; subrahmanya_teja@utexas.edu; jaydeep@austin.utexas.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSSC.2020.3041502.

Digital Object Identifier 10.1109/JSSC.2020.3041502

0018-9200 © 2020 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 January 18,2021 at 10:12:23 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

signal when it is low. The proposed CTD approach combines the throughput benefits obtained by skipping the zero-magnitude input activations and eliminating the dead time for the input activations, which have a magnitude less than the full-scale value ($2^n - 1$ for n-bit input). The simulation results on the AlexNet CNN over 1000 ImageNet images show a throughput improvement of 88.60% and 94.46% to time-encode an 8-bit input activation in two different CTD modes compared with the conventional PWM-based time-domain approach [8], [9].

3) Pooling-aware convolution (PAC) technique is proposed to reduce the number of redundant MACs for the convolution layers that are followed by a max-pooling layer. For example, in the 2×2 max-pooling operation, the maximum value out of the four adjacent convolution outputs is chosen and applied as an activation to the subsequent convolution layers. In other words, the remaining three non-maximum MAC values are discarded. In the proposed PAC approach, these redundant MACs are identified during the time-domain MAC operation and are not completely computed to save energy. A comparison logic is added to keep a track of all four MAC values while accumulating the products of inputs and weights with reduced precision. In the scenario where a given MAC value lags behind the maximum MAC value (out of four MAC values) by a certain pre-defined threshold, the given MAC value is not further computed; thereby resulting in a fewer number of MACs. The simulation results on the AlexNet CNN over 1000 ImageNet images show up to 31.47% (21.79%) reduction in the number of non-zero input activations MACs with a top-five classification accuracy loss of 0.60% (0.90%) in two different PAC modes.
4) An energy-efficient data flow (EEDF) for optimal on-/off-chip memory accesses is also proposed for time-domain MAC computation. A configurable static random access memory (SRAM) bank architecture to allocate dynamic storage space to the input activations and weights depending upon the convolution layer architecture is proposed. Furthermore, a convolution-stride-aware mapping of input activations and weights to the SRAM array is proposed to optimize for on-/off-chip memory accesses. In addition, input–output activation and weight-reuse techniques are proposed for on-chip access reduction, and the run-length coding (RLC) technique is used for off-chip memory access reduction. COMPAC data flow results in 86.97% reduced on-chip accesses and 29.74% reduced off-chip accesses compared with an Eyeriss [4], compared at an iso-8-bit precision.
5) The COMPAC CNN engine is demonstrated on the 65-nm CMOS node, implementing the AlexNet CNN, achieving energy efficiency of 1.044 TOPS/W, a throughput of 0.1278 GOPS, and a top-five classification accuracy of 76.90% at 720 mV.

This article is organized as follows. Section II describes the background of this work and discusses the basics of CNNs and motivations for the CTD multi-bit input/weight MAC computation, PAC technique, and EEDF. Section III presents the concept, architecture, design, and implementation details of the proposed COMPAC CNN engine. Section IV presents the 65-nm CMOS test-chip measurement results. Section V concludes this article with the key findings.

II. Background

A. AlexNet CNN

A typical CNN consists of convolution layers, pooling layers, activation layers, and FCN layers. The convolution layer serves as the core layer, which implements multiple trainable filters that are used to extract features, such as edges, gradients, and colors, in an image. The major operation in CNNs is the MAC operation, which is computed by performing the dot product of a weight matrix and a portion of the input image matrix [8]. Each filter is convolved across the width and height of the input activation, computing multiple MACs and producing a 2-D output activation map for a given filter. Typically, an activation layer is used between consecutive convolutional or fully-connected layers. This layer is used to add non-linearity to a neural network. The rectified linear unit (ReLU) is the most commonly used activation layer that returns 0 if it receives any negative input and returns the same value if it is positive [10]. Pooling (sub-sampling) is performed to reduce the data size feeding into the next layer [11], [12]. Max-pooling is commonly used for the sub-sampling operation that chooses the maximum MAC value across a small output window (e.g., 2×2) [12]. The FCN layers are usually used in the end, once the data size of input feature maps is reduced after passing through the pooling layers. A typical FCN layer comprises of a finite number of feature maps, each of size 1×1. Each of these feature maps is connected to all the feature maps of the previous layer. For example, the AlexNet CNN contains eight trainable layers; the first five are the convolution layers, and the remaining three are the fully connected layers [3]. The output of the last fully connected layer is fed to a 1000-way soft-max layer that produces a distribution over the 1000 class image labels. Max-pooling layers follow the first, second, and fifth convolution layers. The ReLU non-linearity is applied to the output of every convolution and FCN layer.

B. Prior Work

As MAC operations constitute a significant portion of the total CNN power budget [6], various methods have been explored for compact data representation to improve energy efficiency. The data in digital domain are represented as multi-bit digital vector [4], [13]–[18]. This form of data representation is implemented using a higher number of nets, translating into a high dynamic switching capacitance (C_{DYN}) and, consequently, higher power and area. In the analog-domain approach, data are represented as a continuously varying voltage signal. Various design techniques using charge manipulation schemes and analog-to-digital (A/D) converters have been proposed to realize efficient MAC computations in a CNN accelerator. Analog approaches [19]–[21] compute MAC operation in analog voltage domain using

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 18,2021 at 10:12:23 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

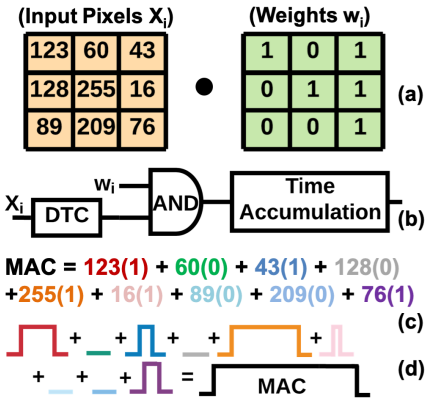


Fig. 1. Concept of the time-domain MAC operation [9]. (a) Dot product operation of input matrix and weight matrix. (b) Time-domain MAC circuit concept. (c) MAC operation in the digital domain. (d) MAC operation in the time domain.

a SRAM array, capacitors, and data converters (ADCs and DACs). However, finite voltage headroom and the sensitivity of circuit parameters to a slight change in analog-domain signals limit the voltage scalability, thereby degrading the MAC accuracy [20]. In the frequency-domain approach, data are represented as signals with varying frequencies using ring oscillators or *RC* loaded circuits. In [23] and [24], MAC operations are computed in the frequency domain using a digital controlled oscillator (DCO) with either resistor or capacitor loading. Various nodes of a ring oscillator are loaded with different capacitor banks to represent multi-bit weights altering the *RC* time constant of the oscillator. Thus, large-range frequency generators/modulators limit the performance scalability of such frequency-domain approaches. Moreover, larger magnitude inputs would result in higher toggle activity incurring higher switching power. In the time-domain computing approach, data can be represented as the PWM inputs or pulse position modulated inputs or a combination of thereof. In [24], the time-domain analog-digital mixed-signal processing approach is proposed to implement a time-domain binary neural network. A time-mode adder circuit is proposed in [25] to realize a trans-linear principle in time.

In our earlier work [8], [9], a multi-bit digital bit-stream is encoded as a single PWM signal with its pulsewidth representing the magnitude of the input data. The multi-bit data are compacted to a single data signal, resulting in reduced C_{DYN} compared with the conventional digital data representation. We presented a detailed comparison between the time-domain computing approach and the sequential digital implementation, performing 1-/8-bit weight and 8-bit input MAC operations, and observed 2.09×-2.32× higher energy efficiency and 2.22×–3.45× smaller area for the time-domain implementations [9]. Furthermore, this approach leverages standard digital gates with full rail-to-rail voltage swing outputs. This mitigates the reduced voltage headroom challenges encountered in an analog-voltage-domain approach enabling ultralow voltage operation. In addition, this approach does not require multiple clock sources (unlike frequency-domain approaches), which results in reduced C_{DYN}. The concept of the time-domain MAC computation using PWM inputs is illustrated in Fig. 1. In this example, the MAC operation is performed by adding the products of a 3 × 3 weights matrix and input pixel matrix [see Fig. 1(a) and (c)]. An input pixel value is encoded into a PWM signal using a digital-to-time converter (DTC). Then, the input PWM signal is multiplied by the corresponding weight bit using an AND gate. This time-encoded product signal is then passed onto the time accumulator circuit [see Fig. 1(b)] to compute MAC operation in the time domain. Sequentially, all input pixel values are encoded as PWM signals, and its product with respective weight bit is loaded on the time accumulator circuit. Thus, all time-encoded products are applied onto the time accumulator circuit, and the width of these timing pulses gets accumulated to realize the MAC operation in the time domain [see Fig. 1(d)].

C. Motivation for CTD Multi-Bit Input and Weight MAC Operation

Complex CNNs, such as AlexNet [3] and VGG [26], compute multi-bit input and multi-bit weight MAC operations to achieve the state-of-art classification accuracy on complex data sets, such as 1000-class ImageNet [3]. In our previous work [8], [9], MAC operation with binary weights is demonstrated. Earlier proposed techniques to perform multi-bit weight time-domain MAC operation using either multiple MDLs or a single configurable-length MDL incur high power/area along with classification accuracy loss; thereby making the time-domain MAC approach less feasible for complex CNNs [8], [9]. There arises a critical need to devise better MDL designs to perform energy-efficient multi-bit input/weight MAC operation in the time domain to make it feasible for complex CNNs. In this article, the TRS technique is proposed to perform the multi-bit input and multi-bit weight MAC operation with minimal MDL time residue, minimal area, and power overhead while maximizing energy efficiency.

Moreover, one of the tradeoffs of the time-domain computation is its sequential nature, resulting in low MAC throughput. In the conventional time-domain approach, as discussed in our previous work [8], [9], the PWM signal is generated synchronously consuming $2^{(n-1)}$ clock cycles to time-encode an n-bit input. PWM signal is held high for time duration proportional to its magnitude and held low for the remaining time duration (dead time), thereby resulting in a significant throughput degradation. In order to overcome this low throughput limitation, a CTD approach is proposed in this work for compact input encoding, in which the total time taken to time-encode an input is proportional to its magnitude (realized by eliminating the dead time using a negative-edge and zero-value detector circuit).

D. Motivation for PAC Operation

The CNN inference process requires the computation of an enormous number of MAC operations, thereby making it challenging to implement in the energy-constrained edge devices. For example, widely known CNNs, such as AlexNet [3] and VGG [26], comprise of 60 million and 138 million filter weights, respectively, which are used to compute 724 million and ≈15.5 billion MAC operations [4]. Thus, reducing the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 18,2021 at 10:12:23 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

number of MACs without compromising the classification accuracy is highly desirable. In a typical CNN, a max-pooling layer is used for the sub-sampling operation that chooses a maximum MAC value across a small output window. For example, in a 2 × 2 max-pooling operation, the maximum value out of the four convolution outputs is chosen and fed to the subsequent convolution layers. In other words, the remaining three values are not used. In this article, a PAC approach is proposed, in which these redundant MACs are identified during the MAC operation and are not completely computed to save energy. A comparison logic is added to keep a track of all four MACs while accumulating the products of inputs and weights with reduced precision. A given MAC value is not further computed if it lags behind the maximum MAC value by a certain threshold, thereby reducing # MACs.

E. Motivation for EEDF

A significant portion of the total accelerator energy is spent in accessing the parameters and activations from either on-chip or off-chip memory. Since deep CNNs comprise a large number of parameters, it is not feasible to store all the activations and weights on-chip. This results in significant energy consumption in off-chip memory accesses, which is one of the critical factors affecting energy efficiency. The energy spent to access the data from the off-chip memory is around 200× compared with the MAC computation energy [4]. Despite pruning the input data and filtering out zero weights and/or inputs, a large number of MAC operations are still performed; which leads to a large number of off-chip accesses. The other critical component of power consumption in these networks is the on-chip memory buffer (SRAM) accesses. The energy spent to access the data from an on-chip memory buffer is around 6× compared with the MAC computation energy [4]. Moreover, the number of on-chip data accesses is significantly more than the number of off-chip dynamic random access memory (DRAM) accesses, making the on-chip memory access an important concern for energy-efficient edge computation.

III. Proposed COMPAC CNN Engine Design

A. Architecture Overview

The proposed COMPAC CNN engine supports 32 filters, a CTD controller, multiple address generators, registers, and a 67-kB on-chip SRAM to perform a pooling-aware 8-bit input and 8-bit signed weight MAC operation for the AlexNet CNN in the time domain (see Fig. 2). The overall operation is briefly discussed as follows. First, the input activations and the weights of 32 filters are accessed from the off-chip DRAM memory (Xilinx VC707 FPGA resources) using a 32-bit input data bus and stored in the global on-chip buffer (SRAM) with a storage capacity of 67 kB. The input activations are then accessed from the global buffer and stored locally in 144 input activation cyclic shift registers (each 32-bit wide). The address generator and controller are used to access the input activations from the on-chip SRAM and efficiently broadcast them to the activation registers such that each activation is accessed only once from on-chip SRAM and stored in multiple registers,

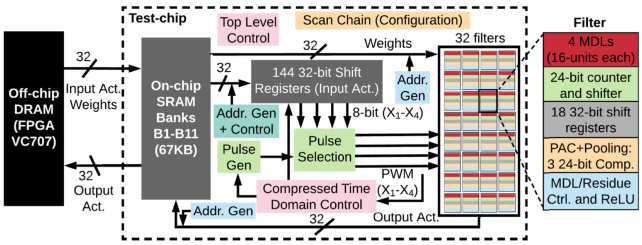


Fig. 2. Proposed COMPAC CNN engine system-level architecture.

thereby optimizing the total number of on-chip accesses. The pulse generator generates 16 free-running PWM signals, one of which gets selected in the pulse-selection module depending upon the input activation value. To improve the throughput, a CTD approach is proposed in this work, where the pulse generation occurs asynchronously, to eliminate the dead time using a CTD controller. The four time-encoded PWM inputs for X_1–X_4 are sent to all 32 filters to perform the MAC operations in the time domain. Each filter comprises of four 16-unit long MDLs, a 24-bit counter, a 24-bit shifter, nine 32-bit shift registers for weight bit storage, nine 32-bit cyclic shift registers to store the sign bit of the weight, three 24-bit comparators to perform comparison during PAC and max-pooling operations, an ReLU activation controller, and an MDL/residue controller for the reliable and correct MDL operation. The 8-bit input/weight MAC operation is computed sequentially by accumulating the products of time-encoded input with the most significant weight bit and repeating it for the next less-significant weight bits. The weight bit and the sign bit are optimally accessed from the on-chip global buffer using an address generator and stored locally in 18 32-bit registers of each filter. The pooled output value of each filter is stored in the global-buffer, and once the MAC computation for all the input activations stored on-chip for 32 filters is completed, the output activations are sent through a 32-bit output data bus and stored off-chip (FPGA VC707). This process is repeated until all the MAC operations for the entire input activation map and for all the filters are computed, for each convolution layer.

B. Multi-Bit Input and Weight MAC Using MDL With TRS

1) Background: Time-Domain MAC Computation Using MDL: The concept of an MDL to perform 1-bit weight time-domain MAC operation with reduced switching capacitance for an energy-efficient edge computing was discussed in our previous work [8], [9]. MDL is comprised of a string of gated delay units (termed MDL units). The MDL design is derived from the concept of a time register used in high-precision time-to-digital converters, which can perform time addition and time storage using a string of delay cells [27]. Each MDL unit is controlled by an EN signal. The product of time-encoded input activation and weight bit acts as an EN signal. MDL acts as a forward (backward) delay line for positive (negative) weight when EN is held high and retains the MDL state when EN is held low. The 0→1 transition is detected at either end of the MDL (node A or node E) when a

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 18,2021 at 10:12:23 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

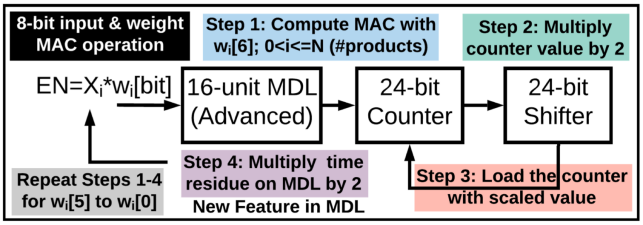


Fig. 3. 8-bit input/weight MAC operation in time-domain using a single 16-unit MDL (with TRS), counter, and shifter.

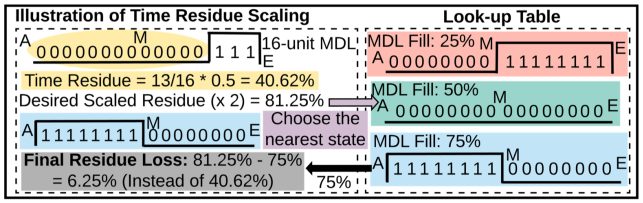


Fig. 4. Proposed concept of TRS on MDL for multi-bit weight/input MAC operation.

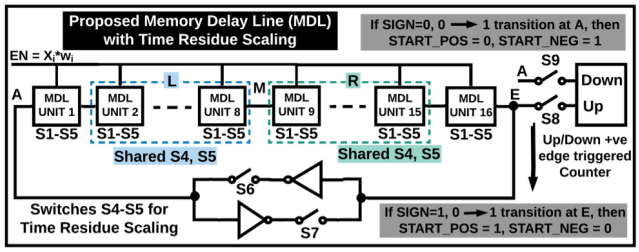
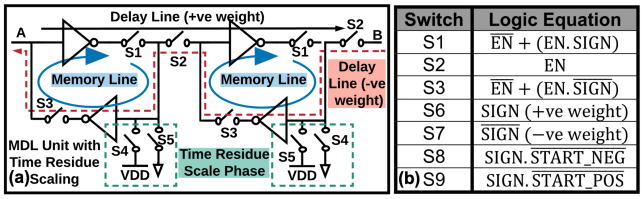


Fig. 5. Proposed MDL with TRS features for multi-bit weight/input MAC operation.



Switch	Logic Equation
S1	$\overline{EN} + (EN.SIGN)$
S2	EN
S3	$\overline{EN} + (EN.\overline{SIGN})$
S6	SIGN (+ve weight)
S7	$\overline{SIGN}$ (−ve weight)
S8	SIGN.START_NEG
(b) S9	SIGN.$\overline{START_POS}$

Fig. 6. (a) Proposed MDL unit with TRS features for multi-bit weight/input MAC operation. (b) Switch configurations for the proposed MDL supporting multi-bit weight/input MAC operation.

string of "0's" followed by a string of "1's" is propagated. The delay variations in MDL due to process variations (random errors) or temperature changes (systematic error) are mitigated by the addition of a calibration unit in MDL. Thus, a finite MDL length along with an up–down counter can be used to perform signed time accumulation of a long duration to perform MAC operation.

2) Proposed Multi-Bit Input/Weight MAC Operation in Time Domain: The multi-bit weight time-domain MAC operation is performed by sequentially accumulating the products of time-encoded input with each weight bit, using a single fixed-length MDL (see Fig. 3), thereby overcoming the limitations of multi-bit weight time-domain MAC, as discussed in our previous work [9]. The counter value and the time residue on MDL are scaled by a factor of 2 to match its bit precision, before accumulating the product of the input with the next less-significant weight bit (see Fig. 3). The counter value is scaled by a factor of 2 using a shifter (left shift by 1), whereas time residue on MDL is scaled by a factor of 2 using the TRS technique, as discussed in Section III-B3. This process is repeated until the products of time-encoded input with all the weight bits are accumulated (see Fig. 3).

3) Concept of TRS: The time residue on each MDL needs to be scaled by a factor of 2, before accumulating the products of input activation with the next less-significant weight bit on MDL (to match the precision). This scaling has been realized efficiently using a lookup table approach (see Fig. 4). The TRS lookup table comprises of three MDL states when MDL is 25% fill, 50% fill, and 75% fill. The state of the MDL is first determined by observing the start (node A), middle (node M), and end (node E) nodes of the MDL, and then, the scaled residue state is chosen from one of the TRS lookup table MDL states (see Fig. 4). For instance, the time residue on MDL is 40.62% of the MDL full-length delay, which needs to be scaled by a factor of 2. The desired scaled residue is $40.62\% \times 2 = 81.25\%$. The nearest residual value of this desired scaled value is chosen, which is 75% of the full-length MDL delay, thereby minimizing the time residue loss (see Fig. 4). The simulation results of the AlexNet CNN on 1000-class ImageNet data set have shown ≈3% improvement in the top-five classification accuracy with the proposed TRS technique compared with the baseline MDL approach (wo/TRS).

4) Implementation of TRS in MDL: The bi-directional MDL is designed to support the TRS for multi-bit weight MAC operation (see Fig. 5). The state of each MDL unit [see Fig. 6(a)] is controlled by observing the start (node A), middle (node M), and end (node E) nodes of the MDL such that the time residue loss is minimized to achieve high classification accuracy. MDL unit acts as a forward (backward) delay unit for positive (negative) weight when EN is held high and retains MDL state when EN is held low during normal MDL operation (controlled using switches S1—S3) and sets its logic node value to 0 or 1 to scale the residue on MDL during the TRS phase (controlled using switches S4 and S5). Fig. 6(b) lists down the switch configurations (S1–S3 and S6–S9) for the memory and delay phases of the MDL operation. The basic operation of the memory and delay phases is briefly discussed in Section III-B1. A comprehensive account of the working principle of MDL has been presented in our previous work [9].

In order to realize energy and area-efficient implementation of the TRS, only three nodes (nodes A, M, and E) are observed to determine the desired MDL state with scaled time residue. There are 128 TRS controllers (one per each MDL), each occupying a core area of 96 um^2. Thus, the TRS modules occupy 0.0123 mm^2, contributing to a 0.71% increase in the total core area. The power and latency overheads of the TRS are insignificant compared with the MAC accumulation phase, which comprises of many clock cycles [equals filter dimensions × channels × 7.15 cycles (to time-encode 8-bit input)]. This efficient lookup table-based TRS approach has

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 18,2021 at 10:12:23 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE I
LOOKUP TABLE FOR THE PROPOSED TRS APPROACH

MDL State Vector and Phase				Orig. Residue Loss (%MDL full length delay)	Switches S4/S5 (Output MDL State Vector) and Counter Incr./Decr.					Final Residue Loss (% MDL)
A	M	E	Phase		Unit 1	Units 2-8	Units 9-15	Unit 16	Counter	
0	1	1	+ve	0-25%	1/0	1/0	0/1	0/1	No	<±25%
0	0	1	+ve	25-50%	0/1	0/1	1/0	1/0	No	<±25%
1	0	0	+ve	50-75%	1/0	1/0	0/1	0/1	Incr.	<±25%
1	1	0	+ve	75-100%	1/0	1/0	1/0	0/1	Incr.	<50%
1	1	0	-ve	0-25%	0/1	0/1	1/0	1/0	No	<±25%
1	0	0	-ve	25-50%	1/0	1/0	0/1	0/1	No	<±25%
0	0	1	-ve	50-75%	0/1	0/1	1/0	1/0	Decr.	<±25%
0	1	1	-ve	75-100%	0/1	1/0	1/0	1/0	Decr.	<50%

resulted in ≤±25% of MDL full-scale delay, thereby results in high classification accuracy. The logic values on nodes A, M, and E are observed to determine the initial state (time residue) of the MDL. The phase of the MDL, i.e., whether it is holding a positive or negative residue, is also determined from the MDL controller. The switch S4 and S5 values for the MDL unit 1, units 2–8, and units 9–16 along with the counter increment/decrement signal are generated from the TRS controller to scale the time residue on MDL such that the final residue loss is minimized (see Table I). Besides ensuring minimal accuracy loss, it is also ensured that weights can be reused to reduce the on-chip data accesses. This requires a co-design of the TRS controller and weight address generator to scale the time-domain approach for multi-bit weight MAC operation. It should be noted that the residue loss can be further minimized by designing a relatively complex TRS controller with a higher area/power overhead, which observes more number of MDL internal nodes and outputs more sets of S4/S5 switch values to precisely control MDL state. This would lead to lower time residue loss and better accuracy.

C. Proposed CTD Approach

1) Concept of CTD Approach: In the conventional time-domain approach as discussed in our previous work [8], [9], the PWM signal is generated synchronously consuming $2^{(n-1)}$ clock cycles to time-encode an n-bit input. PWM is held high for the time duration proportional to its magnitude and low for the remaining time duration (dead time), thereby resulting in a significant throughput degradation [see Fig. 7(a)]. A CTD approach is proposed for the compact input encoding in which the total time taken to encode an input is proportional to its magnitude only [see Fig. 7(c)]. This is achieved by eliminating the dead time (time duration when a PWM signal is held low). The proposed CTD approach combines the throughput benefits obtained by skipping the zero-magnitude input activations [see Fig. 7(b)] and eliminating the dead time for the input activations that have a magnitude less than the maximum value ($2^n - 1$ for an n-bit input).

2) Implementation of CTD Approach: The proposed CTD approach is implemented using a negative-edge and zero-value detector circuit (see Fig. 8). The proposed COMPAC CNN engine performs an on-chip 2 × 2 max-pooling by computing four MAC values simultaneously. Each MAC value is obtained by computing a dot product of input activation (X_1–X_4) with

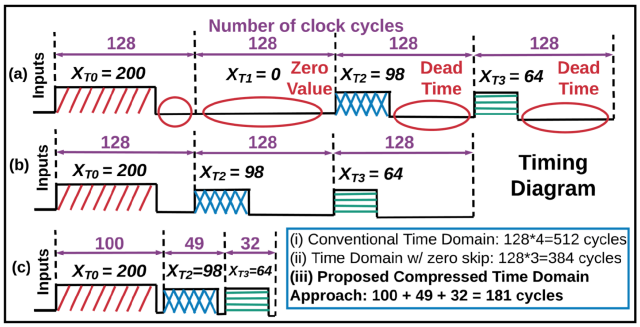


Fig. 7. (a) Conventional time domain approach. (b) Conventional time domain approach with zero-skipping. (c) Proposed concept of the CTD approach in the COMPAC CNN engine.

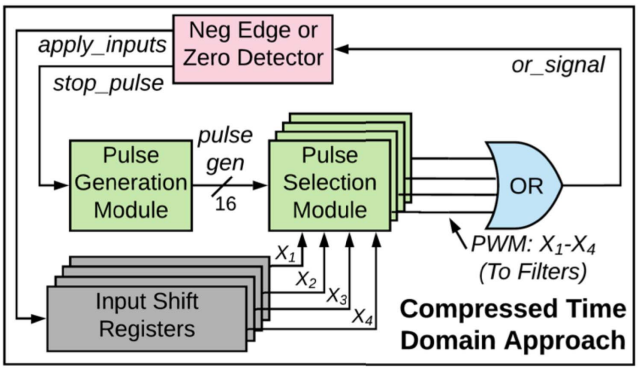


Fig. 8. Circuit implementation of the CTD approach in the COMPAC CNN engine.

the weight matrix. These input activations are stored in the input activation shift registers: 16 free-running PWM signals are generated in the pulse generation module such that the pulsewidth of each consecutive signal increases by *to*, where *to* equals the half period of the input clock. Thus, the pulsewidth of the PWM signal varies from 0 × *to* to 15 × *to*.The pulse generation scheme of the proposed work generates different PWMs based on synchronous input clock operation. The pulse generation approach is not based on a logic path delay difference, which is susceptible to pulse shrinking and expansion due to process variations. Thus, no such shrinking and expansion effects are expected in pulse generation. The pulse-selection module comprises one 16:1 MUX. The 4-bit input activation value serves as the four select lines of this 16:1 MUX to select one of the 16 PWM signals, based on input activation magnitude. Four pulse-selection modules are used to time-encode four 4-bit input activation values (X_1–X_4). These time-encoded input activation signals are passed through a four-input OR gate to find the time-encoded input with a maximum pulsewidth (*or_signal*). This signal is passed to the negative edge and the zero value detector to determine the negative edge or zero-magnitude value. Once any of these two events occur, a trigger signal (*stop_pulse*) is generated on the next edge of the input clock (adding a delay of half input clock period), and the pulse generator is stopped on the next subsequent clock edge. Then, the next input activations are applied to the next subsequent clock edge by shifting

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 18,2021 at 10:12:23 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

the input activations from the shift registers, by asserting an *apply_inputs* signal. Finally, the pulse generator starts on the next subsequent clock edge.

Thus, the proposed CTD approach time encodes the input activations in an asynchronous manner and incurs a time overhead of two input clock cycles (half-cycle each for trigger generation, stopping the pulse generator, applying the next inputs, and starting the pulse generator) to ensure a reliable time encoding (without any glitches) of input activations. Thus, the time taken to encode input activation equals (max(X_1 to X_4) $\times$to) + 2 $\times$ input clock cycle period (time overhead) in the proposed CTD approach, whereas it takes ($2^{(n-1)} \times to$) + 1 $\times$ input clock cycle period (time overhead to ensure no race around condition and/or glitch formation) in the baseline MDL approach [8], [9].

The conventional time-domain approach, as discussed in our previous work [8], [9], employs a pulse generator module (generating 16 pulses in MSB phase and 16 pulses in LSB phase) and a pulse selector module (four 2:1 Muxes and a 16:1 MUX), whereas the proposed CTD approach employs a smaller pulse generator module (16 pulses only), a smaller pulse selector module (16:1 MUX only), and a CTD controller. The CTD circuit occupies a core area of 920 um^2, whereas the conventional time-domain pulse approach occupies 642 um^2. Thus, CTD results in a core area overhead of 278 um^2, which is 0.016% of the total core area. Furthermore, CTD modules consume a total power of 18.72 μW, whereas the conventional time-domain approach consumes 14.78 μW, thereby resulting in a power overhead of 3.94 μW and amounting to $\approx$3% of the total core power.

3) Simulation Results of CTD on AlexNet CNN: The proposed concept of the CTD approach is validated for all the five convolution layers of the AlexNet CNN on 1000 images of the ImageNet validation data set (1000 class). One image is randomly chosen from each class of the validation data set. The time encoding of an 8-bit input (X[7:0]) consumes 14.71 input clock cycles on an average (over five convolution layers) using CTD, whereas the conventional time-domain approach consumes 129 cycles when simulated on AlexNet over 1000 images, thereby improving the throughput by 88.60% compared with the conventional time-domain approach [see Fig. 9(a)]. To further improve the throughput, an 8-bit input is time encoded using CTD in two phases: four MSBs (X[7:4]) and four LSBs (X[3:0]) of input activation, as shown in Fig. 9(b). Thus, the time encoding of 8-bit input (X[7:0]) in two phases (4 bits each) on the AlexNet over 1000 images consumes 7.15 input clock cycles on an average, further improving the throughput by 51.40%, compared with the CTD encoding of an 8-bit input activation in a single phase.

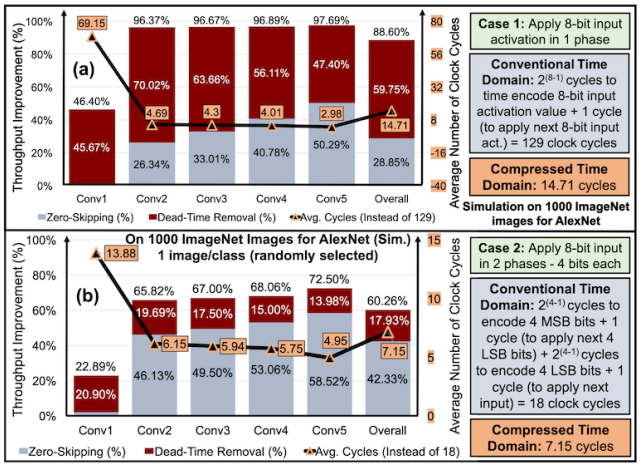


Fig. 9. Throughput Improvement in the proposed CTD approach to time-encode 8-bit input activation in (a) one phase of 8 bit and (b) two phases of 4 bit.

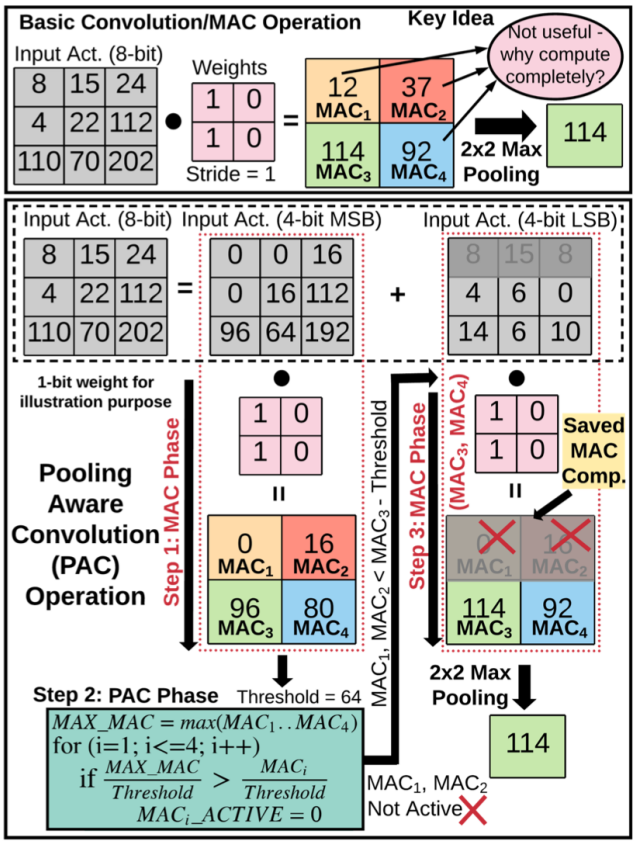


Fig. 10. Proposed concept of the PAC operation.

D. PAC to Reduce # of MACs

1) Concept of PAC: A generic CNN comprises of a pooling layer to perform a sub-sampling operation. For instance, in 2 $\times$ 2 max-pooling, only the maximum value among the four (2 $\times$ 2) MACs is fed to the next layer, and the remaining MAC outputs are not used (see Fig. 10). In the proposed PAC approach, these redundant MACs are identified during the MAC operation and are not completely computed to save energy. MAC operation is computed in multiple phases (accumulating products of input/weight with a smaller bit width in each phase), separated by a comparison phase. In the comparison phase, the MAC values that are smaller than the maximum MAC value by a certain predefined threshold are

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 18,2021 at 10:12:23 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

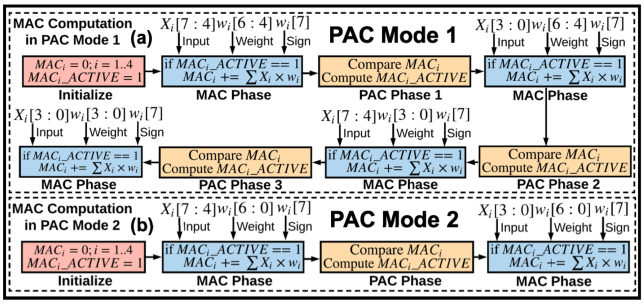


Fig. 11. MAC computation process in (a) PAC Mode 1 and (b) PAC Mode 2.

TABLE II

THRESHOLD VALUES FOR DIFFERENT RUNS AND ALEXNET CONV. LAYERS IN (A) PAC MODE 1 AND (B) PAC MODE 2

(a) Run	Threshold Values for PAC Mode 1 - AlexNet CNN								
	Layer – Conv1			Layer – Conv2			Layer – Conv5		
	Pha-se1	Pha-se2	Pha-se3	Pha-se1	Pha-se2	Pha-se3	Pha-se1	Pha-se2	Pha-se3
Run1	2^{16}	2^{15}	2^{10}	2^{12}	2^{12}	2^{9}	2^{10}	2^{11}	2^{7}
Run2	2^{15}	2^{14}	2^{10}	2^{11}	2^{12}	2^{8}	2^{10}	2^{11}	2^{7}
Run3	2^{15}	2^{14}	2^{10}	2^{11}	2^{11}	2^{7}	2^{10}	2^{11}	2^{7}
Run4	2^{15}	2^{14}	2^{10}	2^{10}	2^{10}	2^{6}	2^{10}	2^{11}	2^{7}
Run5	2^{14}	2^{13}	2^{9}	2^{9}	2^{10}	2^{6}	2^{10}	2^{10}	2^{6}
Run6	2^{13}	2^{12}	2^{8}	2^{8}	2^{9}	2^{5}	2^{9}	2^{9}	2^{5}
Run7	2^{12}	2^{12}	2^{8}	2^{8}	2^{8}	2^{4}	2^{8}	2^{8}	2^{4}
Run8	2^{12}	2^{12}	2^{8}	2^{7}	2^{8}	2^{4}	2^{8}	2^{8}	2^{4}

(b) Run	Threshold Values for PAC Mode 2 - AlexNet		
	Layer Conv1	Layer Conv2	Layer Conv5
RunA	2^{12}	2^{10}	2^{8}
RunB	2^{12}	2^{10}	2^{7}
RunC	2^{12}	2^{9}	2^{6}
RunD	2^{11}	2^{8}	2^{5}
RunE	2^{10}	2^{7}	2^{5}
RunF	2^{10}	2^{6}	2^{6}
RunG	2^{10}	2^{6}	2^{5}

eliminated, and the remaining bits of these MACs are not computed in subsequent phases, thereby reducing the total number of bit-wise MACs. For example, the MAC operation with four MSBs (X[7:4]) of the input activation and weight is first computed (see Fig. 10). This is followed by a comparison operation that identifies and eliminates those MAC values (MAC1 and MAC2), which are smaller than the maximum value by a certain predefined threshold. The non-eliminated MAC values (MAC3 and MAC4) are then fully computed by accumulating the products of four LSBs (X[3:0]) of the inputs and weight.

2) PAC Modes in COMPAC CNN Engine: In this work, two modes of PAC are proposed for the 8-bit input activation and 8-bit weight MAC operation. In PAC mode 1, input activation and weight values are applied in four MAC phases (4 bits of input and weight in each phase), with a comparison PAC phase in between each MAC phase [see Fig. 11(a)], whereas PAC mode 2 comprises of two MAC phases (computing MAC for 4-bit input and 8-bit weight in each phase) separated by a comparison PAC phase [see Fig. 11(b)]. It should be noted that the chosen threshold values are multiples of 2^n such that the shifters and comparators can be reused during the PAC phase operation, thereby eliminating the need for additional circuitry.

3) Simulation Results of PAC on AlexNet CNN: The simulation results on the Alex Net CNN over 1000 ImageNet validation data set images (one image randomly chosen from each class) for different thresholds (see Table II) show up to 31.47% (21.79%) reduction in the number of non-zero input activations MACs with a top-five classification accuracy loss of 0.60% (0.90%) and a top-one classification accuracy loss of 2.20% (1.90%) for the PAC modes 1 (2), respectively (see Fig. 12). The reduction in the number of MACs translates

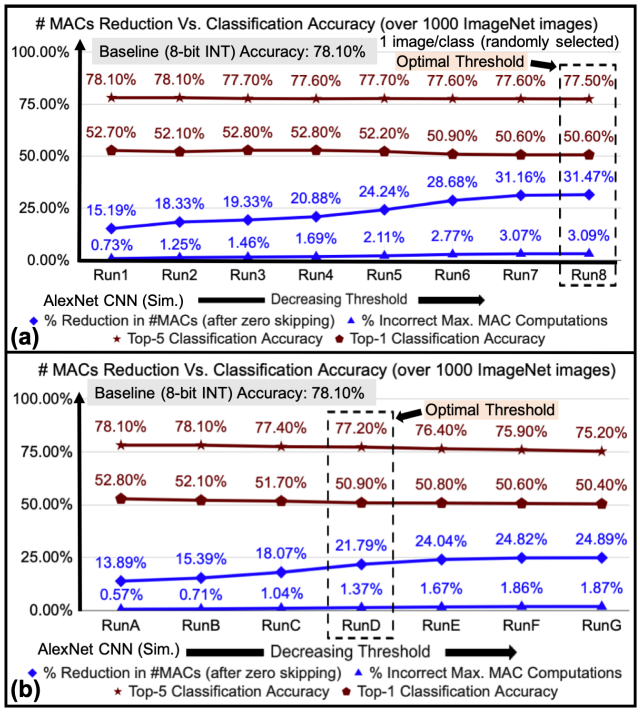


Fig. 12. Reduction in # of MACs and classification accuracy on AlexNet for different threshold values in (a) PAC Mode 1 and (b) PAC Mode 2.

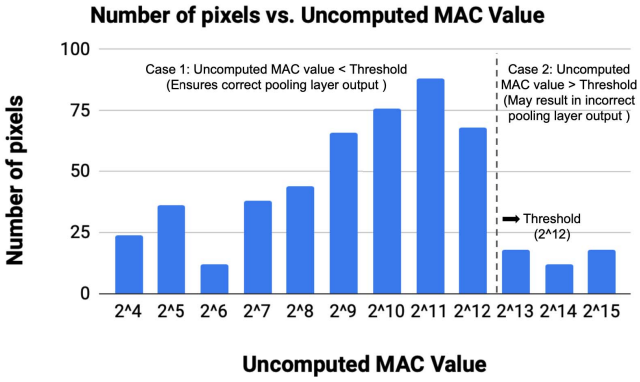


Fig. 13. Illustration of selecting optimal threshold value.

to power savings since the MDLs remain idle for those MACs that are not completely computed. Thus, MDL power is saved by the same amount by which the number of MAC operations is reduced. The power overhead due to shifters and comparators is insignificant compared with MAC computation since the PAC phase (duration: two cycles) is followed by a MAC accumulation phase, comprising of many cycles [equals the filter dimensions $*$ channels $*$ 50.05 (cycles for 1 MAC)]. Thus, the power overhead of PAC circuit elements can be ignored since they remain inactive most of the time, and effective power savings equal reduction in MDL power.

Table II lists down the threshold values used in PAC modes 1 and 2. The optimal values of these threshold values are determined using an empirical model discussed as follows. First, for each PAC phase in PAC modes 1 or 2,

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 18,2021 at 10:12:23 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE III
COMPARISON OF NUMBER OF MACS, ACCURACY, AND ON-CHIP ACCESSES OF THE PROPOSED PAC MODES 1 AND 2 WITH BASELINE, NON-ZERO INPUT, AND PC [28] APPROACHES

Operation Mode – AlexNet CNN	#MACs (Million)	Top-5 Acc. (%)	Top-1 Acc. (%)	Incorrect Max MAC (%)	On-chip Access (MB)
Baseline	666.0	78.10%	52.80%	0.00%	13.58 (1.00 x)
Non-zero input	325.3	78.10%	52.80%	0.00%	13.58 (1.00 x)
Prec. Cascading	202.5	70.50%	43.60%	22.13%	19.60 (1.44 x)
PAC Mode 1	222.9	77.50%	50.60%	3.09%	21.80 (1.60 x)
PAC Mode 2	270.1	77.20%	50.90%	1.37%	14.67 (1.08 x)

the uncomputed MAC value (MAC value—partially computed MAC value till a given PAC phase) is determined for each pixel for a given convolution layer output over ImageNet validation data set images. Here, the uncomputed MAC value signifies the remaining MAC value, which would be added to the partially computed MAC value after a given PAC phase. Next, the frequency bin distribution of these uncomputed MAC values for all the pixel values for a given layer is plotted (see Fig. 13), where the X-axis represents the range of uncomputed MAC values, and the Y-axis represents the frequency (number of uncomputed MAC values). The lowest value on the X-axis, which covers most of the area under the graph (most of the bins), is chosen as the optimal threshold value for a given PAC phase and convolution. The uncomputed MAC values that lie to the left of this chosen threshold value on the graph (case 1 in Fig. 13) would not result in incorrect output values after max-pooling layer operation, thereby ensuring minimal accuracy loss while maximizing the reduction in the number of MACs.

There is a tradeoff between classification accuracy and a reduction in the number of MACs. A higher threshold value results in a lesser reduction in the number of MACs, thereby having a low chance of predicting the incorrect maximum MAC value, leading to lesser accuracy loss, as shown in Fig. 12. Table III summarizes the number of MACs, top-one/top-five classification accuracy, and the on-chip SRAM access overhead for the AlexNet CNN without/with PAC modes. The proposed PAC approach results in an on-chip SRAM access overhead of 60.53% (8.03%) for the PAC modes 1 (2) (see Table III). It should be noted that the proposed PAC approach does not affect the off-chip DRAM accesses. Considering the on-chip access overhead, PAC still results in overall energy reduction. PAC mode 2 results in a 21.79% reduction in the number of MAC operations, which translates to a reduction of 145.12 million MAC operations (total MACs = 666M for AlexNet). This is achieved at an on-chip access overhead of 14.67–13.58 MB = 1.09 MB (refer Section IV-C). Assuming that the on-chip access energy is 6× times the MAC compute operation [4], it translates to the MAC compute overhead of $1.09 * 1024 * 1024 * 6 = 6.86$ million MACs. Therefore, the net MAC savings using the proposed PAC Mode 2 approach are $145.12 - 6.86 = 138.26$ million MACs (equals 20.76% of total MACs). Thus, the tradeoffs between the reduced number of MACs, classification accuracy loss, and on-chip access overhead need to be carefully analyzed. Note that the proposed PAC technique is independent of underlying MDL-based time-domain implementation and can be applied for other digital/analog MAC implementations.

4) Comparison of PAC With Precision Cascading (PC) [28]: Kim and Seo [28] proposed a PC approach to reduce the bit-wise MACs for the convolution layers that are followed by a max-pooling layer. The goal of the PC approach is similar to the proposed PAC approach, i.e., not to fully compute the redundant MACs. In the PC approach, the MAC value is computed in eight rounds, computing MAC for 8-bit weight and 1-bit input value in each round. The maximum MAC value is predicted if it exceeds the remaining MAC values by at least 1. Thus, the PC approach may lead to higher MAC savings at reduced classification accuracy compared with the proposed PAC approach since it has higher chances of predicting incorrect maximum MAC value having a discard threshold of only 1. We compared the PC approach with the proposed PAC approach for the # of MACs and classification accuracy on the AlexNet CNN. The simulation results on the AlexNet CNN over 1000 ImageNet data set images show that the PC approach results in 9.15% fewer # of MACs (202.5 million instead of 222.9 million) at a cost of 7% drop (70.5% instead of 77.5%) in the top-five classification accuracy compared with the PAC-1 mode proposed in this work (see Table III).

E. EEDF for Optimal Data Movement

In deep neural networks, such as AlexNet and VGG, millions of MAC operations are computed using a large number of weight parameters and input activations. Significant energy is spent on accessing these parameters and activations. All the activations and parameters cannot be stored on-chip due to limited on-chip SRAM capacity, resulting in significant off-chip access energy. COMPAC proposes an EEDF to reduce the on-/off-chip accesses. The input activations and weights for any given convolution stride are stored on an on-chip SRAM, as given by (1), such that all the column values for the accessed row are utilized in the MAC operation, thereby minimizing the total number of on-chip accesses (see Fig. 14). This is achieved by saving the slice (part) of input activation/weight in each row such that the width (w) and height (h) of the slice do not exceed the convolution stride. The depth (d) of the slice is chosen such that it maximizes the storage utilization of the SRAM array. In the proposed COMPAC CNN engine implementing the AlexNet CNN, 256 column-wide SRAM banks are used. The w, h, and d values (width, height, and depth of the slice storing input activations and weight parameters in a single SRAM row) are set as 4, 2, 3 and 1, 1, 32 for the convolution layers 1 and 2–5, respectively

$$w \times h \times d \times \text{data-width} \leq \#\text{ of columns in SRAM Array} \quad \text{where, w and h} \leq \text{convolution stride.} \tag{1}$$

Besides optimally mapping the parameters and activations to the SRAM array for the time-domain MAC computation, the proposed COMPAC CNN engine comprising of 11 SRAM banks (B1–B11) with an on-chip capacity of 67 kB provides configurable bank architecture to maximize the input activation and weight reuse for different convolution layers of the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 18,2021 at 10:12:23 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

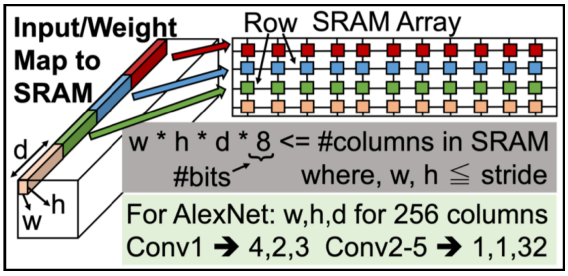


Fig. 14. Convolution-stride-aware data storage technique for optimal on-/off-chip accesses.

TABLE IV
CONFIGURABLE SRAM BANK STRUCTURE TO STORE INPUT ACTIVATIONS, WEIGHTS, AND OUTPUT ACTIVATIONS FOR ALEXNET CNN

SRAM Banks: B1-B7 (8KB), B8-B9 (4KB), B10 (2KB), B11 (1KB)						
Layer	Input Activation		Weights		Output Activation	
	Banks	Capacity	Banks	Capacity	Banks	Capacity
Conv1	B1-B5, B8	44 KB	B6-B7, B10	18 KB	B9, B11	5 KB
Conv2	B1, B8	12 KB	B2-B7, B10	50 KB	B9, B11	5 KB
Conv3	B1, B8, B9	16 KB	B2-B7, B10	50 KB	B11	1 KB
Conv4	B1, B8	12 KB	B2-B7, B9, B10	54 KB	B11	1 KB
Conv5	B1, B8	12 KB	B2-B7, B9, B10	54 KB	B11	1 KB

AlexNet CNN. The configurable bank architecture helps in achieving the optimum number of accesses by allocating more space to the input activations for convolution layer 1 and to the weight parameters for the layers 2–5 (see Table IV).

1) On-Chip Access Reduction Techniques: The on-chip SRAM accesses are optimized using a configurable SRAM bank architecture that supports the convolution-stride-aware mapping of the input activations and weights. Furthermore, input activations, weights, and partial sum data movement are further optimized, as discussed in the following.

The input activation values (X_1–X_4) stored on the SRAM array are accessed and temporarily stored in the input activation shift registers. The inputs are accessed from these registers while performing time encoding and the MAC operations. Typically, a convolution stride is less than a filter window size. Therefore, a lot of input activation values (X_1–X_4) are replicated when stored in these shift registers. To save the on-chip input activation accesses, the inputs from the SRAM array are broadcasted to these shift registers such that each input activation is accessed only once and stored in multiple input shift registers (if required), as shown in Fig. 15. Furthermore, inputs (X_1–X_4) are shared across 32 filters (F1–F32 implemented in COMPAC) to perform four MACs per each filter (see Fig. 15).

Each filter comprises four MAC engines, each supporting an MDL, a counter, and a shifter to perform the MAC operation. Four MAC operations are performed using the same weight (stored in weight register of a filter) and different input activation (X_1–X_4). Thus, the weight value is reused across the four MACs for each filter, as shown in Fig. 16. As discussed earlier, multi-bit input/weight MAC operation is performed sequentially by adding the product of time-encoded input pulse with each weight bit (wt[6]–wt[0]). The MAC operation is

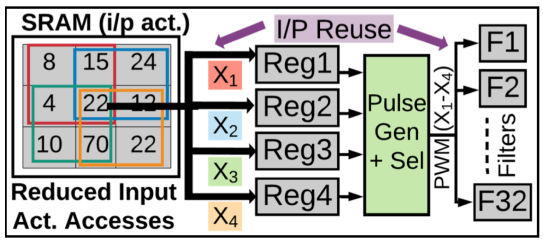


Fig. 15. Proposed input activation reuse techniques in COMPAC CNN engine.

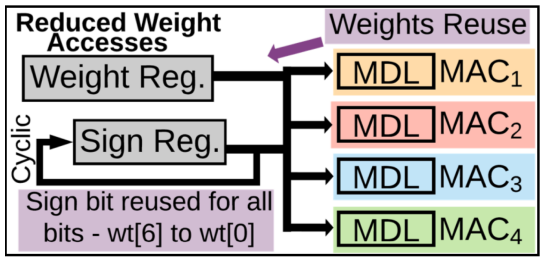


Fig. 16. Proposed weight-reuse techniques in the COMPAC CNN engine.

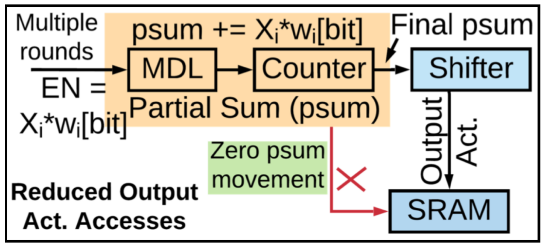


Fig. 17. Proposed output activation reuse techniques in the COMPAC CNN engine with zero partial sum movement.

computed by accumulating products of input pulsewidth with the highest significant weight bit (stored locally in weight shift register), and the products of these inputs with less-significant weight bits are computed in subsequent cycles. However, the weight sign bit (wt[7]) is accessed only once from the SRAM and stored locally in a weight sign cyclic shift register. The sign bits are rotated to reuse for the different weight bits (wt[6]–wt[0]), thereby reducing the on-chip access energy in accessing the sign bit of the weight (see Fig. 16). The weight/input activation reuse and configurable bank architecture schemes are implemented using the address generators, controllers, and bank select logic that occupy 0.025 mm^2 (1.44% of the total core area). The power overhead of these elements can be ignored since only 0.00431 cycles (discussed in Section IV-C) are consumed on an average for each MAC operation (compute time equals 50.05 cycles) to access the on-chip data.

In the proposed COMPAC CNN engine, the MAC operation is computed in the time domain by sequentially accumulating the products of input activation with each weight bit. The partial sum value is stored locally in the counter and is not sent (received) to (from) the global buffer. Once all the products of input activation and weight are accumulated, the fully

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 18,2021 at 10:12:23 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

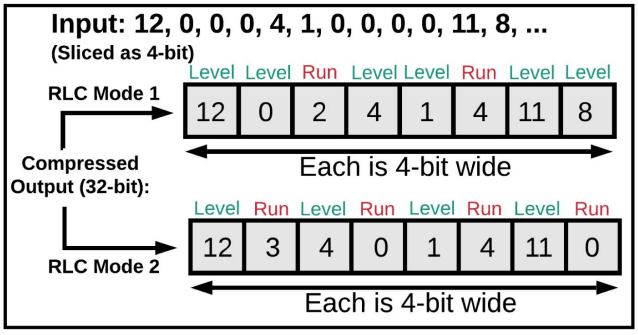
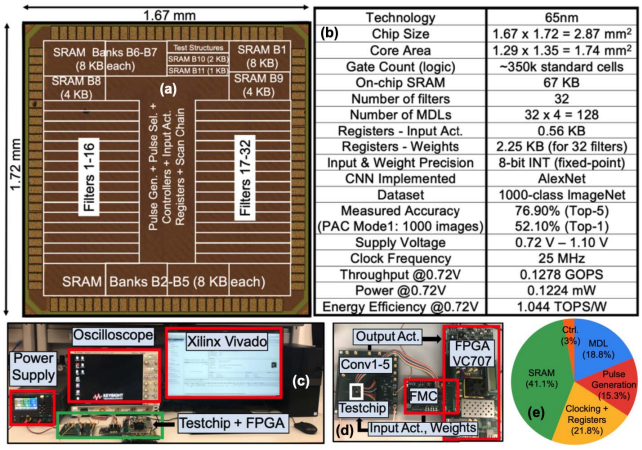


Fig. 18. Run length compression codes for the proposed COMPAC CNN engine in RLC modes 1 and 2.



	Technology	65nm
(b)	Chip Size	1.67 x 1.72 = 2.87 mm^2
	Core Area	1.29 x 1.35 = 1.74 mm^2
	Gate Count (logic)	~350k standard cells
	On-chip SRAM	67 KB
	Number of filters	32
	Number of MDLs	32 x 4 = 128
	Registers - Input Act.	0.56 KB
	Registers - Weights	2.25 KB (for 32 filters)
	Input & Weight Precision	8-bit INT (fixed-point)
	CNN Implemented	AlexNet
	Dataset	1000-class ImageNet
	Measured Accuracy (PAC Mode1: 1000 images)	76.90% (Top-5) 52.10% (Top-1)
	Supply Voltage	0.72 V – 1.10 V
	Clock Frequency	25 MHz
	Throughput @0.72V	0.1278 GOPS
	Power @0.72V	0.1224 mW
	Energy Efficiency @0.72V	1.044 TOPS/W

Fig. 19. (a) 65-nm test-chip die micrograph. (b) Summary table. (c) and (d) FPGA-based lab measurement setup. (e) Test-chip power breakdown.

computed MAC output is written to the on-chip global buffer (see Fig. 17). This results in a significant reduction in the on-chip memory accesses of the output activations since there is no partial sum movement.

2) Off-Chip Access Reduction Techniques: The off-chip DRAM accesses are optimized using a configurable SRAM bank architecture supporting convolution-stride-aware mapping of input activations and weights. In addition, the RLC technique [4] can be used in the proposed COMPAC approach to further reduce the off-chip DRAM accesses, by exploiting the zeros in the input activations (due to preceding ReLU activation layer) and weights. It should be noted that RLC is not implemented in our test chip. However, simulations supporting two different RLC codes are performed to evaluate the off-chip DRAM access savings. Fig. 18 shows both the modes of RLC encoding proposed for the COMPAC CNN engine. There are two types of fields in an RLC compressed output—"Run" and "Level." The consecutive number of zeros to a maximum run length of 15 is represented using a 4-bit number in the Run field, whereas a 4-bit value is inserted directly in the level field. Two (four) runs and Six (four) levels are packed into a 32-bit word, which is sent through a 32-bit input data bus in RLC modes 1 (2), respectively (see Fig. 18). These two RLC modes are supported to maximize the overall compression (minimize the total off-chip accesses) for the weights/input activations for all the convolution layers of the AlexNet. RLC Mode 1 is used to compress the weights for all the convolution layers 1–5 and input activations of convolution layer 1 (since the input activations have a lesser number of zeros), whereas RLC mode 2 is used to compress input activation for the remaining (2–5) convolution layers of the AlexNet CNN (since the input activations for these layers have a large number of zeros due to the preceding ReLU activation layer). The simulation results of the RLC access savings on the AlexNet CNN over the ImageNet data set are discussed later in Section IV-D2.

IV. Measured Results

A. Measurement Setup and Test-Chip Summary

Fig. 19 shows the die micrograph, test-chip summary, and FPGA-based test measurement setup of the proposed COMPAC CNN engine implemented using commercial 65-nm CMOS technology. The test chip implements the AlexNet CNN to classify 1000-class ImageNet data set and occupies a total core area of 1.74 mm^2. The overall CNN inference and training methodology consists of: 1) training input data using TensorFlow/Keras [29] software framework; 2) feeding trained filter weights and input activation values for the convolution layers 1–5 to the test chip using Xilinx Virtex-7 FPGA VC707 Evaluation board [30] and Xilinx FMC XM105 Debug Card [31]; 3) performing on-chip MAC, averaging, and pooling operations for convolution layers 1–5 of the AlexNet; and 4) FCN and soft-max layers computation in software (TensorFlow). Due to limited on-chip global buffer capacity and the number of filters, the input activations, the weights, and the output activations (pooled MAV value) are not entirely stored on-chip to compute all the MACs for any given convolution layer. The test chip supports MAC computation for 32 filters and a portion of input activation for any given convolution layer. Thus, multiple iterations are performed, each time feeding in weights (for 32 filters) and a portion of input activations from the Xilinx FPGA/FMC board to the test chip using a 32-bit input data bus and storing the output activations from the test chip on the Xilinx FPGA using a 32-bit output data bus. It should be noted that the input activation and weight address generators/controllers are designed in such a way that it optimizes for data accesses for AlexNet CNN, and the off-chip data access address controller supports only AlexNet CNN. These controllers can be made flexible for supporting other CNNS for optimal data accesses.

B. Test-Chip Characterization

The functionality of scaling the time residue on the MDL is validated on the 65-nm test chip. Fig. 20 shows the oscilloscope captured waveforms of the MDL nodes—A, M, and E for the time accumulation and TRS modes. MDL operation is performed while operating at a 400-mV core voltage and a 1-V pad voltage. It can be observed that the state vector on MDL is advanced resulting in transitions during the time accumulation phase, whereas the state of MDL is changed

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 18,2021 at 10:12:23 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

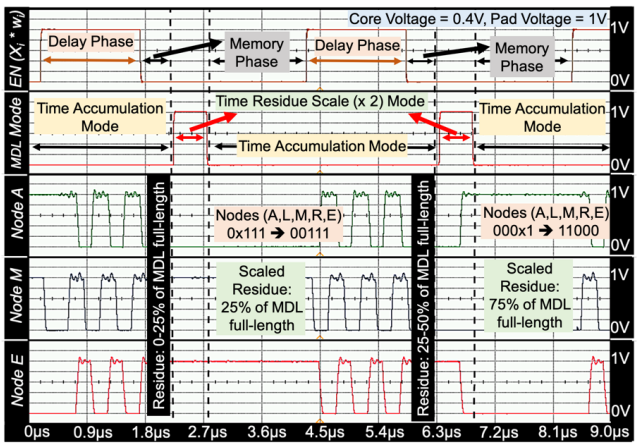


Fig. 20. Experimental demonstration of TRS on bidirectional MDL on 65-nm test chip.

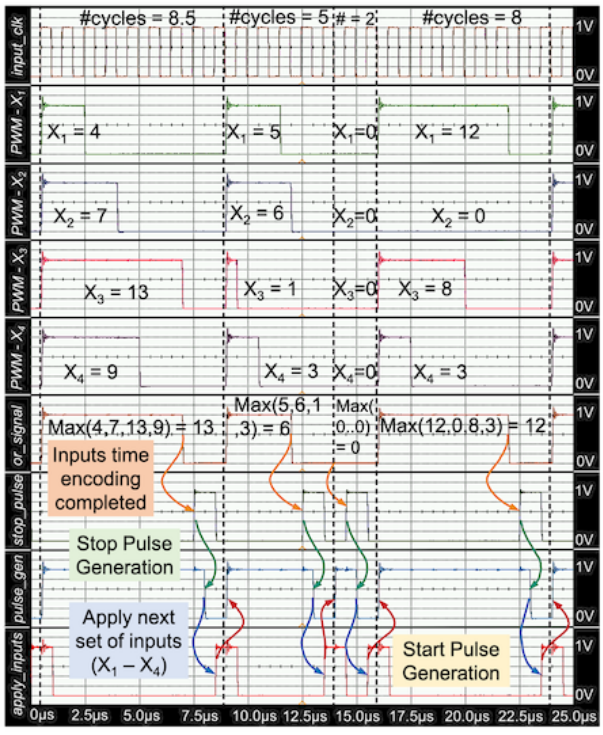


Fig. 21. Experimental demonstration of the CTD approach while operating at 0.5-V core Vcc and 1-V pad Vcc on the 65-nm test chip.

to one of the predefined MDL states (using the TRS lookup table), which has the time residue closest to the actual scaled (×2) residue. For instance, the actual time residue during the two TRS phases lies in the range of 0%–25% and 25%–50% of the MDL full-length delay, which is scaled to 25% and 75% of the MDL full-length delay values, respectively, to scale the time residue, as shown in Fig. 20. These waveforms confirm the successful operation of the TRS approach.

The experimental demonstration of the CTD approach in COMPAC CNN engine on a 65-nm test chip, while operating at a 0.5-V core Vcc, a 1-V pad Vcc, and a 1-MHz input clock

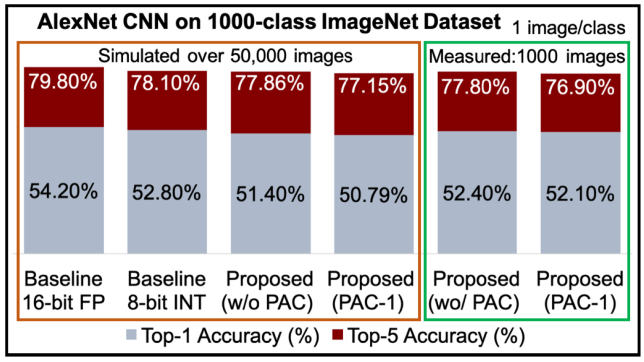


Fig. 22. Measured and simulated classification accuracy of the proposed COMPAC CNN engine for AlexNet CNN on the 1000-class ImageNet data set.

frequency, is shown in Fig. 21. Four 4-bit input activations (X_1–X_4) are encoded in the time domain by selecting the appropriate PWM signal from the pulse generation module, as shown by the oscilloscope captures PWM X_1–X_4. The *stop_pulse* control signal is generated on the next clock edge whenever a negative edge or zero value is observed in the *or_signal* (one of the time-encoded input signals with the maximum pulsewidth). The pulse generation stops on the next subsequent clock edge, and a new set of inputs is applied to the pulse-selection module on the next to next subsequent clock edge by shifting the values from the input activation shift registers using an *apply_inputs* control signal. Finally, the pulse generator starts its operation on the next clock edge, thereby adding a total time overhead of two clock cycles in the CTD approach. These waveforms (see Fig. 21) confirm the successful operation of the proposed concept of the CTD approach to improve the throughput in time encoding the input activations.

C. Accuracy, Throughput, and Energy-Efficiency Results

The test-chip measurements on the AlexNet CNN over ImageNet data set images (randomly chosen 1 image from each class) achieve a top-five classification accuracy of 76.90% (77.80%) and a top-one classification accuracy of 52.10% (52.40%) with (without) the PAC approach (see Fig. 22). In order to quantify COMPAC's CNN engine classification performance on the entire 50 000 ImageNet validation data set images, the non-ideal circuit effects of MDL are incorporated into the COMPAC CNN/Tensorflow simulation framework. The simulation results achieve a top-five classification accuracy of 77.15% (77.86%) and a top-one classification accuracy of 50.79% (51.40%) with (without) the PAC approach (see Fig. 22), showing a classification accuracy loss within 1% compared with the 8-bit fixed-point software implementation.

The proposed COMPAC CNN engine achieves a peak throughput of 0.1278 GOPS on the AlexNet while operating at a 0.72-V Vcc and a 25-MHz frequency and consuming 0.1224 mW of power (measured), thereby making it suitable for an energy-efficient edge computation [see Fig. 19(b)]. The throughput is calculated as follows. There are 32 filters on the test chip, each comprising of four MDLs. Thus,

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 18,2021 at 10:12:23 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE V
ON-/OFF-CHIP ACCESSES FOR THE PROPOSED COMPAC CNN ENGINE WITHOUT AND WITH PAC MODES FOR ALEXNET CNN

Layer	On-chip SRAM Accesses (MB)												Off-chip DRAM Accesses (MB)							
	Without PAC				PAC Mode 1				PAC Mode 2				Without RLC				With RLC			
	Input Act.	Wei-ghts	Output Act.	Total	Input Act.	Wei-ghts	Output Act.	Total	Input Act.	Wei-ghts	Output Act.	Total	Input Act.	Wei-ghts	Output Act.	Total	Input Act.	Wei-ghts	Output Act.	Total
Conv1	2.67	0.05	0.07	2.79	5.34	0.12	0.07	5.53	2.67	0.12	0.07	2.86	0.66	0.05	0.07	0.78	0.62	0.04	0.07	0.73
Conv2	3.62	0.39	0.05	4.06	7.24	0.88	0.05	8.17	3.62	0.88	0.05	4.55	0.98	0.39	0.05	1.42	0.63	0.33	0.05	1.00
Conv3	2.53	0.84	0.11	3.48	2.53	0.84	0.11	3.48	2.53	0.84	0.11	3.48	1.48	0.84	0.11	2.43	0.97	0.76	0.11	1.84
Conv4	1.27	0.63	0.07	1.97	1.27	0.63	0.07	1.97	1.27	0.63	0.07	1.97	0.74	0.63	0.07	1.44	0.47	0.60	0.07	1.14
Conv5	0.85	0.42	0.01	1.28	1.69	0.95	0.01	2.65	0.85	0.95	0.01	1.81	0.49	0.42	0.01	0.92	0.28	0.41	0.01	0.70
Total	**10.94**	**2.33**	**0.31**	**13.58**	**18.07**	**3.42**	**0.31**	**21.80**	**10.94**	**3.42**	**0.31**	**14.67**	**4.35**	**2.33**	**0.31**	**6.99**	**2.97**	**2.14**	**0.31**	**5.41**

128 MDLs are operational at a time, each computing MAC operation. Since each MAC operation is considered as two operations (MAC), the total number of operations computed in parallel is $128 \times 2 = 256$ operations. In the CTD approach (case 2), as discussed in Section III-C, the total time taken to compute a 8-bit input and a weight MAC operation is 7.15 cycles (to time-encode an 8-bit input) $\times 7$ (sequential accumulation with each weight bit, excluding the sign bit) $= 50.05$ cycles. Thus, 256 operations are computed in 50.05 clock cycles. The total latency overhead due to on-/off-chip accesses is 0.007 cycles/operation (calculated by dividing the total on-/off-chip accesses by a product of data bus width and the total number of MACs). This is calculated as follows. Without RLC, the off-chip DRAM access is 6.99 MB to compute 666M MAC operations for AlexNet (see Table V). The data bus is 32-bit wide. This translates to $(6.99 \times 1024 \times 1024 \times 8)/(32 \times 666 \times 10^6) = 0.00275$ cycles per MAC. The weights and input activations are accessed on-chip for each layer in parallel; hence, the latency overhead is the max (cycles took to store inputs in registers and cycles took to store weights in registers) = cycles took to store inputs in registers (since input access > weight access). Thus, the total clock cycles consumed to access the input activations for 666 million MAC operations in AlexNet CNN are 10.94-MB/32-bit wide bus (see Table V), translating to $(10.94 \times 1024 \times 1024 \times 8)/(32 \times 666 \times 10^6) = 0.00431$ cycles/MAC. Adding the on-chip and off-chip access overhead, the latency overhead in accesses is $(0.00275 + 0.00431) = 0.007$ cycles per MAC operation.

Thus, the measured throughput of the proposed COMPAC CNN engine is 256 operations/$(50.05 + 0.007$ cycles $\times$ 40-ns clock period) $=$ 0.1278 GOPS. The measured peak energy efficiency is then calculated by dividing the throughput by the measured on-chip (excluding the leakage and off-chip DRAM access power), as 0.1278 GOPS/0.1224 mW $=$ 1.044 TOPS/W. It should be noted that the operating frequency of 25 MHz is limited by the measurement test-infrastructure and not fundamentally by the MDL-based time-domain approach. Based on the post place and route simulations, the proposed COMPAC design is functional up to 800 MHz (potentially $32\times$ higher throughput).

D. On-/Off-Chip Data Accesses for AlexNet CNN

1) On-Chip Data Accesses: Table V summarizes the on-chip data accesses of the proposed COMPAC CNN engine for input activations, weights, and output activations, over all the five convolution layers of the AlexNet CNN with and without PAC. The COMPAC data flow optimizes for input activation/weight reuse and eliminates the partial sum data movement. COMPAC data flow results in 13.58 MB of on-chip data accesses to compute 666 million MAC operations for the convolution layers 1–5 of the AlexNet CNN, leading to on-chip access of 0.0214 bytes per 8-bit input/weight MAC operation. As discussed earlier, PAC modes 1 (2) lead to an on-chip access overhead of 60.53% (8.03%), thereby leading to 21.80 MB (14.67 MB) of on-chip data accesses. PAC mode 1 incurs higher overhead since both the weights and inputs are accessed again, whereas only weights are accessed again in PAC mode 2, leading to a lower on-chip access overhead, compared with the baseline (wo/PAC) approach.

2) Off-Chip Data Accesses: Table V summarizes the off-chip data accesses of the proposed COMPAC CNN engine for the input–output activations and weights over all the five convolution layers of the AlexNet CNN, with and without the RLC technique. The proposed COMPAC data flow supports configurable SRAM banks with 67-KB capacity, optimal mapping methodology to store the input activations and weights for any given convolution stride on the SRAM bank, and RLC techniques, to optimize for the off-chip data accesses. The proposed COMPAC data flow results in 6.99 MB (5.41 MB) of off-chip data accesses to compute 666 million MAC operations for the convolution layers 1–5 of the AlexNet CNN, leading to off-chip access of 0.011 (0.0085) bytes per 8-bit input/weight MAC operation without (with) RLC technique, respectively.

E. Comparison With Prior Energy-Efficient CNN Accelerators

Table VI compares the proposed COMPAC CNN engine with prior energy-efficient test chips [4], [14]–[18], implementing the AlexNet CNN. The proposed COMPAC CNN engine occupies the least area despite implementing it at a relatively old CMOS technology node (65 nm) compared with other test chips. The proposed COMPAC CNN supports an on-chip global buffer of the least capacity and achieves an energy efficiency of 1.044 TOPS/W (comparable and better) compared with the other test chips at the iso-technology node.

The proposed COMPAC CNN engine results in 86.97% reduced on-chip accesses (13.58 MB instead of 104.25 MB)

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 18,2021 at 10:12:23 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE VI

COMPARISON OF THE PROPOSED COMPAC CNN ENGINE WITH PRIOR TEST CHIPS IMPLEMENTING ALEXNET CNN

Reference	Tech. (nm)	Core Area (mm^2)	On-chip SRAM (KB)	Data Width (#bits)	Top-5 Acc. (%)	Energy Efficiency (TOPS/W)
EYERISS [4]	65	12.25	181.5	16	N/A	0.246
ENVISION [14]	28	1.87	144	1-16	N/A	0.80 - 3.80
STICKER [15]	65	7.80	170	8	N/A	1.038
SNAP [16]	16	2.40	280.6	16	N/A	3.86
UNPU [17]	65	16.00	256	8	N/A	4.30
QUEST [18]	40	121.5	7864.3	4	76.70	0.877
This work	**65**	**1.74**	**67**	**8**	**76.90**	**1.044**

TABLE VII

COMPARISON OF ON-/OFF-CHIP ACCESSES FOR THE PROPOSED COMPAC CNN ENGINE WITH EYERISS [4] AND UNPU [17] APPROACHES

Layer	On-chip Access Comparison (MB)				Off-chip Access Comparison (MB)			
	Eyeriss (16-bit)	Proposed CNN (8-bit)			Eyeriss (w/ RLC)	UNPU	Proposed (wo/ RLC)	Proposed (w/ RLC)
		wo/ PAC	PAC Mode 1	PAC Mode 2	181.5 KB 16-bit	256 KB 8-bit	67 KB 8-bit	67 KB 8-bit
Conv1	18.50	2.79	5.53	2.86	5.00	0.77	0.78	0.73
Conv2	77.60	4.06	8.17	4.55	4.00	0.89	1.42	1.00
Conv3	50.20	3.48 (No pooling layer)			3.00	1.59	2.43	1.84
Conv4	37.40	1.97 (No pooling layer)			2.10	0.97	1.44	1.14
Conv5	24.90	1.28	2.65	1.81	1.30	0.64	0.92	0.70
Total	**208.50**	**13.58**	**21.80**	**14.67**	**15.40**	**4.86**	**6.99**	**5.41**

TABLE VIII

IMPACT OF THE KEY IDEAS OF THE PROPOSED COMPAC CNN ENGINE ON PERFORMANCE, ENERGY EFFICIENCY, AND ACCURACY METRICS

Idea	Through -put	Power	Energy Efficiency	Top-5 Accuracy	Area Overhead	Access Energy
TRS	No effect (<0.1%)	No effect (<0.1%)	No effect (<0.1%)	Increase (3%)	Increase 0.71%	No effect
CTD	Increase (94.46%)	Increase ~3%	Increase (>90%)	No effect	Increase 0.016%	No effect
PAC	No effect	Decrease 22% MDL	Increase 22% comp.	Decrease <1%	No effect	Increase 8%on-chip
EEDF	No effect	Increase ~3%	Increase Access Energy Eff.	No effect	Increase 1.44%	Decrease 87(30)% on(off)chip

compared with an Eyeriss approach [4], considering iso-8-bit precision (see Table VII). The proposed COMPAC CNN engine with PAC also results in significant on-chip access savings: 79.10% and 85.93% in PAC modes 1 and 2, respectively (see Table VII); thereby making the proposed COMPAC CNN engine much more energy efficient in terms of on-chip data access energy compared with Eyeriss [4].

The proposed COMPAC CNN engine results in 29.74% reduced off-chip accesses (5.41 MB instead of 7.70 MB) compared with an Eyeriss approach [4], with lesser on-chip SRAM buffer capacity (67 kB instead of 90.75 kB) and considering iso-8-bit precision (see Table VII). The proposed COMPAC CNN engine compared with UNPU design [17] has 11.32% increased off-chip accesses (5.41 MB instead of 4.86 MB) at iso-8-bit precision. However, the on-chip SRAM capacity of the UNPU approach is approximately four times that of the proposed CNN engine (256 instead of 67 kB), which resulted in reduced off-chip accesses. Thus, the proposed COMPAC CNN engine is much more energy-efficient in terms of off-chip data access energy compared with the Eyeriss [4] and UNPU [17] design approaches under iso-bit width and iso/smaller on-chip SRAM capacity conditions. However, it should be noted that the proposed COMPAC CNN engine design has been optimized for AlexNet, whereas Eyeriss [4] and UNPU [17] designs are more flexible in terms of supporting various CNNs (and even RNNs for UNPU).

V. CONCLUSION

In this article, a COMPAC CNN engine for energy-efficient edge AI computing is demonstrated in 65-nm CMOS technology. Four major ideas are proposed in this work, which results in better performance, accuracy, and overall energy-efficiency (see Table VIII). The proposed COMPAC CNN engine supports TRS in the MDL to perform an energy-efficient multi-bit input and weight MAC operation in the time domain while still achieving high classification accuracy. A CTD approach is proposed and deployed in COMPAC to improve throughput in time encoding of the input activation. The simulation results on the AlexNet CNN over 1000 ImageNet images show a significant throughput improvement, consuming on, an average, 14.71 and 7.15 input clock cycles to time-encode an 8-bit input activation in two different CTD modes (see Fig. 9). The PAC technique is proposed to reduce the number of MACs. The simulation results on the AlexNet CNN over 1000 ImageNet images show up to 31.47% (21.79%) reduction in the number of non-zero input activations MACs with a top-five classification accuracy loss of 0.60% (0.90%), top-one classification accuracy loss of 2.20% (1.90%), and an on-chip access overhead of 60.53% (8.03%) for the PAC modes 1 (2), respectively (see Fig. 12 and Table III). The tradeoffs between classification accuracy, reduction in the number of MACs, and on-chip access need to be carefully comprehended to achieve optimal overall energy savings. An EEDF for optimal on-/off-chip memory accesses is also discussed in this article. COMPAC data flow results in 86.97% reduced on-chip SRAM accesses and 29.74% reduced off-chip DRAM accesses compared with an Eyeriss [4], considering iso-8-bit precision. The 65-nm CMOS test chip implementing the AlexNet CNN achieved an energy efficiency of 1.044 TOPS/W and a throughput of 0.1278 GOPS at 720 mV while operating at 25 MHz. The top-five classification accuracy of 76.90% measured over 1000 ImageNet images and 77.15% by simulating over 50 000 ImageNet images is achieved with the proposed COMPAC approach. The simulation results taken into account MDL circuit non-idealities over 50 000 ImageNet validation set images show a classification accuracy loss within 1% compared with the 8-bit fixed-point software implementation.

REFERENCES

[1] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[2] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1701–1708.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 18,2021 at 10:12:23 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[4] Y. H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," in *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, Jan. 2017.

[5] B. Moons and M. Verhelst, "An energy-efficient precision-scalable ConvNet processor in 40-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 52, no. 4, pp. 903–914, Apr. 2017.

[6] T.-J. Yang, Y.-H. Chen, and V. Sze, "Designing energy-efficient convolutional neural networks using energy-aware pruning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5687–5695.

[7] M. Horowitz, "Computing's energy problem (and what we can do about it)," *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2014, pp. 10–14.

[8] A. Sayal, S. Fathima, S. S. T. Nibhanupudi, and J. P. Kulkarni, "All-digital time-domain CNN engine using bidirectional memory delay lines for energy-efficient edge computing," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2019, pp. 228–230.

[9] A. Sayal, S. S. T. Nibhanupudi, S. Fathima, and J. P. Kulkarni, "A 12.08-TOPS/W all-digital time-domain CNN engine using bi-directional memory delay lines for energy efficient edge computing," *IEEE J. Solid-State Circuits*, vol. 55, no. 1, pp. 60–75, Jan. 2020.

[10] D. Stursa and P. Dolezel, "Comparison of ReLU and linear saturated activation functions in neural network for universal approximation," in *Proc. 22nd Int. Conf. Process Control (PC19)*, Štrbské Pleso, Slovakia, 2019, pp. 146–151.

[11] J. Zhou, W. Xu, and R. Chellali, "Analysing the effects of pooling combinations on invariance to position and deformation in convolutional neural networks," in *Proc. IEEE Int. Conf. Cyborg Bionic Syst. (CBS)*, Beijing, China, Oct. 2017, pp. 226–230.

[12] C. Lee, P. Gallagher, and Z. Tu, "Generalizing pooling functions in CNNs: Mixed, gated, and tree," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 863–875, Apr. 2018.

[13] J. Sim *et al.*, "A 1.42 TOPS/W deep convolutional neural network recognition processor for intelligent IOE systems," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Jan./Feb. 2016, pp. 264–265.

[14] B. Moons, R. Uytterhoeven, W. Dehaene, and M. Verhelst, "14.5 envision: A 0.26-to-10TOPS/W subword-parallel dynamic-voltage-accuracy-frequency-scalable convolutional neural network processor in 28nm FDSOI," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2017, pp. 246–247.

[15] Z. Yuan *et al.*, "Sticker: A 0.41-62.1 TOPS/W 8Bit neural network processor with multi-sparsity compatible convolution arrays and online tuning acceleration for fully connected layers," in *Proc. IEEE Symp. VLSI Circuits*, Honolulu, HI, USA, Jun. 2018, pp. 33–34.

[16] J. Zhang, C. Lee, C. Liu, Y. S. Shao, S. W. Keckler, and Z. Zhang, "SNAP: A 1.67–21.55 TOPS/W sparse neural acceleration processor for unstructured sparse deep neural network inference in 16nm CMOS," in *Proc. Symp. VLSI Circuits*, Kyoto, Japan, Jun. 2019, pp. C306–C307.

[17] J. Lee, C. Kim, S. Kang, D. Shin, S. Kim, and H.-J. Yoo, "UNPU: An energy-efficient deep neural network accelerator with fully variable weight bit precision," *IEEE J. Solid-State Circuits*, vol. 54, no. 1, pp. 173–185, Jan. 2019.

[18] K. Ueyoshi *et al.*, "QUEST: multi-purpose log-quantized DNN inference engine stacked on 96-MB 3-D SRAM using inductive coupling technology in 40-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 54, no. 1, pp. 186–196, Jan. 2019.

[19] S. K. Gonugondla, M. Kang, and N. Shanbhag, "A 42pJ/decision 3.12TOPS/W robust in-memory machine learning classifier with on-chip training," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2018, pp. 490–492.

[20] J. Zhang, Z. Wang, and N. Verma, "In-memory computation of a machine-learning classifier in a standard 6T SRAM array," *IEEE J. Solid-State Circuits*, vol. 52, no. 4, pp. 915–924, Apr. 2017.

[21] A. Biswas and A. P. Chandrakasan, "Conv-RAM: An energy-efficient SRAM with embedded convolution computation for low-power CNN-based machine learning applications," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2018, pp. 488–490.

[22] M. Liu, L. R. Everson, and C. H. Kim, "A scalable time-based integrate-and-fire neuromorphic core with brain-inspired leak and local lateral inhibition capabilities," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Austin, TX, USA, Apr. 2017, pp. 1–4.

[23] A. Amravati, S. B. Nasir, S. Thangadurai, I. Yoon, and A. Raychowdhury, "A 55nm time-domain mixed-signal neuromorphic accelerator with stochastic synapses and embedded reinforcement learning for autonomous micro-robots," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2018, pp. 124–126.

[24] D. Miyashita, S. Kousai, T. Suzuki, and J. Deguchi, "Time-domain neural network: A 48.5 TSOp/s/W neuromorphic chip optimized for deep learning and CMOS technology," in *Proc. IEEE Asian Solid-State Circuits Conf. (A-SSCC)*, Nov. 2016, pp. 25–28.

[25] R. J. D'Angelo and S. R. Sonkusale, "A time-mode translinear principle for nonlinear analog computation," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 62, no. 9, pp. 2187–2195, Sep. 2015.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[27] K. Kim, W. Yu, and S. Cho, "A 9 bit, 1.12 ps resolution 2.5 b/Stage pipelined Time-to-Digital converter in 65 nm CMOS using time-register," *IEEE J. Solid-State Circuits*, vol. 49, no. 4, pp. 1007–1016, Apr. 2014.

[28] M. Kim and J.-S. Seo, "Deep convolutional neural network accelerator featuring conditional computing and low external memory access," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Boston, MA, USA, Mar. 2020, pp. 1–4.

[29] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," 2016, *arXiv:1603.04467*. [Online]. Available: http://arxiv.org/abs/1603.04467

[30] Xilinx. *Xilinx Virtex-7 FPGA VC707 Evaluation Kit*. Accessed: Dec. 2, 2020. [Online]. Available: https://www.xilinx.com/products/boards-and-kits/ek-v7-vc707-g.html

[31] *Xilinx FMC XM105 Debug Card*. Accessed: Dec. 2, 2020. [Online]. Available: https://www.xilinx.com/products/boards-and-kits/hw-fmc-xm105-g.html



Aseem Sayal (Member, IEEE) received the bachelor's degree in electrical and electronics engineering from Delhi Technological University (formerly Delhi College of Engineering), Delhi, India, in 2013, and the M.S. and Ph.D. degrees in electrical and computer engineering from The University of Texas at Austin, Austin, TX, USA, in 2017 and 2020, respectively.

From 2013 to 2015, he worked as a Physical Design CAD Engineer at Qualcomm India Pvt. Ltd., India. He was also an intern at Apple Inc., Cupertino, CA, USA, and Google LLC, Sunnyvale, CA, USA, where he was involved in developing clock tree simulation methodology and average power estimation solutions for long-running arbitrary workloads, respectively. He is currently working as an ML Design Engineer with the Platforms Datacenter Chip Team, Google LLC. He has filed five patents and published 15 articles in referred journals and conferences. His graduate research was focused on designing energy-efficient hardware accelerators for machine learning applications and developing EDA design methodologies for heterogeneously integrated secured ASICs.

Dr. Sayal was a recipient of the Chancellor Gold Medal and Meritorious Student Award from Delhi Technological University in 2013.



Shirin Fathima received the bachelor's degree in electronics and communication engineering from the Birla Institute of Technology and Science, Pilani (BITS, Pilani), Pilani, India, in 2017, and the M.S. degree in electrical and computer engineering from The University of Texas at Austin, Austin, TX, USA, in 2019.

She worked on designing energy-efficient hardware accelerators for image classification in her master's thesis project. Currently, she is working at Apple Inc., Cupertino, CA, USA, as an SoC Physical Design Timing Engineer.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 18,2021 at 10:12:23 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



SS Teja Nibhanupudi (Student Member, IEEE) received the bachelor's degree in electrical engineering from IIT Gandhinagar, Gandhinagar, India, in 2016. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX, USA.

From 2016 to 2017, he worked as a Project Associate at IIT Gandhinagar, where he was involved in the development of high-voltage process flow for Semi-Conductor Laboratory, Ajitgarh, India. His research is focused on circuit-device co-optimization using novel materials, such as RRAM, IMT, and STTRAM.



Jaydeep P. Kulkarni (Senior Member, IEEE) received the B.E. degree from the University of Pune, Pune, India, in 2002, the M.Tech. degree from the Indian Institute of Science (IISc), Bengaluru, India, in 2004, and the Ph.D. degree from Purdue University, West Lafayette, IN, USA, in 2009, all in electronics/electrical engineering.

From 2009 to 2017, he was with Intel Circuit Research Lab, Hillsboro, OR, USA, where he worked on energy-efficient integrated circuit technologies. He is currently an Assistant Professor in electrical and computer engineering at The University of Texas at Austin, Austin, TX, USA, and a fellow of the AMD Endowed Chair in Computer Engineering and a fellow of the Silicon Labs Chair in Electrical Engineering. He has filed 35 patents and published 85 articles in referred journals and conferences. His research is focused on machine learning hardware accelerators, in-memory computing, emerging nano-devices, hardware security, heterogeneous/3-D integration, and cryogenic computing.

Dr. Kulkarni is a member of the ACM. He received the Best M.Tech. Student Award from IISc, the Intel Foundation Ph.D. Fellowship Award, the SRC Best Paper and Inventor Recognition Awards, the Purdue Outstanding Doctoral Dissertation Award, seven Intel Divisional Recognition Awards, the 2015 IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS Best Paper Award, the SRC Outstanding Industrial Liaison Award, Micron foundation faculty awards, and the Intel Rising Star Faculty Award. He has served as the Conference General Co-Chair for 2018 ISLPED, and is currently participating in the technical program committees of CICC, ICCAD, DAC, and AICAS conferences. He is currently serving as the Chair for the IEEE Central Texas SSCS/CAS Joint Chapter. He is also serving as an Associate Editor for the IEEE SOLID STATE CIRCUIT LETTERS and the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 18,2021 at 10:12:23 UTC from IEEE Xplore. Restrictions apply.

Context- and Sequence-Aware Convolutional Recurrent Encoder for Neural Machine Translation

Ritam Mallick
Delhi Technological University
Rohini, Delhi-42
India
ritammallick.mallick9@gmail.com

Seba Susan
Delhi Technological University
Rohini, Delhi-42
India
seba_406@yahoo.in

Vaibhaw Agrawal
Delhi Technological University
Rohini, Delhi-42
India
vaibhaw2731@yahoo.com

Rizul Garg
Delhi Technological University
Rohini, Delhi-42
India
gargrizul@gmail.com

Prateek Rawal
Delhi Technological University
Rohini, Delhi-42
India
prateekrawal11@gmail.com

ABSTRACT

Neural Machine Translation model is a sequence-to-sequence converter based on neural networks. Existing models use recurrent neural networks to construct both the encoder and decoder modules. In alternative research, the recurrent networks were substituted by convolutional neural networks for capturing the syntactic structure in the input sentence and decreasing the processing time. We incorporate the goodness of both approaches by proposing a convolutional-recurrent encoder for capturing the context information as well as the sequential information from the source sentence. Word embedding and position embedding of the source sentence is performed prior to the convolutional encoding layer which is basically a n-gram feature extractor capturing phrase-level context information. The rectified output of the convolutional encoding layer is added to the original embedding vector, and the sum is normalized by layer normalization. The normalized output is given as a sequential input to the recurrent encoding layer that captures the temporal information in the sequence. For the decoder, we use the attention-based recurrent neural network. Translation task on the German-English dataset verifies the efficacy of the proposed approach from the higher BLEU scores achieved as compared to the state of the art.

KEYWORDS

Neural machine translation, convolutional, recurrent, encoder, decoder, context, sequence-to-sequence model

1 INTRODUCTION

Neural Machine Translation (NMT) model is a sequence-to sequence converter that translates a variable length source sentence to a variable length target sentence using only neural networks. The precursor of the current state of the art in NMT can be attributed to Kalchbrenner and Blunsom (2013) [1] and Sutskever *et al.* (2014) [2]; both works condition the probability of each word in the target sentence based on the source sentence representation. Separate recurrent neural networks were used in [2] for processing the source sentence and for predicting the target sentence. The problem with these encoder-decoder models, as they are called [3], is the drop in performance with increase in the length of the source sentence and the presence of unknown words [4]. Bahdanau *et al.* proposed a solution involving a soft-search through sentences to find a suitable segment of the source sentence that can be translated effectively [5]. His approach, called the attention mechanism, is popularly incorporated into current NMT models [8, 17]. An alternative to recurrent neural network is the convolutional neural network (CNN) [1]. Unlike recurrent networks, CNN enables parallelization and faster processing. Encoder-decoder models using CNN were proved effective in translating phrases in the source sentence to suitable target sentences [6, 7]. CNN based NMT models could not, however, match the performance of the state of the art in recurrent neural network based NMT models [3]. Our work integrates convolutional and recurrent layers for constructing the NMT encoder, in order to extract both the context and the temporal information in the source sentence. The organization of this paper is as follows. Section 2 presents the proposed approach. Section 3 discusses the experimental setup section 4 analyzes the results of the translation task. Section 5 summarizes the paper.

2 PROPOSED MODEL

A few approaches, in the past, have combined convolutional and recurrent architectures for NMT. Quasi-RNN (QRNN) proposed by Bradbury *et al.* [9] alternates convolutional layers with the recurrent pooling function that is non-trainable. Its translation performance is lower than the attention-based models which our model outperforms. Another example is the Convolution over Recurrent model (CoveR) [10] in which convolutional layers are added at the output of the RNN encoder to capture the context information. In our work, we propose a novel convolutional-cum-recurrent encoder (shown in Fig.1) to combine the goodness of both approaches. The resultant encoder is both context-aware and sequence-aware. Bahdanau *et al.*'s RNN with attention

mechanism [5] constitutes the decoder module in our NMT system. All layers are fully trainable.

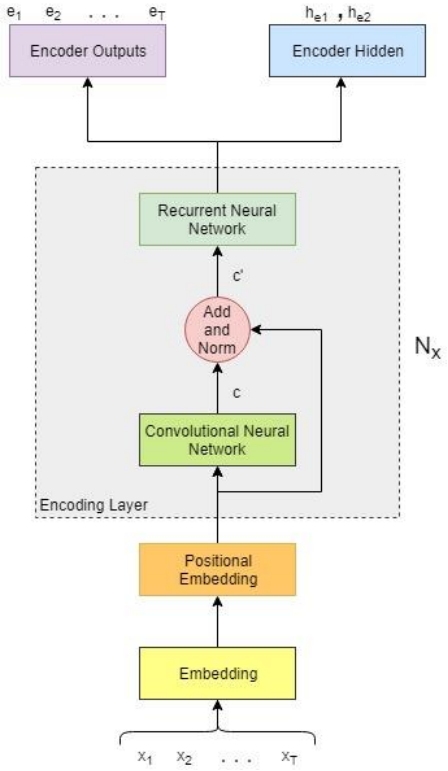


Figure 1: Proposed Convolutional Recurrent encoder model

3.1 Problem definition

Consider that the source sentence of length *T* is denoted by the sequence $\{x_1,...,x_T\}$ and the translated target sequence of length *M* is denoted by $\{y_1,...,y_M\}$. For the sentence pair (**x**, **y**) the task is to learn a probability distribution $p(y_i|x_1,...,x_T,y_{1:i-1})$ that predicts the next target word y_i, given the partial translation $\{y_1,...,y_{i-1}\}$ and the encoded source sentence. The probability of the target sentence **y** given a source sentence **x**, is given by

$$p(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{M} p(y_i|x_1,...,x_T,y_{1:i-1}) \quad (1)$$

3.2 Encoder

A novel NMT encoder architecture is proposed in our paper that comprises of the convolutional neural network in the first subsection followed by the recurrent neural network in the second subsection. All layers are fully trainable. The CNN and the RNN subsections of the encoder are explained next in more detail.

3.2.1 CNN subsection

Akin to the work of Gehring *et al.* on convolutional encoders [6], in a bid to generate position-aware encoding vectors, the word embedding **l** is augmented (summed up) with position embedding **p** [14] to record the position of each token in the source sentence. Let $a_i \in \mathbb{R}^{1\times d}$ be the *d*-dimensional augmented embedding for the i^{th} discrete token (*d*=512), that is given as input to the convolutional network. The advantage of using the convolutional layers first is the n-gram feature extraction that captures context by applying a filter of size *n* on the embedding ensuing from the source sentence **x**. Multiple layers are vertically stacked to capture context from longer sentences, as shown in Fig. 2.

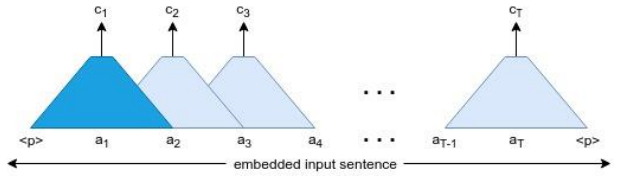


Figure 2: Stacking multiple convolutional layers to capture context

To facilitate efficient learning, we use skip connections [11] from the input of the convolutional layer to its output. Each convolutional layer is followed by a non-linearity (*tan h* function). The rectified convolutional encoded vector *c* is summed up with the word and position embedding through a residual connection, and normalized by layer normalization procedure [12]. The normalized sequential stream of vectors *c'*, corresponding to the input tokens in the source sentence, are fed as input to the recurrent part of the encoder where these are encoded as hidden state vectors that constitute the input to the decoder module.

3.2.2 RNN subsection

Recurrent neural networks are used to encase the temporal sequence of words in the source sentence that are represented by word embeddings, a significant advance over the bag-of-words representation [15]. The temporal pattern of the encoded sequence is learnt by the RNN subsection shown in Fig. 3. The normalized output of the CNN part of the encoder is fed as input to the bidirectional recurrent network which forms the second subsection of our encoder module. There are two RNN encoders, e_1 for processing the input sentence in the forward direction and the encoder e_2 that processes the sentence in the reverse direction. The output of the bidirectional RNN is the concatenation of the forward pass and backward pass outputs at each time-step *t*.

$$e_t = \left[\overrightarrow{e_t} : \overleftarrow{e_t}\right] \quad (2)$$

The encoder output is the annotation obtained for the input source sentence, which is passed as the input to the RNN decoder. The forward pass output is computed as a function of the previous hidden state and the convolutional input.

$$\overrightarrow{e_t} = f\left(c_t', h_{e_1}^{<t-1>}\right) \quad (3)$$

Likewise, the backward pass output is computed.

$$\overleftarrow{e_t} = f\left(c_t', h_{e_2}^{<t+1>}\right) \quad (4)$$

Here, f is a non-linear activation function. We use Gated Recurrent Unit (GRU) [3] for implementing the non-linear recurrent function, which is a simplified version of the Long Short Term Memory (LSTM) [13].

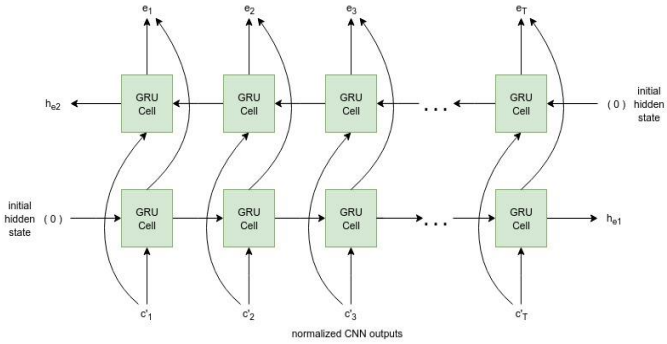


Figure 3: RNN subsection of the encoder

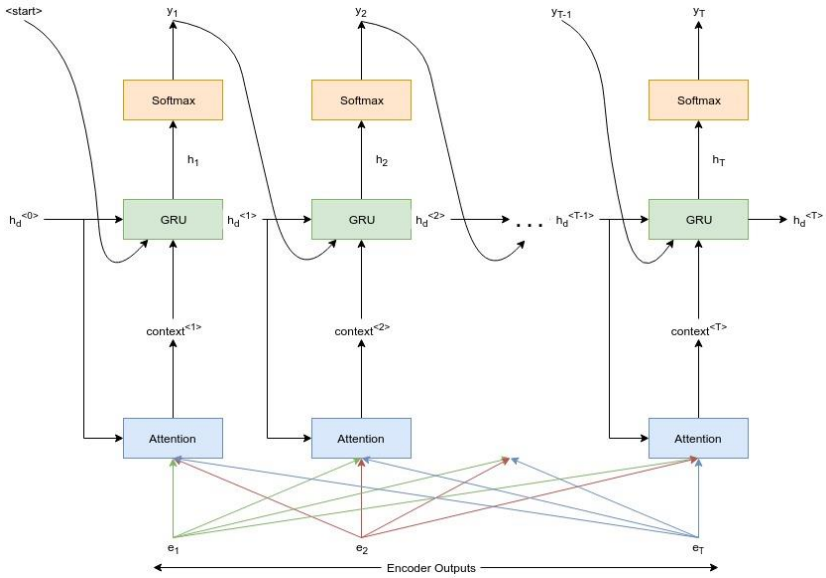


Figure 4: Decoder architecture

3.3 Decoder

Our decoder module (shown in Fig. 4) comprises of a recurrent neural network with attention mechanism [5] that computes a probability distribution for every word in the target sentence. The probability of the output target word is computed as a function of the previously predicted word and the context vector. A soft weight mechanism is used to evaluate the context vector from the encoder output. The context vector at time-step *i*, is defined as the weighted sum of the encoder hidden states computed in (2).

$$context^{<i>} = \sum_{t=1}^{T} \alpha_{it} e_t \quad (5)$$

The soft attention weights are computed over the entire source sentence of length *T* as

$$\alpha_{it} = \frac{e^{score(h_d^{<i-1>}, e_t)}}{\sum_{h=1}^{T} e^{score(h_d^{<i-1>}, e_h)}} \quad (6)$$

$$score\left(h_d^{<i-1>}, e_t\right) = \tanh\left(W_1 h_d^{<i-1>} + b_1 + W_2 e_t + b_2\right) \quad (7)$$

To compute the probability distribution of the target sentence given the source sentence as shown in (1), a word-by-word prediction pattern is followed. The probability of the next target word is computed as a function of the previously predicted word and the context vector as

$$p\left(y_i \middle| x_1, ..., x_T, y_{1:i-1}\right) = soft\max(g(h_j)) \quad (8)$$

where,

$$h_j = f(h_{j-1}, [context^{<j>} : y_{j-1}]) \quad (9)$$

4 RESULTS

Our experiments were conducted on 176,692 sentence pairs of the German-English dataset (*Tatoeba project*) [16]. 163,957 sentence pairs were used for training and 3900 sentence pairs for testing. 5% of the data was used for validation, that served the purpose of tuning of hyper-parameters such as the size of the embedding vector, number of hidden units and number of encoding layers. The batch size was set to 128 sentence pairs. We compare our model, implemented using Python on a Tesla K80 GPU, with that of Kalchbrenner and Blunsom (2013) [1] that used a CNN for the encoder and a RNN for the decoder, that we call as CNN-1. The next work we compare with, is the pure RNN model of Bahdanau *et al*. (2014) [5] that incorporates attention mechanism. We also compare our work with the convolutional NMT model of Gehring *et al*. (2016) [7], referred to as CNN-2. The results are compiled in Table 1. The presence or absence of position embedding and attention mechanism in these models is also indicated. We conducted an empirical study on the variation of BLEU scores when the number of stacked convolutional encoding layers is varied from 1 to 5. The highest BLEU score of 30.6 is obtained when x=3. For the encoder and decoder RNNs, the Adadelta optimizer is used with learning rate 0.1 and threshold 1e-6. We find that our convolutional-cum-recurrent encoder model yields the highest BLEU scores of 30.6 with position embedding and 27.9 without position embedding. In comparison, the purely CNN model of CNN-2 gives the second-best performance.

5 CONCLUSION

In our work, we seek to improve the RNN encoder-decoder NMT model by incorporating encoding by convolutional neural networks in the first part of the encoder to capture the syntactic structure in the source sentence. The CNN-encoded output is embossed with the embedding vector and the result is given as input to the RNN part of the encoder. The RNN hidden states are used to compute the context vector using soft attention, which helps the decoder RNN in translating the target sentence by computing the probability for each target word given the previous one. Our NMT model outperforms existing models for experiments based on the German-English dataset. Adaptation of our model to Indic languages forms the next phase of our project.

Table 1: Comparison of our model with best performing models

Models	Encoder			Decoder		BLEU
	Position embedding	CNN	RNN	Attention	RNN	
CNN-1	No	Yes	No	No	Yes	22.5
RNN	No	No	No	Yes	Yes	26.3
CNN-2	Yes	Yes	No	Yes	Yes	29.2
Ours	Yes	Yes	No	Yes	Yes	30.6

REFERENCES

[1] Kalchbrenner, Nal, and Phil Blunsom. "Recurrent continuous translation models." In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1700-1709. 2013.

[2] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." In *Advances in neural information processing systems*, pp. 3104-3112. 2014.

[3] Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." *arXiv preprint arXiv:1406.1078* (2014).

[4] Cho, Kyunghyun, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. "On the properties of neural machine translation: Encoder-decoder approaches." *arXiv preprint arXiv:1409.1259* (2014).

[5] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." In *3rd International Conference on Learning Representations, ICLR 2015*. 2015.

[6] Gehring, Jonas, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. "Convolutional sequence to sequence learning." *arXiv preprint arXiv:1705.03122* (2017).

[7] Gehring, Jonas, Michael Auli, David Grangier, and Yann N. Dauphin. "A convolutional encoder model for neural machine translation." *arXiv preprint arXiv:1611.02344* (2016).

[8] Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. "Effective approaches to attention-based neural machine translation." *arXiv preprint* arXiv:1508.04025 (2015).

[9] Bradbury, James, Stephen Merity, Caiming Xiong, and Richard Socher. "Quasi-recurrent neural networks." *arXiv preprint arXiv:1611.01576* (2016).

[10] Dakwale, Praveen, and Christof Monz. "Convolutional over recurrent encoder for neural machine translation." *The Prague Bulletin of Mathematical Linguistics* 108, no. 1 (2017): 37-48.

[11] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778. 2016.

[12] Ba, Jimmy Lei, Jamie Ryan Kiros, and Geoffrey E. Hinton. "Layer normalization." *arXiv preprint arXiv:1607.06450* (2016).

[13] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9, no. 8 (1997): 1735-1780.

[14] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." In *Advances in neural information processing systems*, pp. 5998-6008. 2017.

[15] Susan, Seba, and Juli Keshari. "Finding significant keywords for document databases by two-phase Maximum Entropy Partitioning." *Pattern Recognition Letters* 125 (2019): 195-205.

[16] http://manythings.org/anki [Last accessed on 15th Dec 2020]

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/311394834

Cyclone Frame Prediction by Gaussian Mixture Modeling of the Three Penultimate Time-Lapse Frames

Article *in* Journal of the Indian Society of Remote Sensing · December 2016

DOI: 10.1007/s12524-016-0644-8

CITATIONS
0

READS
16

5 authors, including:



Seba Susan
Delhi Technological University
78 PUBLICATIONS **422** CITATIONS

SEE PROFILE



Akshay Takhi
NITIE-National Institute of Industrial Engineering
1 PUBLICATION **0** CITATIONS

SEE PROFILE

All content following this page was uploaded by Seba Susan on 16 November 2020.

The user has requested enhancement of the downloaded file.

This is the initial version of an article submitted to the *Journal of the Indian Society of Remote Sensing*. The final version of the article after revisions which includes details on both the cyclones *Nargis* (2008) and *Hudhud* (2014) that hit the Indian sub-continent, are given at https://link.springer.com/article/10.1007/s12524-016-0644-8. The published paper is available for direct online view at: http://rdcu.be/nodN

Cyclone Frame Prediction by Gaussian Modeling of Most Recent Spatio-Temporal Data

Seba Susan, Achin Saxena, Anuvart Badhwar, Akshay Takhi, Abhishek Varshney
Department of Computer Science & Engineering,
Delhi Technological University, New Delhi, India-110042
Email: seba_406@yahoo.in

***Abstract*— A fast cyclone frame prediction is proposed in this paper that models the most recent spatio-temporal data in the cyclone video using a Gaussian Mixture followed by the fuzzy regression procedure outlined in [1]. Unlike the previous work which models the entire history on a per pixel basis we model the entire 3D space within the last three frames making the process faster and more accurate. The increase in accuracy is attributed to the fact that cyclones evolve over time making the most recent data the only one most reliable to predict the next frame. We also propose a technique of finding the optimal number of components in the Gaussian mixture based on maximum likelihood principle. Our results on the recently occurred Nargis cyclone that hit the Indian seas show a high accuracy of frame prediction.**

Keywords- Fuzzy Gaussian Regression

I. INTRODUCTION

The prediction of future frames of a cyclone video is crucial to timely intervention on the part of the authorities for preventive measures and relief operations. Since the frames of the video captured by the satellite are about 30 secs apart, future frame prediction based on the current video frame/frames finds relevance for guiding rescue operations. Several cyclone frame prediction schemes can be found in literature. In [1] Verma and Pal proposed Gaussian mixture modeling of the entire history of each pixel for each of the color R,G,B channels. A fuzzy regression procedure was used to successfully predict the future frame intensities with the weights of the rules being the prior probabilities of the Gaussian components. An improvisation that involves the learning of cyclone data from past history pixel by pixel using artificial neural networks and dimensionality reduction was proposed in [2]. However the study of extreme storms in [3,4] indicates that the severe cyclone behaviour deepens explosively with no warning whatsoever and a prediction based on the current situation alone would increase the accuracy of cyclone frame prediction. Also the recent 3D data in the entire spatial domain needs to be studied for predicting the cyclone effect at a point. Based on this axiom we propose to model the 3D data in recent frames over the entire spatial domain using a single Gaussian Mixture Model, with the optimal number of mixture components determined for from the maximum likelihood graph. Case studies are abundant in literature on existing cyclone prediction models such as NOGAPS in the Pacific [5], COAMPS-TC of US-Navy [6] and TC-LAPS of Australian Meteorology centre [7]. All studies generally observe that the prediction is heavily dependent on the reference frames. The organization of this paper is as follows: The proposed cyclone frame prediction scheme is described in Section II, the results are discussed in Section III and the conclusions are drawn in Section IV.

This is the initial version of an article submitted to the *Journal of the Indian Society of Remote Sensing*. The final version of the article after revisions which includes details on both the cyclones *Nargis* (2008) and *Hudhud* (2014) that hit the Indian sub-continent, are given at https://link.springer.com/article/10.1007/s12524-016-0644-8. The published paper is available for direct online view at: http://rdcu.be/nodN

II. PROPOSED CYCLONE FRAME PREDICTION USING FUZZY REGREESION

We consider the last three frames overall to predict the next frame. The recent past of each pixel together with its current value comprise its 11 features. For a pixel of the current (and available) frame t denoted by (x,y,t), its past would comprise of (x,y,t-1), (x,y-1,t-1), (x-1,y,t-1), (x,y+1,t-1), (x+1,y,t-1), (x,y,t-2), (x,y-1,t-2), (x-1,y,t-2), (x,y+1,t-2), (x+1,y,t-2) and together the 11 intensities would form the feature vector. The Gaussian mixture modeling of the 11D data samples follow next. A single model is created opposed to the procedure in [1] where a pixel by pixel modeling is done. The Gaussian mixture model probabilities are used as the weights of the fuzzy rule base regression [1] used for the future frame prediction of (x,y,t+1) based on the 10-D input of (x,y,t), (x,y-1,t), (x-1,y,t), (x,y+1,t), (x+1,y,t), (x,y,t-1), (x,y-1,t-1), (x-1,y,t-1), (x,y+1,t-1), (x+1,y,t-1) obtained from past two frames t and t-1.

The algorithm for proposed cyclone frame prediction is given below customized to predict the future frame from the past three frames.

- *For predicting the 193rd frame from the 192nd, 191st and 190th frames, do:*
  - *For each R,G,B, channel do:*

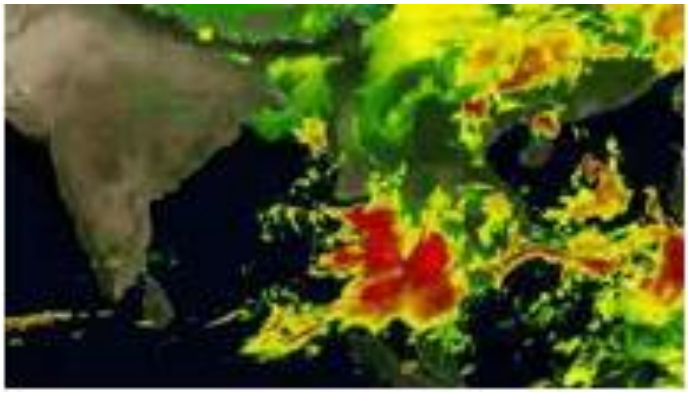
1. Create a eleven-dimensional feature set for each pixel position comprising of the center pixel of the 192nd frame, center pixel and 4 neighboring pixels of the 191st frame and the center pixel and 4 neighboring pixels of the 190th frame.
2. Model the 11-D data from all pixel positions using a Gaussian Mixture Model with the number of components initialized to 2.
3. Plot the log-likelihood of the mixture model as an instantaneous graph.
4. If the log-likelihood has attained a peak that is positioned between two valley points, then the number of mixture components is set to the number of components at the location of the peak, else increase the number of components by one and repeat steps 2 and 3.
5. Based on the mean, standard deviation and probability weights of the components of the Gaussian mixture model that is optimized by Expectation Maximization, we achieve next frame intensity prediction by using fuzzy regression.

- *Combine the RGB values to form a color image and compute the metric values pertaining to error of prediction*

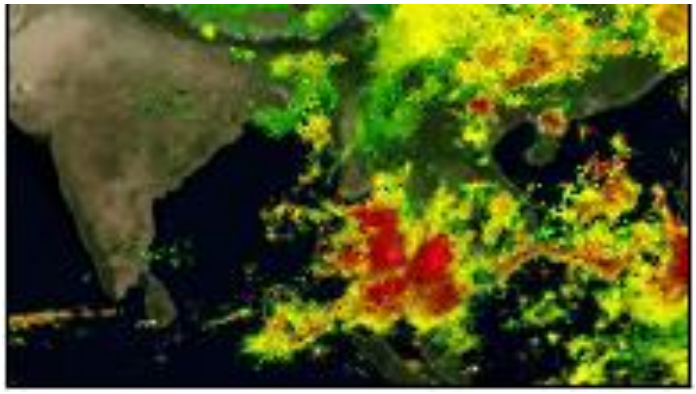
III. EXPERIMENTAL RESULTS AND DISCUSSIONS

The satellite video of *Nargis*, the severe cyclone that hit the shores of Myanmar in 2008 is used for our experimentation. The coding was done in MATLAB 7.10 on a 2.6 GHz PC. The proposed algorithm successfully predicted the 193rd video frame as shown in Fig. 1 as compared to the original fuzzy rule based Gaussian regression [1]. For the results of [1] we set the number of components to four (as per example of a case given in [1]) per mixture per pixel to save computations and time. We have also compared our result for the same setting as well as optimal number of components which is automated procedure without any manual intervention. The saving in computations over [1] is tremendous since the optimality is for a single model while in [1] the optimality is done for each pixel in the image. The prediction accuracy in terms of the CIM index for R,G,B channels are shown in Table 1 and reiterate our claim that the proposed method predicts cyclone frames most accurately from more relevant and recent reference frames. Comparison with the fuzzy rule based Gaussian regression [1] indicates a superior performance for our method.

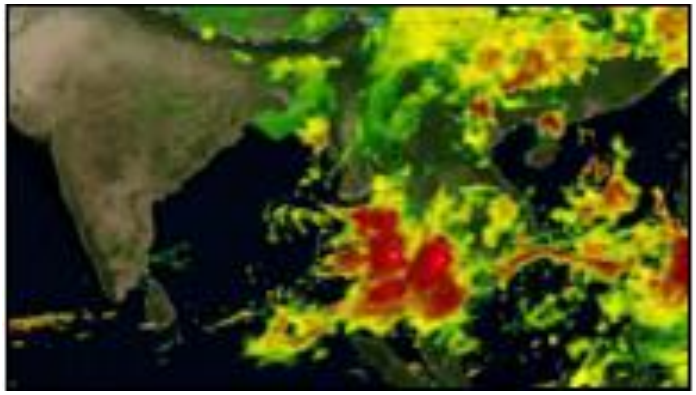
This is the initial version of an article submitted to the *Journal of the Indian Society of Remote Sensing*. The final version of the article after revisions which includes details on both the cyclones *Nargis* (2008) and *Hudhud* (2014) that hit the Indian sub-continent, are given at https://link.springer.com/article/10.1007/s12524-016-0644-8. The published paper is available for direct online view at: http://rdcu.be/nodN



(a)

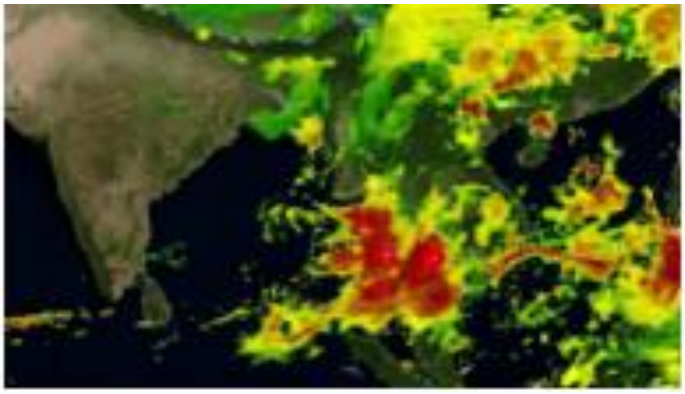


(b)



(c)

This is the initial version of an article submitted to the *Journal of the Indian Society of Remote Sensing*. The final version of the article after revisions which includes details on both the cyclones *Nargis* (2008) and *Hudhud* (2014) that hit the Indian sub-continent, are given at https://link.springer.com/article/10.1007/s12524-016-0644-8. The published paper is available for direct online view at: http://rdcu.be/nodN



(d)

Fig. 1 (a) Original 193rd frame (b) Frame predicted using fuzzy rule based Gaussian regression with four components per pixel (c) Frame predicted using proposed method with four components per pixel (d) Frame predicted using proposed method with optimal number of components

Table 1: Accuracy in prediction for various methods for the 193rd frame using CIM index

Method	CIM-Red	CIM-Green	CIM-Blue
Fuzzy rule based Gaussian regression [1]	0.9806	0.9327	0.7874
Proposed method	0.9984	0.9882	0.7927

IV. CONCLUSION

A fast and accurate cyclone frame prediction is proposed in this paper that uses Gaussian mixture modeling aand fuzzy rule based regression on the most recent spatio-temporal data in the cyclone video. The optimal number of components of the mixture is determined from the first maxima of the log-likelihood of the mixture. Comparison of our results with existing methods on a recent cyclone video that affected the Indian Peninsula affirms the efficiency of our approach.

REFERENCES

[1] Verma, Nishchal K., and N. R. Pal. "Prediction of satellite images using fuzzy rule based Gaussian regression." In *Applied Imagery Pattern Recognition Workshop (AIPR), 2010 IEEE 39th*, pp. 1-8. IEEE, 2010.

[2] Verma, Nishchal K. "Future image frame generation using Artificial Neural Network with selected features." In *Applied Imagery Pattern Recognition Workshop (AIPR), 2012 IEEE*, pp. 1-8. IEEE, 2012.

[3] Economou, Theodoros, David B. Stephenson, and Christopher AT Ferro. "Spatio-temporal modelling of extreme storms." *The Annals of Applied Statistics* 8, no. 4 (2014): 2223-2246.

[4] Edwards, Roger. "Tropical cyclone tornadoes: A review of knowledge in research and prediction." *E-Journal of Severe Storms Meteorology* 7, no. 6 (2012).

[5] Fiorino, Michael, James S. Goerss, Jack J. Jensen, and Edward J. Harrison Jr. *An evaluation of the real-time tropical cyclone forecast skill of the Navy Operational Global Atmospheric Prediction System in the western North Pacific*. No. NRL-W-JA-431-006-92. NAVAL RESEARCH LAB MONTEREY CA, 1993.

[6] Hendricks, Eric A., Jonathan R. Moskaitis, Yi Jin, Richard M. Hodur, James D. Doyle, and Melinda S. Peng. "Prediction and diagnosis of Typhoon Morakot (2009) using the Naval Research Laboratory's mesoscale tropical cyclone model." *Terrestrial, Atmospheric and Oceanic Sciences* 22, no. 6 (2011): 579-594.

[7] Davidson, Noel E., and Harry C. Weber. "The BMRC high-resolution tropical cyclone prediction system: TC-LAPS." *Monthly Weather Review* 128, no. 5 (2000): 1245-1265.

View publication stats

2020 IEEE International Conference for Innovation in Technology (INOCON)
Bengaluru, India. Nov 6-8, 2020

Design and Analysis of a Three-Phase Grid-Connected Solar PV Inverter with Proportional-Integral Controller

Kemoh Lumei
Electrical Engineering Department
Delhi Technological University
New Delhi, India
lumeijr.kemoh@yahoo.com

***Abstract*—In detail, the design and analysis of a three-phase grid-connected PV electrical converter are well discussed in this paper. Inverter provides DC power to AC power and it's helpful in powering electronic and electrical instrumentation rated at the ac mains voltage. Before developing the inverter, a mathematical model for the PV system let alone proportional-integral controller and perturbs and observe MPPT formula were designed to access the varied characteristics of the PV module. Considering the implementation space of this paper finding, a PV system is presently the foremost reliable energy supply to deal with the energy crisis in Liberia. Liberia's PV power offers high potential to stimulate economic progression. The last section of this paper presents the results of the integrated system that enables the readers to grasp the varied conditions that were thought-about throughout this research analysis.**

Keywords— Photovoltaic (PV), Maximum Power Point Tracking (MPPT), Perturb and Observe (P&O), Proportional-Integral (PI), MatLab and Liberia Electricity Corporation (LEC).

I. INTRODUCTION

Currently, solar energy is one of the renewable energy that the Republic of Liberia has been blessed with. Solar power is eco-friendly and really straightforward to be enforced as compared to alternative renewable energy sources. Solar power is often accustomed to meet the direct desires of electricity with the assistance of a PV system and power electronic devices. With the fast utilization of PV array, energy generated is transferred to the grid because of the value and life potency limitations of the energy storage system. The PV array converts the radiation into voltage and also the electrical converter converts the DC energy into AC energy, which is fed to the grid. The technical term for the electrical converter is "grid-connected inverter". The grid connected inverter is connected to the power grid and it is used to converter AC power to DC power.

A. Grid-Connected Inverter

The three-phase grid-connected power converter completes the power supply where the power is converted from DC to AC. To appreciate this, the power generated by the PV system is regenerated from DC to AC by a grid-connected electrical converter. It synchronizes current and frequency according to the national or local grid. The output voltage is booted from the grid so that it is slightly higher than the grid voltage for additional voltage. "The three-phase grid-connected electrical converter is considered exclusively for grid-connected applications that do not require cooperative power storage systems" [5]. The three-phase power converter is commonly used for high power application, although the project aims to encourage individual households to use the three-phase grid-connected electrical converter.

B. Paper Background And Problem Statement

The warfare of 1989 and 2003 severely affected Liberia's social and economic development. The influx of donor partners and international communities notably has not significantly improved the social and economic state of affairs of the Liberian people. However, like many countries in Africa, heaps of other developing countries have a terrible energy sector. Liberia as an African nation presently has a population of 4.8 million. The current energy services are supported by hydroelectric machines and diesel oil production. Even with the efforts of the Liberian government and its development partners, over eighty percent (80%) of the full population does not have access to electricity. Consistent with the United States of America Agency for International Development (USAID), April 16, 2020, Africa power Sheet, said about (12%) of Liberians have access to electricity. The full generating unit capability of the national power grid is one hundred and twenty-six megawatts (126 MW). Thirty-eight megawatts (38 W) of diesel and heavy fuel oil (HFO) units, and hydropower plants is eighty eight megawatts (88 MW). This paper aims to work with the national stakeholders for the design of a "three-phase PV system grid-connected inverter" with the help of the Liberian government to meet the electricity demand.

II. INTEGRATED SYSTEM DESIGN AND CONFIGURATION

A three-phase grid-connected inverted is currently considered as the most reliable means to transmit DC power to an AC grid. An inverter basically converts DC power to AC. Illustrated in figure 2 is a three-phase grid-connected solar PV inverter. The system is basically made up of 50kW solar PV system, boost converter with a Perturb and Observe MPPT model, three-phase inverter and a power grid. The system ac voltage is 400 volts but for inverter safety factor and voltage variation in a practical application we considered twenty percent (20%) of the system voltage to quantify the DC-DC boost converter and inverter. And Figure 1 is the block diagram for a solar PV system connected to the grid. As it is seen, the inverter has a separate control. For the PV array, the DC – DC boost converter is used to step-up the DC voltage that is

978-1-7281-9744-9/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:13:32 UTC from IEEE Xplore. Restrictions apply.

required for the inverter input voltage. This is done with the help of the P&O MPPT algorithm.

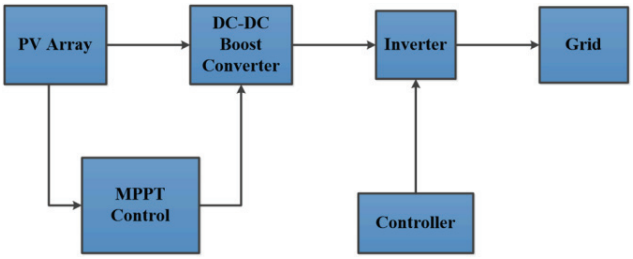


Fig. 1. Solar PV System Connected to Grid Block Diagram

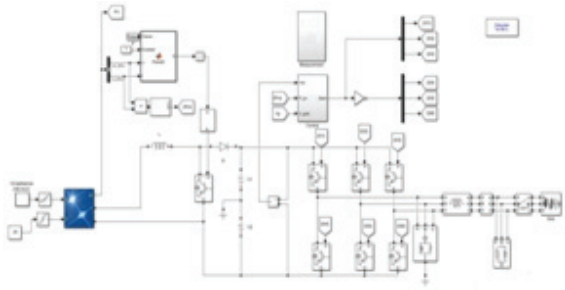


Fig. 2. Three-Phase Grid-Connected Inverter

A. Solar PV System

For this paper, a solar PV array block from the Matlab Simulink 2019 library was used. The name for the simple solar PV array is 1Soltech 1STH-350WH. Table 1 and table 2 clearly provided in detail the technical specification of the PV module that was used for this paper. From table 2 it is illustrated that the module maximum power is 349.59W. Also table 1 shows that the maximum voltage and current are 43V and 8.13A respectively. With reference to table 1 the open circuit voltage and short-circuits current of the PV model is 51.5V and 9.4A respectively.

TABLE I. SOLAR PV SYSTEM SPECIFICATION

Number of Modules Per String	***5***
Number of String Per Array	***29***
Number of Array	***1***
String Voltage	***215V***
Array Power	***50kW***
Number of Boost Converter	***1***

TABLE II. SOLAR PV MODULE SPECIFICATION (1SOLTECH 1STH-350WH)

Maximum Power	***349.59W***
Open Circuit Voltage (V_{OC})	***51.5V***
Voltage at Maximum Power Point (V_{MPP})	***43V***
Temperature Coefficient of VOC (% deg C)	***-0.36***
Cell Per Module (N Cell)	***80***
Short-Circuit Current (I_{SC})	***9.4A***
Current at Maximum Power Point (I_{MPP})	***8.13A***
Temperature Coefficient of I_{SC}	***0.09***
Light-Generated I_L	***9.4447A***
Diode Saturation Current I_O	***$3.2328X10^{-10}A$***
Diode Ideality Factor	***1.04***
Shunt Resistance RSH	***47.9694ohms***
Series Resistance RS	***0.22828ohms***

B. DC-DC Boost Converter

A DC-DC boost converter comprises of inductor, capacitor and controllable semiconductor switch. Table 3 is the boost converter specification. From table 3, it is clearly written that the inductor is been rated at 0.4352mH and the capacitor has a rating of 1.282mF. The controllable semiconductor switch has a switching frequency of 4000Hz. For this system the boost converter steps up the output voltage of the Solar PV system from 215V to 870.9552V. Therefore, the output voltage of the inverter is 870.9552V. It is a good estimation for the inductor ripple current to be 20% to 40% of the output current and the capacitor voltage to be 1% of the output voltage in this case were the boosting difference is large.

TABLE III. DC-DC BOOST CONVERTER SPECIFICATION

Number of DC-DC Boost Converter	***1***
Converter Efficiency	***0.9***
Duty Cycle	***0.778***
Switching Frequency	***4000Hz***
Ripple Current	***93.0232A***
Power	***50kW***
Input Voltage	***215V***
Output Voltage	***870.9552V***

C. Grid-Connected Inverter

The grid-connected inverter was modeled in Matlab Simulink 2019. It has six IGBTs (insulated-gate bipolar transistor). The internal resistance is 1mohms and the snubber resistance is 100kohms. Table 4 is the grid-connected inverter technical specification. Table 4 provides information like grid voltage, modulation index, inverter switching frequency and inverter rated power. The input voltage of the inverter is the same as the output voltage of the boost converter. From the simulated results, the inverter has an input voltage of 870.9552V and an output voltage of 400v. Also the system has an output ac current that is about 135ampere. During the modeling stage, the inverter was connected to a grid that has wye to ground configuration. The grid frequency is 50Hz and the phase to phase voltage (vrms) is 500v.

TABLE IV. INVERTER SPECIFICATION

Number of DC-DC Boost Converter	***1***
Modulating Index	***0.9***
Switching Frequency	***4000Hz***
Frequency Modulation	***84***
Power	***50kW***
Output Voltage	***400V***
Input Voltage	***870.9552V***

D. Proportional-Integral Controller Analysis

The controller controls and calculates the output signal for each sample time to control the PI controller. There are only two turning parameters for the PI controller. For the PI controller, the sweeping action allows to remove the offset. By removing offset components through PI controllers, it has provided PI controllers with a balance of complexity and capabilities that are widely used by engineers in processing control applications.

$$C_O = C_{O(bias)} + K_C * e(t) + \frac{K_C}{\tau_i}\int e(t)dt \quad (1)$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:13:32 UTC from IEEE Xplore. Restrictions apply.

$$e(t) = SP - PV \quad (2)$$

From the above equation, we defined the following variables. C_O: Controller Output Signal, $C_{o(bias)}$: Controller Bias, e(t): Current Control Error, SP: Set Point, PV: Process Variable Measure, K_C: Grain Controller, a turning parameter and τ_i: Reset Time, a turning parameter.

If we have a three-phase inverter connected to the load grid, it will be connected to the output. The voltage on the output side is considered to be very sensitive and the error signal is the output of the output reference signal and the comparison of the load. The proportional element reduces the error, while the integral element reduces the offset. The proportional element depends on the current error. It is a comprehensive PI controller that improves tracking by minimizing the instantaneous error between the reference and the actual value.

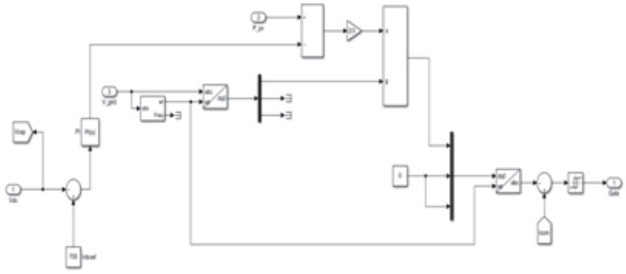


Fig. 3. PI Controller Matlab Model

III. SIMULATION RESULTS

A. Constant Irradiance And Constant Temperature

For the first condition, the irradiance was kept at 1000 W/m^2 while the temperature was kept at 25^o C. The power did not change significantly. And also figure 4 also illustrates that during the constant irradiance and temperature, our controller is very accurate as it is seen in the power, voltage and current graph.

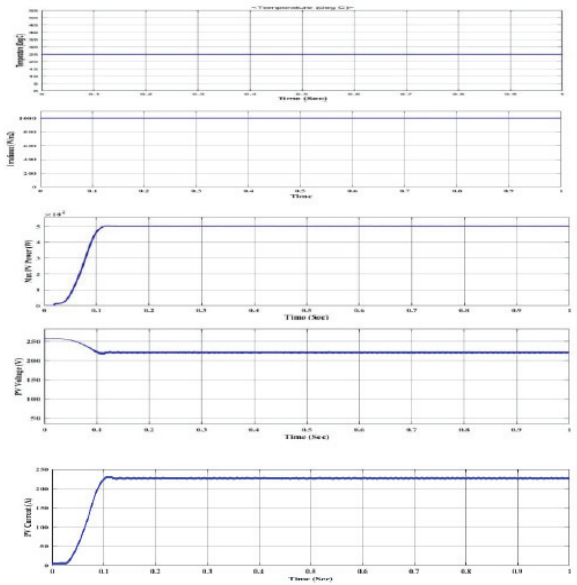


Fig. 4. Constant Irradiance and Temperature

Figure 5 is the constant temperature and irradiance grid power, voltage and current. For this condition the input power is equal to the output power. We didn't experience any form of power variation. With the help of the harmonics filter, our system grid side has a very good power quality. The voltage and current output are pure sinusoidal.

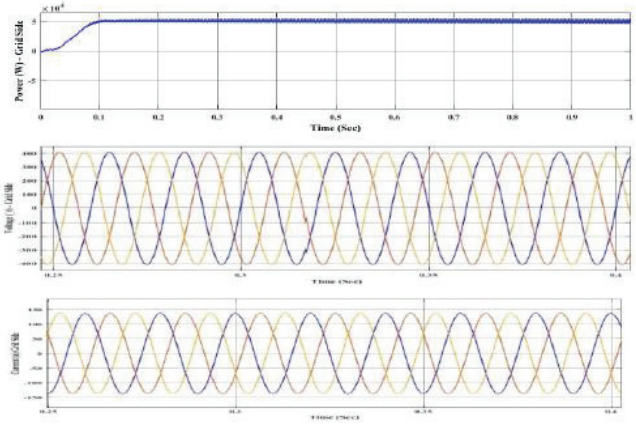


Fig. 5. Constant Irradiance and Temperature Grid Parameters

B. Constant Irradiance and Varying Temperature

As it can be been seen from the figure 5.9, the irradiance was constant while the varying temperature varied from 30^OC, 40^OC to 50^OC. At 0.5 seconds there is a change in power, voltage and current. From figure 6 it is understood that the temperature does not affect the PV parameters and the PI controller has an accurate response with regard to the changing temperature.

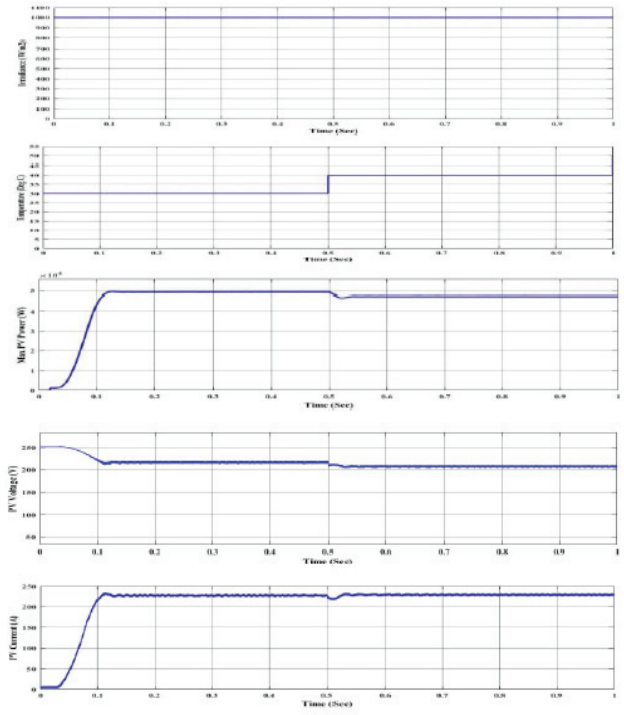


Fig. 6. Constant Irradiance and Varying Temperature

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:13:32 UTC from IEEE Xplore. Restrictions apply.

Figure 7 is the constant temperature and varying irradiance grid power, voltage and current. For this condition the input power is equal to the output power. We experience little form of power variation. With the help of the harmonics filter, our system grid side has a very good power quality. The voltage and current output are also pure sinusoidal for this condition.

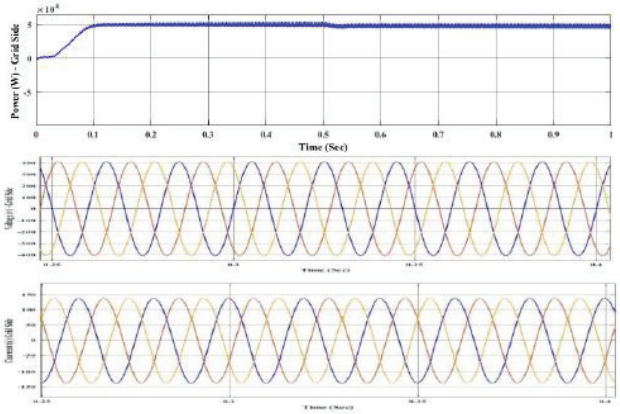


Fig. 7. Constant Irradiance and Varying Temperature Grid Parameters

C. Varying Irradiance and Constant Temperature

Figure 8 illustrates a constant temperature and varying irradiance. The temperature is constant at 25^{O}C while the irradiance varied from 1000 W/m2 to 200 W/m2. Also the power and the current varied proportionally with the irradiance. From figure 5.12 it can clearly be understood that the irradiance affects significantly the PV parameters. Figure 5.12 shows that the PI controller performs very poorly at the time the irradiance was varying

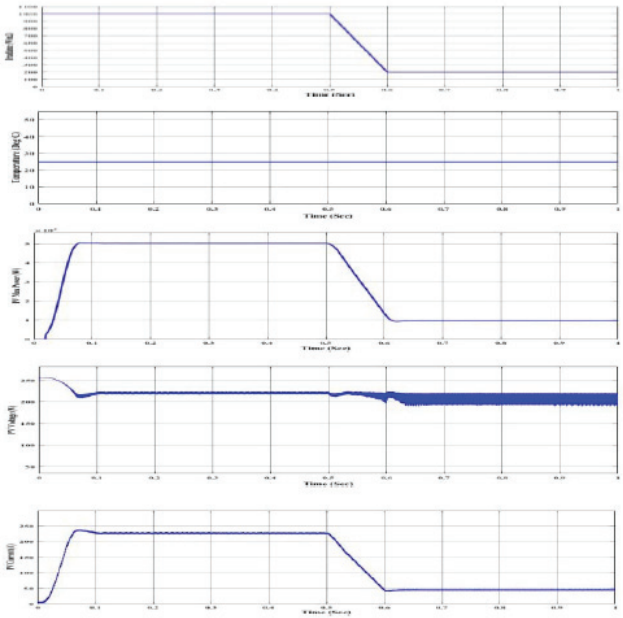


Fig. 8. Varying Irradiance and Constant Temperature

Figure 9 is the varying irradiance and constant temperature grid power, voltage and current. For this condition the input power varies proportional to the output power. We experience little form of power variation. With the help of the harmonics filter, our system grid side has a very good power quality. The voltage and current output are also pure sinusoidal for this condition.

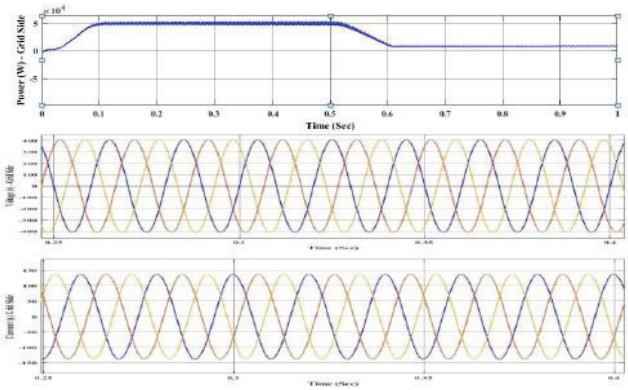


Fig. 9. Varying Irradiance and Constant Temperature Grid Parameters

D. Varying Irradiance and Varying Temperature

Figure 10 illustrates varying irradiance and temperature condition. As it is in the power section of this graph, the power dropped significantly from 0.5 to 0.6 second. Figure 10 also points out the poor performance of the PI controller where the irradiance and temperature are varying.

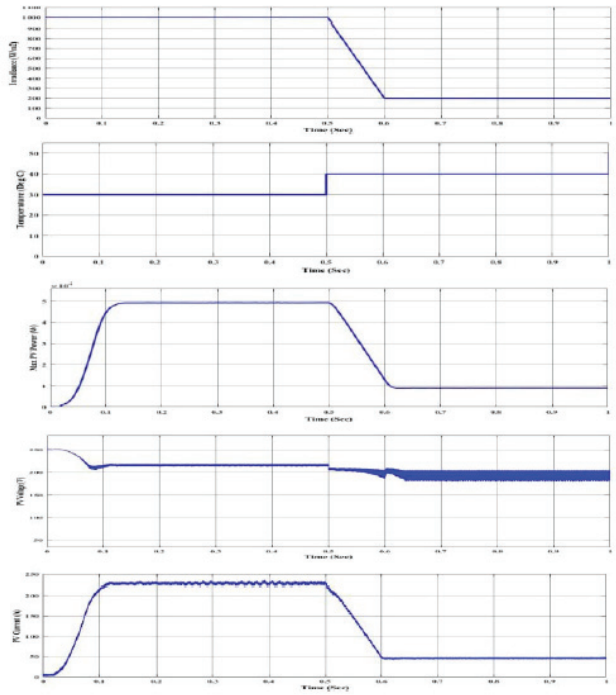


Fig. 10. Varying Irradiance and Temperature

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:13:32 UTC from IEEE Xplore. Restrictions apply.

Figure 11 is the varying irradiance and constant temperature grid power, voltage and current. For this condition the input power varies proportional to the output power. We experience little form of power variation. With the help of the harmonics filter, our system grid side has a very good power quality. The voltage and current output are also pure sinusoidal for this condition.

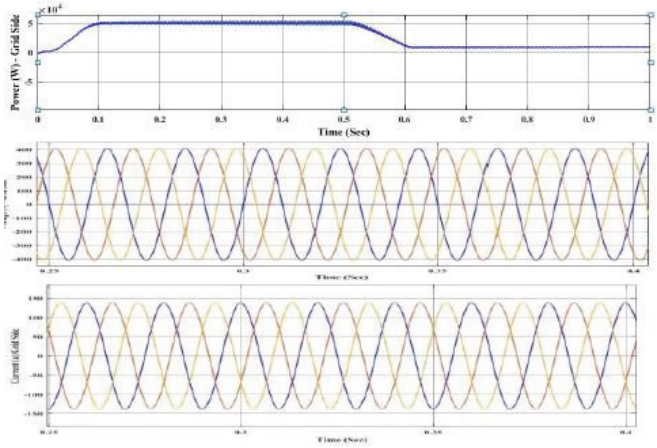


Fig. 11. Varying Irradiance and Temperature Grid Parameters

IV. CONCLUSION

In this paper, the design and analysis of a three-phase grid-connected solar PV inverter with perturb and observe MPPT model, proportional integration controller, solar PV mathematical model, and power grid were technically evaluated. The conditions that were considered for the simulation results were Constant Irradiance and Constant Temperature, Constant Irradiance and Varying Temperature, Varying Irradiance and Constant Temperature and Varying Irradiance and Varying Temperature. For the first condition, the irradiance was kept at 1000W/m2 while the temperature was kept at 25º C. And for the second condition the irradiance was kept at 1000 W/m^2 while the temperature varied between 30º C, 40º C and 50º C. For the third condition the temperature was kept at 25º C while the irradiance varied from 1000 W/m^2 to 100 W/m^2. Lastly, the temperature and irradiance were varied. This paper finding is expected to be implemented across Liberia to address energy efficiency issues. This project emphasizes the need for every business and individual to install a solar PV system connected to the grid to address the power deficit that Liberia is faced with. With model of 50kw power generation it can also be implemented in rural villages across Liberia. It offers several advantages. A solar PV system connected to the national grid enables renewable energy producers to sell power to the national power utility. In the third section of this paper three subsystems were modeled to analyze the performance of the three-phase grid-connected inverter. All of the models were designed using MatLab 2019.

As per what we got from the simulation results, it has been concluded that the integrated system response is fast and control accuracy can be achieved especially at the grid end. The control strategy also allows the inverter power to be controlled independently. At any given point where the irradiance is varying the current and voltage controllers show very poor performance but the grid voltage and current performances are good. Additionally, the analysis in this paper has been done in the time domain. The PI controller can be implemented in a practical project but the control strategy needs to be improved before applying it because the environmental factors change significantly during operations.

REFERENCES

[1] Rao, P V V Rama, 'Multistring Multilevel Inverter Topology For Microgrid Applications', IOSR Journal of Electrical and Electronics Engineering (IOSR-JEEE) E-ISSN:, 2014 (2014), pp.72–75

[2] S. A. Khajehoddin, P. Jain, and A. Bakhshai, "Cascaded multilevel converters and their applications in photovoltaic systems," in 2nd Canadian Solar Buildings Conf., Calgary, Canada, Jun. 10–14, 2007, pp. 1–7.

[3] Chitra and S. Himavathi, "Modeling and experimental validation of solar PV system for cascade H-bridge multilevel inverter," Power, Energy and Control Int. Conf., Dindigul, India, Feb. 6–8, 2013, pp. 60–265.

[4] Rivera, S., Kouro, S., Wu, B., Leon, J.I., Rodriguez, J., and Franquelo, L.G., "Cascaded H-bridge multilevel converter multistring topology for large scale photovoltaic systems," IEEE Industrial Electron. Int. Symp., Gdansk, Poland, Jun. 27-30, 2011, pp. 1837- 1844.

[5] H. Patangia and D. Gregory, "An efficient cascaded multilevel inverter suited for PV application," 35th IEEE Photovolt. Spec. Conf., Honolulu, HI, Jun. 20–25, 2010, pp. 2859–2863.

[6] Rahim, Nasrudin Abdul, Senior Member, and Jeyraj Selvaraj, 'A Novel Multi-String Five-Level PWM, Inverter for Photovoltaic Application', 2011 IEEE International Electric Machines & Drives Conference (IEMDC)

[7] R. Lai, F. Wang, R. Burgos, Y. Pei, D. Boroyevich, B. Wang, T. A. Lipo, V. D. Immanuel and K. J. Karimi, "A systematic topology evaluation method for high density three-phase pwm ac-ac converter", IEEE Transaction on Power Electronics, Vol. 23, NO. 6, November 2008.

[8] E. Rhyne, K. Smith and D. Bratton, "A 3-phase inverter using aferroresonant transformer assembled on a single laminated core", Telecommunications Energy Conference, 14th International 4-8 Oct. 1992.

[9] Md. Nahid Hossain, Tushar Kanti Routh, Abdul Hamid, Md. Iqbal Hossain, Ummal Husnaeen and Bin Yousuf, Design and Development of a Grid Tied Solar Inverter, International Conference on Informatic, Electronics and Vision, 2012

[10] Mowei, Loh Poh Cliang, and Ding Yi, Optimal Design of two Level Inverters, IEEE IPEC, 2010

[11] Kalpesh Y. Raval and V. J. Rupapara, Novel Multilevel Inverter Design with Reduced Device Count, Proceeding of 2018 IEEE International Conference on Current Trends toward Converging Technologies, 2018

[12] Mond Aizenddin Yusof, Muzaidi Othman, Sze Sing Lee, M. A. Rosland and J. H. Leong, Three Phase Multilevel Inverter with Reduced Number of Active Power Semiconductor Switches for Solar PV Modules, 2014 2nd International Conference on Electronic Design (ICED), August 19-12, 2014

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:13:32 UTC from IEEE Xplore. Restrictions apply.

Paper

Designing a Compact Microstrip Antenna Using the Machine Learning Approach

Kanhaiya Sharma and Ganga Prasad Pandey

School of Technology, Pandit Deendayal Petroleum University, Gandhinagar, India

https://doi.org/10.26636/jtit.2020.143520

Abstract—This paper presents how machine learning techniques may be applied in the process of designing a compact dual-band H-shaped rectangular microstrip antenna (RMSA) operating in 0.75–2.20 GHz and 3.0–3.44 GHz frequency ranges. In the design process, the same dimensions of upper and lower notches are incorporated, with the centered position right in the middle. Notch length and width are verified for investigating the antenna. An artificial neural network (ANN) model is developed from the simulated dataset, and is used for shape prediction. The same dataset is used to create a mathematical model as well. The predicted outcome is compared and it is determined that the model relying on ANN offers better results.

Keywords—artificial neural network, dual band, microstrip antenna, notch.

1. Introduction

New developments in the field of wireless communications call for compact, wide-band, high efficiency, multiband, and low cost antennas that are suitable for modern day applications. Microstrip antennas offer numerous advantages, such as small size, low weight and ease of fabrication due to their planar configuration. An antenna of this type is easy to mount on a rigid surface, supports dual-polarization and, unlike solutions of other types, multiband operation. The different techniques relied upon to make the antenna compact [1] without affecting its basic parameters include reactive loading [2], using a monopole antenna with defected ground structure (DGS), and modified patch [3]–[7]. Those methods are capable of reducing the antenna size by up to 70%. Meta-material [8]–[14] and engineered ground structure [15]–[18] may reduce the size of the antenna by up to 74% as well. Other miniaturization methods, like deployment of fractal structures in antenna design [19]–[21] may reduce the size of the antenna by up to 75% and engineered substrates [22]–[24] allow to reduce the size of antenna by 80% and by introducing close-ended or open-ended slots [25]–[30] of various shapes H, L, U, E, V, W, D, S double E, double U may reduce the size of the antenna by up to 86.5%. In some cases, more than one technique is used to make the antenna as compact as possible.

Machine learning (ML) is a subset of artificial intelligence (AI), effectively used in all areas of engineering, especially in communication network automation. ML is a specific type of data analytic techniques enabling machines to learn from experience (training) and to predict new data. A comparison between the traditional and the ML approach is given in Fig. 1. ML is preferred to solve complex mathematical problems with lots of variables and massive amounts of data. Today's methods relied upon while designing antennas require such a high level of expertise, as in most cases the design process requires a trade-off between such antenna parameters as small size and good bandwidth. AI approaches have been used in antenna design and optimization since the 1990s [31]. The most significant advantages that the use of the AI approach in antenna design has over the traditional methods lies in the ability to simultaneously handle multi-objective problems in order to achieve the specified goal and to provide the best automated solution [32]–[35].

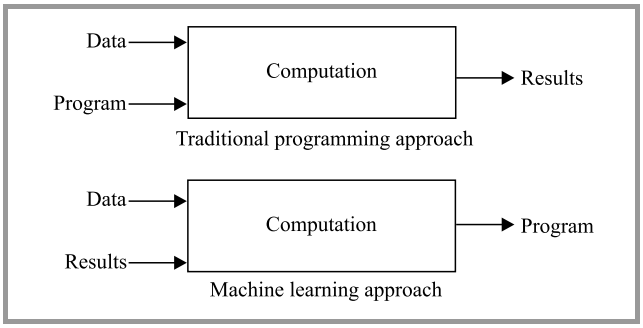


Fig. 1. Computational approaches.

The drawback of analytical-based methods consists in big-computation overheads experienced in the design process. To cope with this, the trial-and-error approach may be used to achieve the specified goals. Slot-loading turns out to be the technique that is most widely used among all aforementioned methods allowing to make antennas compact [28] in size.

2. Synthesis and Analysis Problem for Microstrip Patch Antenna Using ANN

ANN is one of the approaches used in ML to map nonlinear data efficiently [36], [37] based on experience (training),

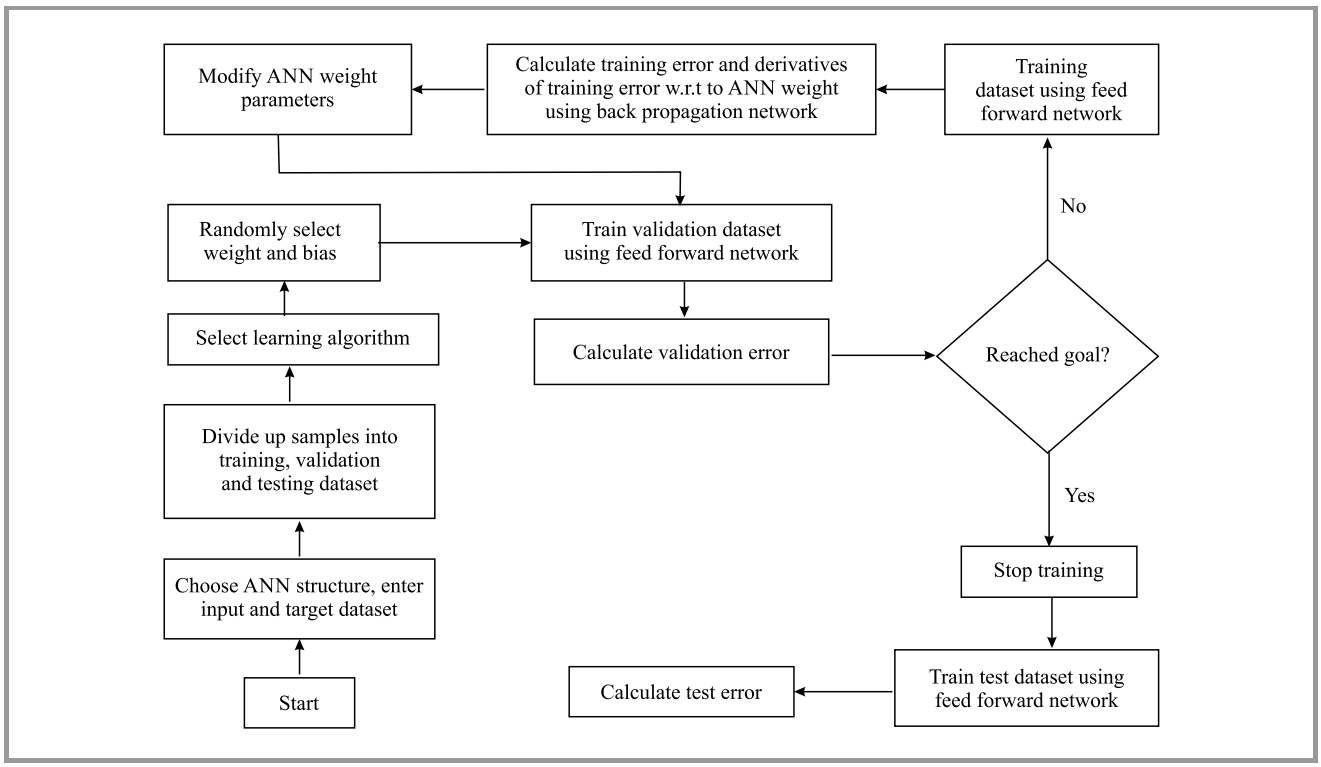


Fig. 2. Flowchart of ANN back-propagation algorithm.

while dealing with new data. Basically, three logic layers are distinguished in ANN architectures. The first is known as the input layer, while the last is referred to as the output layer. Between the input and the output layer, there is a number of hidden layers. Each hidden layer has one or more neurons. The number of the hidden layers, neurons, the activation function and the learning algorithm are all specific to the application. There are no rules to determine the number of hidden layers and neurons, but in most problems two to three hidden layers are used at the most in order to approximate all types of mathematical functions. The performance of ANN model depends upon the data collection, learning algorithm, weight initialization, change in an activation function, etc. In the case of antenna design, data should be collected either through simulations or measurements. The range of data always extends, marginally, beyond the model's utilization range [38]. Initially, data samples are divided into three sets, known as training (with the usage ratio of 70%), testing (15%), and validation (15%). According to the needs of a specific application, the percentages may be different. Next, the network size is chosen, i.e. the number of hidden layers and neurons in each of them is determined. Finally, the algorithm is selected based on feed-forward, back propagation, and feed-forward back propagation, to train the ANN and obtain the model. A flowchart of ANN feed-forward back-propagation algorithms is presented in Fig. 2.

To achieve the minimum mean squared error (MSE) It is necessary to train the model with accurate data. In this case, a total of 80 samples has been collected by simulating H-shaped RMSA using IE3D software. To minimize MSE, different ANN algorithms have been tested by varying the number of hidden layers and neurons in each hidden layer. Finally, the best combination for the proposed model was selected. To train the ANN model, highly accu-

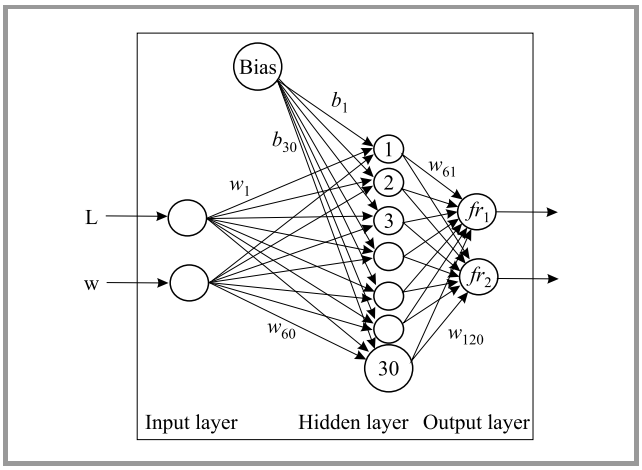


Fig. 3. Proposed ANN model for the analysis of compact H-shaped RMSA.

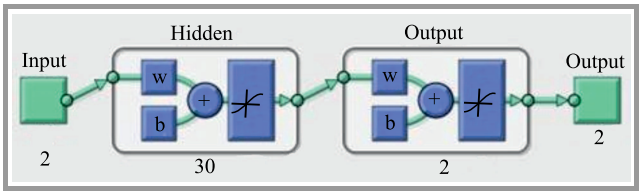


Fig. 4. ANN network architecture to train physical-to-electrical parameter.

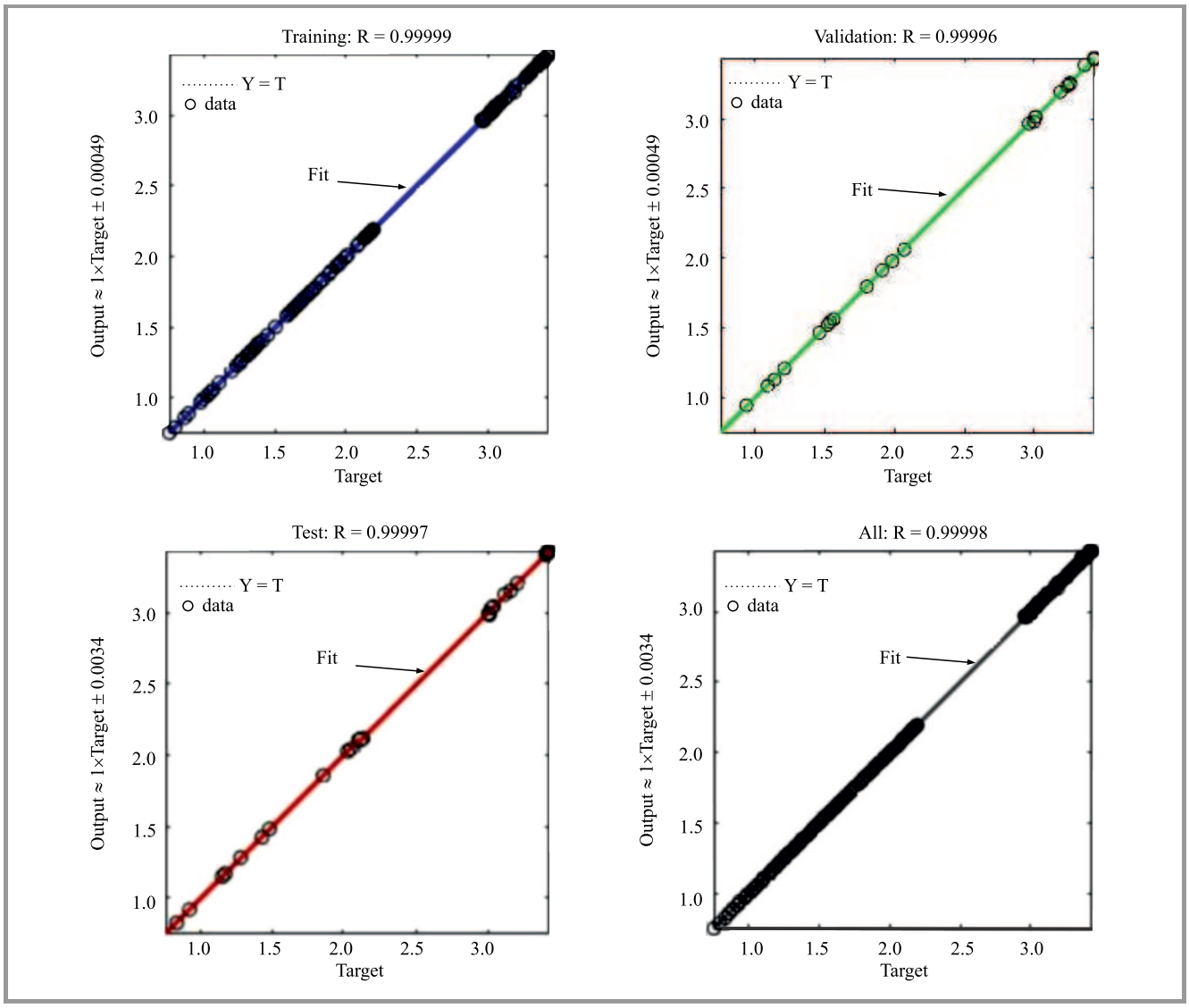


Fig. 5. Regression plot for physical to electrical parameters.

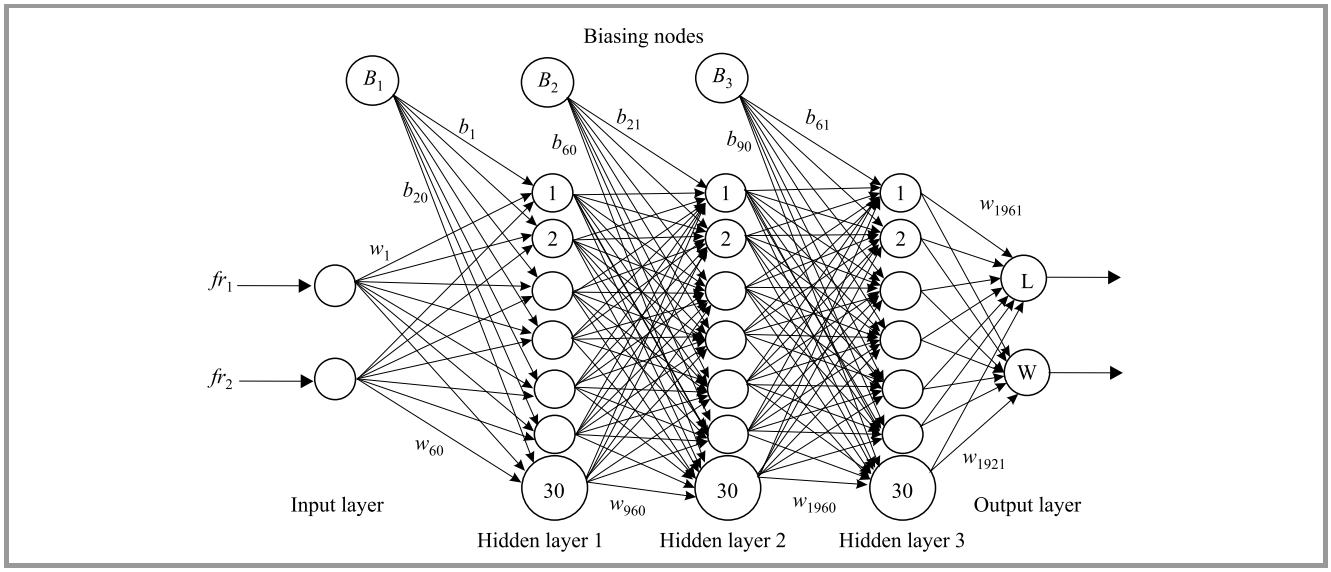


Fig. 6. Proposed ANN model for synthesis of H-shaped compact RMSA.

rate back propagation algorithms with minimum MSE were used [39].

Figure 3 shows a detailed architecture of the hidden layer of the proposed ANN model which uses the train set and predicts the dual resonant frequencies (f_{r1},f_{r2}), as well as the physical aerial parameters, i.e. slot length (L) and width (W) as input values. The Levenberg-Marquardt (LM) back propagation algorithm is used as ANN. ANN network architecture is shown in Fig. 4 with two input and output parameters, one hidden layer and 30 neurons. The regression plot of training, testing, validation, and overall process is shown in Fig. 5.

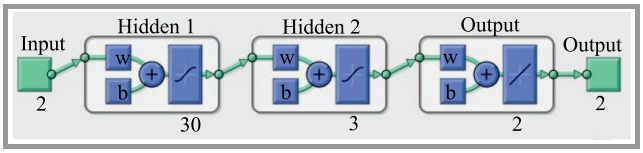


Fig. 7. ANN network architecture to train electrical-to-physical parameters.

To train the ANN model for predicting physical dimensions of antenna, Bayesian Regularization (BR) back propagation algorithm have been used. ANN network architecture is shown in Figs. 6–7, where input and output parameters equals to two, the number of hidden layer is three, each hidden layer consists of thirty neurons. the regression plot of training, testing, and overall process is shown in Fig. 8. In the next step, using the same dataset, a mathematical model was developed to map the dominant resonance mode (TM_{10}) and the higher mode (TM_{02}) of a dual band compact H-shaped patch antenna using the curve fitting tool in Matlab software. Equation 1 represents the dominant mode (TM_{10}) and Eq. 2 represents the higher mode (TM_{02}):

$$fr_1 = 2.279 - 0.145l + .07311w\,, \tag{1}$$

$$fr_2 = 3.37 + 0.1202l - 0.081w - 0.05906l^2 + 0.04877lw + 0.002621l^3 - 0.002256l^2w\,. \tag{2}$$

3. Result and Discussion

Analysis of the H-shaped compact RMSA using the proposed ANN model is performed by taking one hidden layer and 30 neurons, as shown in Fig. 4. 80 samples were taken to train the model, with 70% used for training, 15% for validation and 15% for testing. Figure 5 shows the regression plots for training, validation, testing, and the overall process. The regression R value indicating the correlation between output and target, R value closer to one means that the close relationship. In the ideal case, both outputs and target are equal. It is clear from the evidence shown that the R value for training, validation, and testing of the proposed model is 99.998% accurate, meaning that it is perfect regression model. The synthesis of H-shaped compact RMSA using the proposed ANN model is performed by taking three hidden layers and 30 neurons per hidden layer, as shown in ANN network architecture in Figs. 6–7.

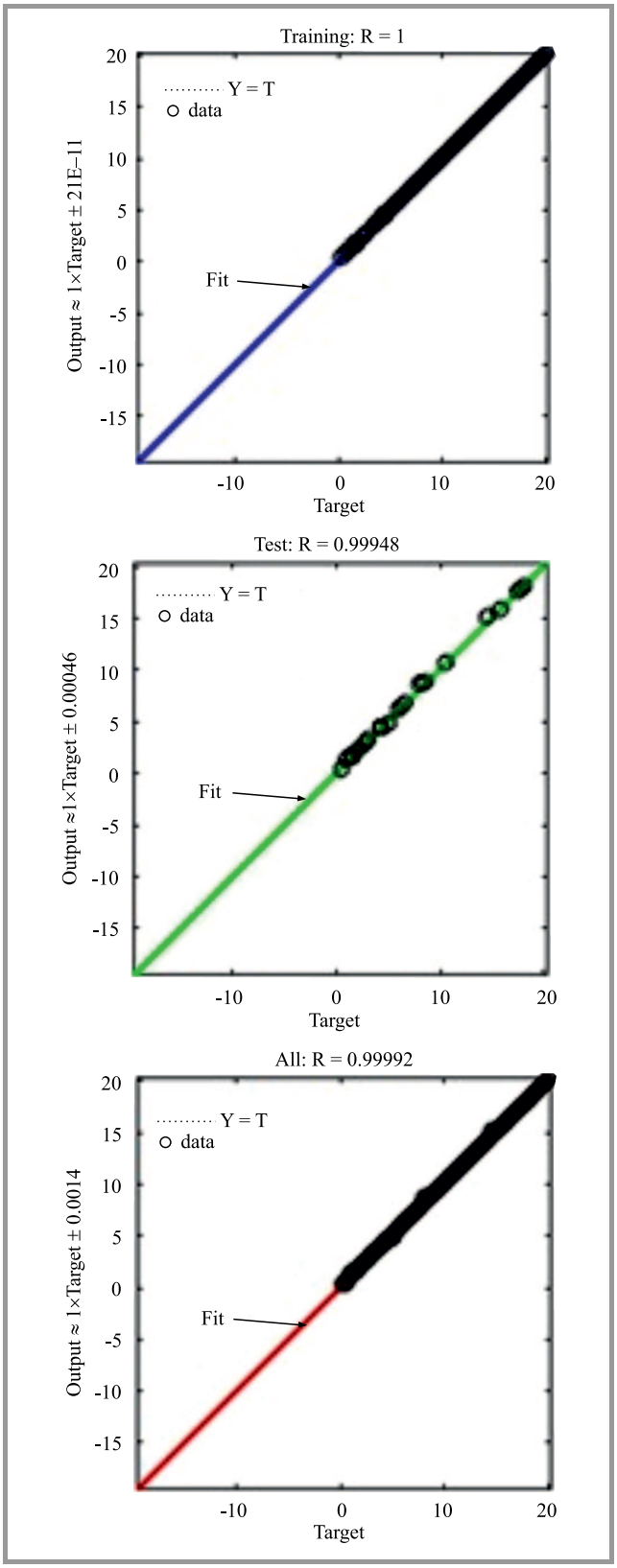


Fig. 8. Regression plot for electrical to a physical parameters.

In this case, 80 samples are used to train the model (85% for training, 15% for testing). Figure 8 shows regression plots for training, testing, and the overall process. It is clear from the evidence shown that the regression R value

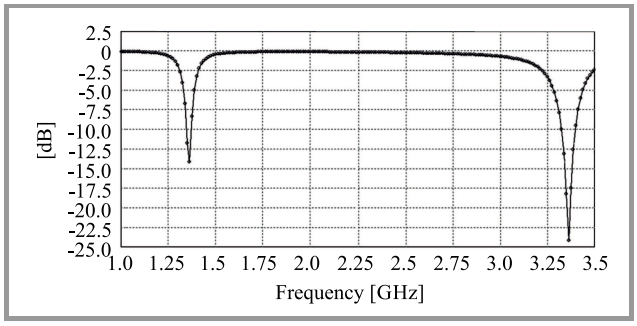


Fig. 9. S_{11} vs. frequency plot for H-shaped RMSA antenna.

for training and testing of the proposed model is 99.992%, which is very close to benchmark value 1. In the ideal scenario, both output and target are equal. In both the cases, the variations are placed precisely on dotted line, as shown in Figs. 5 and 8.

Table 1 shows the result of the proposed ANN model used for analyzing a compact H-shaped RMSA. Except for a few samples, MSE is of the order of 10^{-5}, which indicates that the predicted resonant frequencies are close to simulated results (f_{r1},f_{r2}). Table 2 shows that the MSE is of the order of 10^{-12}, which is very close to actual physical slot dimensions. Table 3 shows a comparison of f_{r1},f_{r2} collected from simulation, curve fitting, and ANN prediction. The dual resonant frequencies predicted by the proposed ANN model are very close to the simulated ones. The relation between rectangular notch dimensions and resonant frequencies f_{r1} and f_{r2} for the dominant mode (TM_{10} and for the higher mode TM_{02}) are shown in Fig. 10. The dimension of notches is increased from 0.25×0.5 mm to 20.00×20.25 mm, in 0.25 mm steps. f_{r1} started falling from 2.20 to 0.75 GHz, while f_{r2} first increased from 3.38 GHz to 3.434 GHz, and then suddenly decreased down to 2.98 GHz, and started increasing again from 2.98 to 3.19 GHz.

To validate the performance of H-shaped dual-band RMSA, one sample was chosen randomly with the same dimensions as those shown in Table 1, with the notch size of 8×14 mm, as well as with identical lower and upper notches centered in the middle. The infinite substrate and the infinite ground plane were assumed and simulated using IE3D software. The performance of the antenna measured in terms of scattering parameter, gain and radiation pattern was analyzed. S_{11} vs. frequency characteristics shown in Fig. 9, indicate that the antenna radiates at 1.360 GHz and 3.366 GHz for two modes (TM_{10} and TM_{02}). From Fig. 11, one may conclude that the gain of the antenna in question measured at 1.360 GHz is 7.572 dBi, and equals, at the upper resonant frequency of 3.366 GHz, 1.00 dBi. E-plane and H-plane radiation patterns at 1.360 GHz and 3.366 GHz with infinite substrate and infinite ground plane dimensions (Fig. 12) show TM_{10} and TM_{02} modes of operation, which is quite normal for any MSA. The proposed aerial offers better results in terms of accuracy and time complexity, with a higher

Table 1
Physical-to-electrical parameters

Prediction of physical parameters (slot length, slot width) to electrical (f_{r1}, f_{r2}) data							
Slot (l,w)		Simulation		ANN prediction		ANN model error	
L	W	f_{r1}	f_{r2}	f_{r1}	f_{r2}	f_{r1}	f_{r2}
0.25	0.5	2.19209	3.38418	2.19419	3.37800	–2.10E–03	6.18E–03
0.5	0.75	2.19128	3.38579	2.19032	3.37734	9.58E–04	8.45E–03
0.75	1	2.18725	3.36158	2.18592	3.37702	1.33E–03	–1.54E–02
1	1.25	2.17918	3.36965	2.18099	3.37718	–1.81E–03	–7.53E–03
1.25	1.5	2.17918	3.36965	2.17556	3.37798	3.62E–03	–8.33E–03
1.5	1.75	2.17111	3.38579	2.16963	3.37965	1.48E–03	6.14E–03
1.75	2	2.159	3.38176	2.16319	3.38238	–4.19E–03	–6.21E–04
2	2.25	2.15819	3.38983	2.15618	3.38633	2.01E–03	3.50E–03
2.25	2.5	2.14689	3.38983	2.14848	3.39150	–1.59E–03	–1.67E–03
2.5	2.75	2.13882	3.3979	2.13989	3.39770	–1.07E–03	2.00E–04
2.75	3	2.13075	3.40194	2.13019	3.40449	5.58E–04	–2.55E–03
3	3.25	2.12268	3.41001	2.11927	3.41125	3.41E–03	–1.24E–03
3.25	3.5	2.10734	3.41808	2.10715	3.41736	1.94E–04	7.18E–04
3.5	3.75	2.0904	3.42373	2.09398	3.42232	–3.58E–03	1.41E–03
3.75	4	2.0791	3.42938	2.07995	3.42589	–8.54E–04	3.49E–03
4	4.25	2.0678	3.42938	2.06515	3.42813	2.65E–03	1.25E–03
4.25	4.5	2.05198	3.42147	2.04960	3.42924	2.38E–03	–7.77E–03
4.5	4.75	2.02986	3.43422	2.03329	3.42955	–3.43E–03	4.67E–03
4.75	5	2.0113	3.42933	2.01622	3.42935	–4.92E–03	–2.47E–05
5	5.25	1.99435	3.42373	1.99848	3.42897	–4.13E–03	–5.24E–03
5.25	5.5	1.98144	3.43019	1.98024	3.42870	1.20E–03	1.49E–03
5.5	5.75	1.96045	3.42938	1.96183	3.42879	–1.38E–03	5.88E–04
5.75	6	1.94915	3.43019	1.94373	3.42941	5.42E–03	7.82E–04
6	6.25	1.92655	3.42938	1.92650	3.43041	4.79E–05	–1.03E–03
6.25	6.5	1.9096	3.43503	1.91041	3.43109	–8.10E–04	3.94E–03
6.5	6.75	1.89266	3.42615	1.89493	3.42987	–2.27E–03	–3.72E–03
6.75	7	1.88055	3.42211	1.87854	3.42462	2.01E–03	–2.51E–03
7	7.25	1.85876	3.41243	1.85964	3.41401	–8.76E–04	–1.58E–03
7.25	7.5	1.83616	3.40678	1.83835	3.39919	–2.19E–03	7.59E–03
7.5	7.75	1.81356	3.40113	1.81694	3.38378	–3.38E–03	1.73E–02
7.75	8	1.79984	3.36158	1.79758	3.37135	2.26E–03	–9.77E–03
8	8.25	1.77966	3.36965	1.78009	3.36310	–4.29E–04	6.55E–03
8.25	8.5	1.75706	3.36723	1.76202	3.35787	–4.96E–03	9.36E–03
8.5	8.75	1.73931	3.35755	1.74105	3.35359	–1.74E–03	3.96E–03
8.75	9	1.71913	3.34948	1.71722	3.34863	1.91E–03	8.49E–04
9	9.25	1.69492	3.34463	1.69299	3.34215	1.93E–03	2.48E–03
9.25	9.5	1.67232	3.32768	1.67094	3.33390	1.38E–03	–6.22E–03
9.5	9.75	1.65052	3.32123	1.65179	3.32382	–1.27E–03	–2.59E–03
9.75	10	1.63035	3.30912	1.63440	3.31195	–4.05E–03	–2.83E–03
10	10.25	1.61824	3.29701	1.61715	3.29858	1.09E–03	–1.57E–03
10.25	10.5	1.59887	3.28249	1.59908	3.28440	–2.10E–04	–1.91E–03
10.5	10.75	1.58192	3.26877	1.58012	3.27032	1.80E–03	–1.55E–03
10.75	11	1.56174	3.25666	1.56061	3.25704	1.13E–03	–3.84E–04
11	11.25	1.53672	3.25424	1.54085	3.24467	–4.13E–03	9.57E–03
11.25	11.5	1.51977	3.23729	1.52096	3.23276	–1.19E–03	4.53E–03
11.5	11.75	1.50121	3.22437	1.50100	3.22060	2.11E–04	3.77E–03
11.75	12	1.48103	3.20823	1.48105	3.20739	–2.26E–05	8.39E–04
12	12.25	1.46086	3.18805	1.46127	3.19248	–4.14E–04	–4.43E–03
12.25	12.5	1.44633	3.16949	1.44182	3.17556	4.51E–03	–6.07E–03
12.5	12.75	1.42857	3.15981	1.42277	3.15689	5.80E–03	2.92E–03
12.75	13	1.40839	3.13963	1.40409	3.13736	4.30E–03	2.27E–03
13	13.25	1.38015	3.11945	1.38564	3.11828	–5.49E–03	1.17E–03
13.25	13.5	1.36723	3.09605	1.36729	3.10072	–5.51E–05	–4.67E–03
13.5	13.75	1.3519	3.08717	1.34903	3.08499	2.87E–03	2.18E–03
13.75	14	1.33172	3.07103	1.33100	3.07076	7.23E–04	2.67E–04
14	14.25	1.31961	3.05892	1.31328	3.05773	6.33E–03	1.19E–03
14.25	14.5	1.29944	3.0565	1.29588	3.04595	3.56E–03	1.05E–02
14.5	14.75	1.27926	3.03874	1.27872	3.03576	5.38E–04	2.98E–03
14.75	15	1.25989	3.02825	1.26165	3.02751	–1.76E–03	7.44E–04
15	15.25	1.25101	3.03874	1.24446	3.02143	6.55E–03	1.73E–02
15.25	15.5	1.22599	3.01695	1.22687	3.01752	–8.80E–04	–5.65E–04
15.5	15.75	1.20904	3.0113	1.20869	3.01522	3.49E–04	–3.92E–03
15.75	16	1.19209	3.0113	1.18994	3.01306	2.15E–03	–1.76E–03
16	16.25	1.1703	3.00646	1.17083	3.00867	–5.34E–04	–2.21E–03
16.25	16.5	1.15012	3.01049	1.15145	3.00000	–1.33E–03	1.05E–02
16.5	16.75	1.13801	2.99839	1.13131	2.98718	6.70E–03	1.12E–02
16.75	17	1.10169	2.95803	1.10966	2.97337	–7.97E–03	–1.53E–02
17	17.25	1.08959	2.96207	1.08634	2.96319	3.25E–03	–1.12E–03
17.25	17.5	1.06215	2.96045	1.06208	2.95967	6.62E–05	7.76E–04
17.5	17.75	1.03955	2.9661	1.03784	2.96245	1.71E–03	3.65E–03
17.75	18	1.01695	2.9774	1.01407	2.96922	2.88E–03	8.18E–03
18	18.25	0.98870	2.97821	0.99068	2.97837	–1.98E–03	–1.62E–04
18.25	18.5	0.97175	2.9887	0.96723	2.99017	4.52E–03	–1.47E–03
18.5	18.75	0.93624	3.00646	0.94302	3.00600	–6.79E–03	4.57E–04
18.75	19	0.92009	3.02663	0.91728	3.02733	2.81E–03	–7.04E–04
19	19.25	0.88701	3.05085	0.88933	3.05488	–2.33E–03	–4.03E–03
19.25	19.5	0.85876	3.0904	0.85881	3.08823	–5.32E–05	2.17E–03
19.5	19.75	0.83051	3.11864	0.82578	3.12581	4.73E–03	–7.17E–03
19.75	20	0.79096	3.18805	0.79066	3.16525	2.95E–04	2.28E–02
20	20.25	0.75141	3.19209	0.75411	3.20409	–2.70E–03	–1.20E–02

Table 2
Electrical-to-physical parameters

Prediction of electrical parameters (f_{r1}, f_{r2}) to physical parameters (slot length, slot width)							
Slot (l, w)		Input to ANN		ANN prediction		Error in prediction)	
L	W	f_{r1}	f_{r2}	f_r1	f_{r2}	f_{r1}	f_{r2}
0.25	0.5	2.19209	3.38418	0.25000	0.50000	4.26E–11	4.26E–11
0.5	0.75	2.19128	3.38579	0.23763	0.48763	0.26237	0.26237
0.75	1	2.18725	3.36158	0.75000	1.00000	7.43E–11	7.43E–11
1	1.25	2.17918	3.36965	1.25000	1.50000	–0.25	–0.25
1.25	4.4	2.17918	3.36965	1.25000	1.50000	5.21E–11	5.21E–11
1.5	1.75	2.17111	3.38579	1.50000	1.75000	5.05E–11	5.05E–11
1.75	2	2.15900	3.38176	1.75000	2.00000	5.57E–11	5.57E–11
2	2.25	2.15819	3.38983	2.05693	2.30693	–0.05693	–0.05693
2.25	2.5	2.14689	3.38983	2.25000	2.50000	5.40E–11	5.40E–11
2.5	2.75	2.13882	3.39790	2.50000	2.75000	4.79E–11	4.79E–11
2.75	3	2.13075	3.40194	2.71920	2.96920	0.03080	0.03080
3	3.25	2.12268	3.41001	3.11056	3.36056	–0.11056	–0.11056
3.25	3.5	2.10734	3.41808	3.25000	3.50000	4.22E–11	4.22E–11
3.5	3.75	2.09040	3.42373	3.50000	3.75000	3.58E–11	3.58E–11
3.75	4	2.07910	3.42938	3.75000	4.00000	3.03E–11	3.03E–11
4	4.25	2.06780	3.42938	4.00000	4.25000	2.51E–11	2.51E–11
4.25	4.5	2.05198	3.42147	4.25000	4.50000	2.95E–11	2.95E–11
4.5	4.75	2.02986	3.43422	4.50000	4.75000	1.48E–11	1.48E–11
4.75	5	2.01130	3.42933	4.75000	5.00000	1.46E–11	1.46E–11
5	5.25	1.99435	3.42373	4.64830	4.89830	0.35170	0.35170
5.25	5.5	1.98144	3.43019	5.25000	5.50000	8.55E–12	8.55E–12
5.5	5.75	1.96045	3.42938	5.50000	5.75000	2.45E–12	2.45E–12
5.75	6	1.94915	3.43019	5.75000	6.00000	4.59E–12	4.59E–12
6	6.25	1.92655	3.42938	6.07401	6.32401	–0.07400	–0.07400
6.25	6.5	1.90960	3.43503	6.25000	6.50000	1.19E–11	1.19E–11
6.5	6.75	1.89266	3.42615	6.55910	6.80910	–0.05909	–0.05909
6.75	7	1.88055	3.42211	6.75000	7.00000	2.72E–11	2.72E–11
7	7.25	1.85876	3.41243	7.00000	7.25000	3.61E–11	3.61E–11
7.25	7.5	1.83616	3.40678	7.25000	7.50000	4.98E–11	4.98E–11
7.5	7.75	1.81356	3.40113	7.50000	7.75000	5.88E–11	5.88E–11
7.75	8	1.79984	3.36158	7.75000	8.00000	–6.34E–12	–6.35E–12
8	8.25	1.77966	3.36965	8.51865	8.76865	–0.51865	–0.51865
8.25	8.5	1.75706	3.36723	8.25000	8.50000	–3.08E–12	–3.08E–12
8.5	8.75	1.73931	3.35755	8.62045	8.87045	–0.12045	–0.12045
8.75	9	1.71913	3.34948	8.75000	9.00000	–7.34E–12	–7.34E–12
9	9.25	1.69492	3.34463	9.00000	9.25000	–9.65E–12	–9.65E–12
9.25	9.5	1.67232	3.32768	9.25000	9.50000	–9.32E–12	–9.32E–12
9.5	9.75	1.65052	3.32123	9.50000	9.75000	–1.20E–11	–1.20E–11
9.75	10	1.63035	3.30912	9.75000	10.0000	–9.44E–12	–9.44E–12
10	10.25	1.61824	3.29701	10.0000	10.2500	–1.13E–11	–1.13E–11
10.25	10.5	1.59887	3.28249	10.2500	10.5000	–2.71E–12	–2.71E–12
10.5	10.75	1.58192	3.26877	10.5246	10.7746	–0.02457	–0.02457
10.75	11	1.56174	3.25666	10.7500	11.0000	–6.22E–12	–6.22E–12
11	11.25	1.53672	3.25424	11.0000	11.2500	–7.51E–12	–7.51E–12
11.25	11.5	1.51977	3.23729	11.2500	11.5000	1.12E–11	1.12E–11
11.5	11.75	1.50121	3.22437	11.5000	11.7500	–1.75E–11	–1.75E–11
11.75	12	1.48103	3.20823	11.7500	12.0000	8.74E–12	8.74E–12
12	12.25	1.46086	3.18805	12.0000	12.2500	–3.49E–12	–3.49E–12
12.25	12.5	1.44633	3.16949	12.2500	12.5000	3.14E–12	3.14E–12
12.5	12.75	1.42857	3.15981	12.5000	12.7500	–1.12E–12	–1.12E–12
12.75	13	1.40839	3.13963	12.7500	13.0000	5.51E–13	5.51E–13
13	13.25	1.38015	3.11945	13.0000	13.2500	7.48E–13	7.48E–13
13.25	13.5	1.36723	3.09605	13.2500	13.5000	–4.42E–13	–4.42E–13
13.5	13.75	1.35190	3.08717	13.5000	13.7500	1.74E–13	1.74E–13
13.75	14	1.33172	3.07103	13.7500	14.0000	–2.08E–12	–2.08E–12
14	14.25	1.31961	3.05892	14.0000	14.2500	–4.21E–12	–4.21E–12
14.25	14.5	1.29944	3.05650	14.2500	14.5000	–6.14E–12	–6.14E–12
14.5	14.75	1.27926	3.03874	14.9728	15.2228	–0.47284	–0.47284
14.75	15	1.25989	3.02825	14.7500	15.0000	–1.61E–10	–1.61E–10
15	15.25	1.25101	3.03874	15.0000	15.2500	–1.07E–10	–1.07E–10
15.25	15.5	1.22599	3.01695	15.2500	15.5000	–1.07E–10	–1.07E–10
15.5	15.75	1.20904	3.01130	15.5000	15.7500	–6.79E–11	–6.79E–11
15.75	16	1.19209	3.01130	15.6709	15.9209	0.07901	0.07901
16	16.25	1.17030	3.00646	16.0000	16.2500	–4.76E–11	–4.76E–11
16.25	16.5	1.15012	3.01049	16.2500	16.5000	–7.06E–11	–7.06E–11
16.5	16.75	1.13801	2.99839	16.5000	16.7500	–2.86E–11	–2.86E–11
16.75	17	1.10169	2.95803	16.7500	17.0000	–5.46E–12	–5.46E–12
17	17.25	1.08959	2.96207	17.0000	17.2500	–5.68E–12	–5.68E–12
17.25	17.5	1.06215	2.96045	17.2500	17.5000	–4.63E–12	–4.63E–12
17.5	17.75	1.03955	2.96610	17.4709	17.7209	0.02905	0.02906
17.75	18	1.01695	2.97740	17.7500	18.0000	–8.48E–12	–8.48E–12
18	18.25	0.98870	2.97821	17.8963	18.1463	0.10368	0.10368
18.25	18.5	0.97175	2.98870	18.2500	18.5000	–2.24E–11	–2.24E–11
18.5	18.75	0.93624	3.00646	18.5000	18.7500	–1.19E–10	–1.19E–10
18.75	19	0.92010	3.02663	18.7500	19.0000	–3.99E–10	–3.99E–10
19	19.25	0.88701	3.05085	19.0000	19.2500	–4.97E–11	–4.97E–11
19.25	19.5	0.85876	3.09040	19.2500	19.5000	–2.32E–11	–2.32E–11
19.5	19.75	0.83051	3.11864	19.5000	19.7500	–1.22E–11	–1.22E–11
19.75	20	0.79096	3.18805	19.7500	20.0000	1.98E–11	1.98E–11
20	20.25	0.75141	3.19209	20.0000	20.2500	1.50E–11	1.50E–11

Table 3
Comparison of resonant frequencies

Comparison of resonant frequencies in various methods							
Slot size		Simulation		Curve fitting		ANN model	
L	W	f_{r1}	f_{r2}	f_{r1}	f_{r2}	f_{r1}	f_{r2}
0.25	0.5	2.19209	3.38418	2.2793	3.3619	2.19323512	3.383955179
0.5	0.75	2.19128	3.38579	2.2613	3.3728	2.190414971	3.377723607
0.75	1	2.18725	3.36158	2.2434	3.3823	2.186759462	3.373556577
1	1.25	2.17918	3.36965	2.2254	3.3907	2.18221719	3.37167556
1.25	1.5	2.17918	3.36965	2.2074	3.3977	2.176752125	3.372181943
1.5	1.75	2.17111	3.38579	2.1894	3.4036	2.170350743	3.374991672
1.75	2	2.159	3.38176	2.1751	3.4084	2.163026757	3.379792164
2	2.25	2.15819	3.38983	2.1535	3.412	2.154820328	3.386051178
2.25	2.5	2.14689	3.38983	2.1355	3.4146	2.145790119	3.393092756
2.5	2.75	2.13882	3.3979	2.1176	3.4161	2.136000000	3.400223563
2.75	3	2.13075	3.40194	2.0996	3.4166	2.12550524	3.406863738
3	3.25	2.12268	3.41001	2.0816	3.4161	2.11434336	3.412633595
3.25	3.5	2.10734	3.41808	2.0636	3.4147	2.102532065	3.417373395
3.5	3.75	2.0904	3.42373	2.0457	3.4123	2.090073149	3.421106618
3.75	4	2.0791	3.42938	2.0277	3.4091	2.076959378	3.423975275
4	4.25	2.0678	3.42938	2.0097	3.405	2.063181506	3.426174076
4.25	4.5	2.05198	3.42147	1.9917	3.4001	2.048733878	3.427898365
4.5	4.75	2.02986	3.43422	1.9738	3.3945	2.033618293	3.429309015
4.75	5	2.0113	3.42933	1.9558	3.3881	2.017846529	3.430510785
5	5.25	1.99435	3.42373	1.9378	3.381	2.001442149	3.431538579
5.25	5.5	1.98144	3.43019	1.9199	3.3732	1.984442103	3.432346835
5.5	5.75	1.96045	3.42938	1.9019	3.3648	1.966898343	3.432799628
5.75	6	1.94915	3.43019	1.8839	3.3558	1.948879197	3.432662761
6	6.25	1.92655	3.42938	1.8659	3.3462	1.930469403	3.431604735
6.25	6.5	1.9096	3.43503	1.848	3.3361	1.911766619	3.429221116
6.5	6.75	1.89266	3.42615	1.83	3.3255	1.892871285	3.425103028
6.75	7	1.88055	3.42211	1.812	3.3144	1.873867609	3.418964553
7	7.25	1.85876	3.41243	1.794	3.3028	1.854798692	3.410809045
7.25	7.5	1.83616	3.40678	1.7761	3.2909	1.835648248	3.401051244
7.5	7.75	1.81356	3.40113	1.7581	3.2786	1.81634694	3.390475111
7.75	8	1.79984	3.36158	1.7401	3.266	1.796810053	3.379982051
8	8.25	1.77966	3.36965	1.7222	3.2531	1.776988061	3.370251304
8.25	8.5	1.75706	3.36723	1.7042	3.2399	1.756898836	3.361519085
8.5	8.75	1.73931	3.35755	1.6862	3.2265	1.736625741	3.353579746
8.75	9	1.71913	3.34948	1.6682	3.2129	1.716290452	3.345948928
9	9.25	1.69492	3.34463	1.6503	3.1992	1.69601866	3.33806789
9.25	9.5	1.67232	3.32768	1.6323	3.1853	1.675911158	3.329468411
9.5	9.75	1.65052	3.32123	1.6143	3.1713	1.656025072	3.319872427
9.75	10	1.63035	3.30912	1.5964	3.1573	1.636366626	3.309226066
10	10.25	1.61824	3.29701	1.5784	3.1433	1.616895951	3.297672933
10.25	10.5	1.59887	3.28249	1.5604	3.1292	1.597542878	3.28547673
10.5	10.75	1.58192	3.26877	1.5424	3.1153	1.578229533	3.272915207
10.75	11	1.56174	3.25666	1.5245	3.1013	1.558892864	3.260178391
11	11.25	1.53672	3.25424	1.5065	3.0875	1.539500355	3.247303294
11.25	11.5	1.51977	3.23729	1.4885	3.0739	1.52005534	3.234163114
11.5	11.75	1.50121	3.22437	1.4705	3.0604	1.500592591	3.22050971
11.75	12	1.48103	3.20823	1.4526	3.0472	1.481167693	3.206053827
12	12.25	1.46086	3.18805	1.4346	3.0342	1.461844345	3.190561503
12.25	12.5	1.44633	3.16949	1.4166	3.0215	1.442682842	3.173944500
12.5	12.75	1.42857	3.15981	1.3987	3.0091	1.423731612	3.15632401
12.75	13	1.40839	3.13963	1.3807	2.9971	1.405022572	3.13805028
13	13.25	1.38015	3.11945	1.3627	2.9854	1.386570215	3.119669014
13.25	13.5	1.36723	3.09605	1.3447	2.9742	1.368373589	3.101839329
13.5	13.75	1.3519	3.08717	1.3268	2.9634	1.350419758	3.085223618
13.75	14	1.33172	3.07103	1.3088	2.9532	1.332686968	3.070378841
14	14.25	1.31961	3.05892	1.2908	2.9434	1.315145999	3.057675597
14.25	14.5	1.29944	3.0565	1.2728	2.9342	1.297758688	3.047257546
14.5	14.75	1.27926	3.03874	1.2549	2.9256	1.280473345	3.039037159
14.75	15	1.25989	3.02825	1.2369	2.9177	1.263217358	3.03271293
15	15.25	1.25101	3.03874	1.2189	2.9104	1.245887804	3.02779209
15.25	15.5	1.22599	3.01695	1.201	2.9038	1.228342041	3.023611391
15.5	15.75	1.20904	3.0113	1.183	2.8979	1.210392669	3.019365567
15.75	16	1.19209	3.0113	1.165	2.8928	1.191815554	3.014176798
16	16.25	1.1703	3.00646	1.147	2.8884	1.172383691	3.007259144
16.25	16.5	1.15012	3.01049	1.1291	2.885	1.151935686	2.9982168
16.5	16.75	1.13801	2.99839	1.1111	2.8824	1.130462957	2.98741289
16.75	17	1.10169	2.95803	1.0931	2.8806	1.108157717	2.976170793
17	17.25	1.08959	2.96207	1.0751	2.8799	1.085352868	2.966523582
17.25	17.5	1.06215	2.96045	1.0572	2.8801	1.062359012	2.960532524
17.5	17.75	1.03955	2.9661	1.0392	2.8813	1.039305812	2.959616637
17.75	18	1.01695	2.9774	1.0212	2.8835	1.016091761	2.964321303
18	18.25	0.988701	2.97821	1.0033	2.8868	0.992442652	2.974524237
18.25	18.5	0.971751	2.9887	0.9853	2.8912	0.968009009	2.989787327
18.5	18.75	0.936239	3.00646	0.9673	2.8963	0.942446684	3.009621486
18.75	19	0.920097	3.02663	0.9493	2.9035	0.915466504	3.033605786
19	19.25	0.887006	3.05085	0.9314	2.9114	0.886861758	3.061400275
19.25	19.5	0.858757	3.0904	0.9134	2.9206	0.856524855	3.092706667
19.5	19.75	0.830508	3.11864	0.8954	2.9311	0.824459185	3.127214353
19.75	20	0.79096	3.18805	0.8774	2.9429	0.790787012	3.164552035
20	20.25	0.751412	3.19209	0.8595	2.956	0.755751204	3.204255063

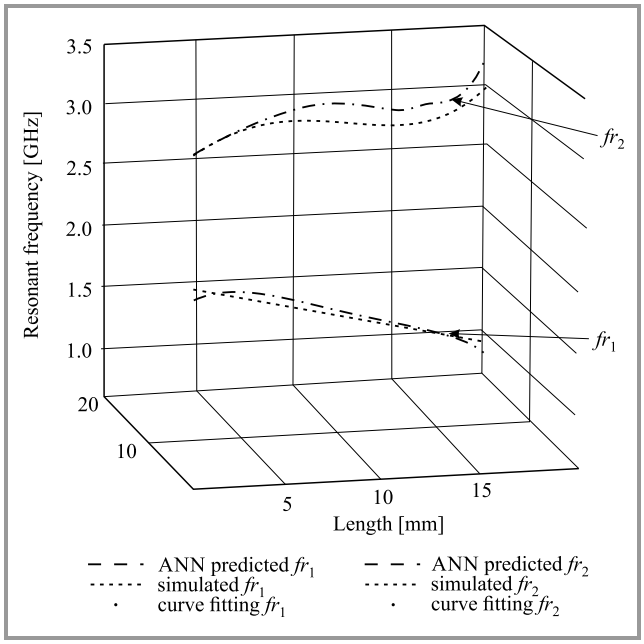


Fig. 10. Notch dimension vs. resonant frequency for TM_{10}, and TM_{02} modes of operation.

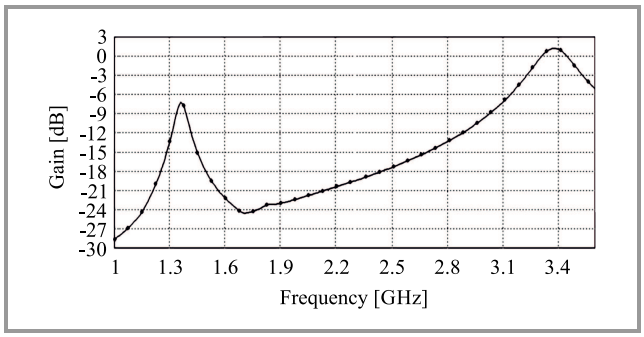


Fig. 11. Maximum gain vs. frequency plot for sample H-shaped RMSA.

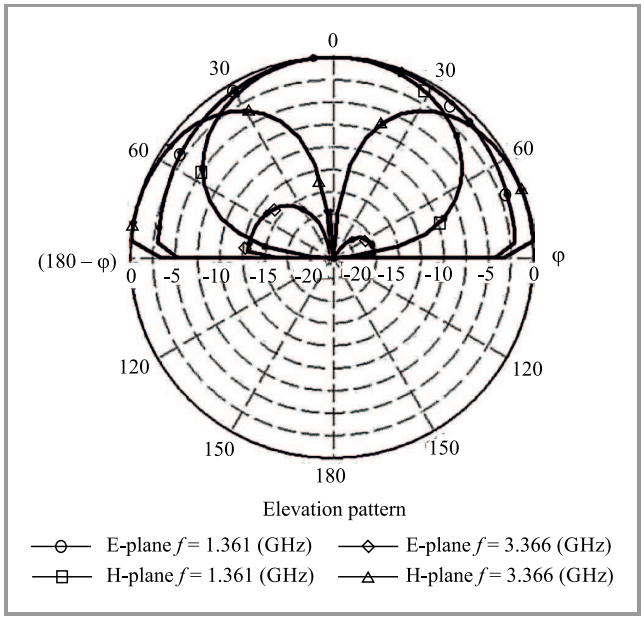


Fig. 12. Radiation pattern for sample H-shaped RMSA.

number of samples, in comparison to examples existing literature [40].

4. Design Specification and Dataset

Design specifications of the proposed dual-band H-shaped compact RMSA with an infinite ground plane are shown in Fig. 13, and the list of parameters used in the design

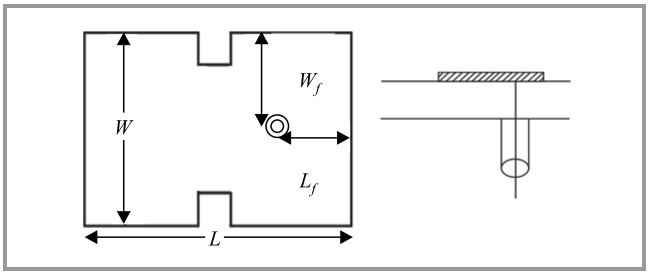


Fig. 13. Top and side view of H-shaped RMSA used for modeling.

is presented in Table 4. Along the radiating edge, two rectangular notches that are equal in dimensions are cut (upper, lower and centered in the middle), starting with the notch dimensions of 0.2×0.5 mm, increased to 20×20.25 mm in fixed steps of 0.25 mm in L and W. 80 samples were collected using IE3D simulation software.

Table 4
Specification of H-shaped RMSA

Parameter	Value
Operating frequency	2.33 GHz
Patch length	$L = 32$ mm
Patch width	$W = 42$ mm
Substrate	FR-4
Substrate thickness	1.6 mm
Dielectric constant	4.4
Loss tangent	0.025
Feed type	Coaxial
Notch length	0.25 to 20 mm
Notch width	0.5 to 20.25 mm
Feed location (L_f, W_f)	6.575 mm, 21 mm

5. Conclusion

Machine learning is one of the methods used to design antennas quickly. The proposed model is simple, time-efficient and does not require any complex mathematical calculations. It may easily predict, with high accuracy, the physical or electrical parameters within the range of data provided for training, for a dual-band (0.75 to 2.20 GHz and 3.0 to 3.44 GHz), compact microstrip antenna that may be used in a variety of wireless applications.

References

[1] M. Fallahpour and R. Zoughi, "Antenna miniaturization techniques: A review of topology- and material-based methods", *IEEE Antenn. and Propag. Maga.*, vol. 60, no 1, pp. 38–50, 2018 (DOI: 10.1109/MAP.2017.2774138).

[2] P. L. Chi, R. Waterhouse, and T. Itoh, "Antenna miniaturization using slow wave enhancement factor from loaded transmission line models", *IEEE Trans. on Antenn. and Propag.*, vol. 59, no. 1, pp. 48–57, 2011 (DOI: 10.1109/TAP.2010.2090452).

[3] A. Foudazi, H. R. Hassani, and S. M. Ali Nezhad, "Small UWB planar monopole antenna with added GPS/GSM/WLAN bands", *IEEE Trans. on Antenn. and Propag.*, vol. 60, no. 6, pp. 2987–2992, 2012 (DOI: 10.1109/TAP.2012.2194632).

[4] V. K. Sambhe, R. N. Awale, and A. Wagh, "Compact U-shape radiating patch with rectangular ground planar monopole antenna", *The J. of Engin.*, vol. 2015, no. 2, pp. 54–58, 2015 (DOI: 10.1049/joe.2014.0269).

[5] X. Ren, S. Gao, and Y. Yin, "Compact tri-band monopole antenna with hybrid strips for WLAN/WiMAX applications", *Microw. and Opt. Technol. Lett.*, vol. 57, no. 1, pp. 94–99, 2015 (DOI: 10.1002/mop.28785).

[6] M.-T. Tan, B.-Z. Wang, and Z.-M. Zhang, "Compact tri-band dual-polarized planar monopole antenna with asymmetrical ground plane and loaded stub", *Progr. in Electromag. Res. Lett*., vol. 58, pp. 89–95 2016 (DOI: 10.2528/PIERL15122004).

[7] Y. Mao, S. Guo, and M. Chen, "Compact dual-band monopole antenna with defected ground plane for Internet of Things", *IET Microw., Antenn. Propag.*, vol. 12, no. 8, pp. 1332–1338, 2018 (DOI: 10.1049/iet-map.2017.0860).

[8] R. O. Ouedraogo, E. J. Rothwell, A. R. Diaz, K. Fuchi, and A. Temme, "Miniaturization of patch antennas using a metamaterial-inspired technique", *IEEE Trans. on Antenn. and Propag.*, vol. 60, no. 5, pp. 2175–2182, 2012 (DOI: 10.1109/TAP.2012.2189699).

[9] Y. Dong and T. Itoh, "Metamaterial-based antennas", *Proceedings of the IEEE*, vol. 100, no. 7, pp. 2271–2285, 2012 (DOI: 10.1109/JPROC.2012.2187631).

[10] A. Bazrkar, A. Gudarzi, and M. Mahzoon, "Miniaturization of rectangular patch antennas partially loaded with μ-negative metamaterials", in *Proc. Int. Conf. on Electron., Biomed. Engin. and its Appl. ICEBEA'2012*, Dubai, United Arab Emirates, 2012, pp. 289–292 [Online]. Available: http://psrcentre.org/images/extraimages/58.%200112256.pdf

[11] F. Raval, Y. P. Kosta, J. Makwana, and A. V. Patel, "Design implementation of reduced size microstrip patch antenna with metamaterial defected ground plane", in *Proc. of Int. Conf. on Commun. and Sig. Process.*, Melmaruvathur, India, 2013, pp. 186–190 (DOI: 10.1109/iccsp.2013.6577040).

[12] R. A. H. Mahdi and S. M. R. Taha, "Miniaturization of rectangular microstrip patch antenna using topology optimized metamaterial", *IEICE Electron. Express*, vol. 14, no. 19, 2017 (DOI: 10.1587/elex.14.20170787).

[13] V. G. Ajay and T. Mathew, "Size reduction of microstrip patch antenna through metamaterial approach for WiMAX application", in *Proc. of Int. Conf. on Wirel. Commun., Sig. Process. and Network. WiSPNET 2217*, Chennai, India, 2017, pp. 379–381 (DOI: 10.1109/WiSPNET.2017.8299782).

[14] J. Chatterjee, A. Mohan, and V. Dixit, "Broadband circularly polarized H-shaped patch antenna using reactive impedance surface", *IEEE Antenn. and Wirel. Propag. Lett.*, vol. 17, no. 4, pp. 625–628, 2018 (DOI: 10.1109/LAWP.2018.2806993).

[15] B. Gupta, S. Nakhate, and M. Shandilya, "A compact wideband microstrip patch antenna with defected ground plane", in *Proc. of 4th World Congr. on Inform. and Commun. Technol. WICT 2014*, Bandar Hilir, Malaysia, 2014, pp. 51–56 (DOI: 10.1109/WICT.2014.7077301).

[16] M. K. Khandelwal, B. K. Kanaujia, S. Dwari, S. Kumar, and A. Gautam, "Analysis and design of dual band compact stacked microstrip patch antenna with defected ground structure for WLAN/WiMAX applications", *AEU – Int. J. of Electron. and Commun.*, vol. 69, no. 1, pp. 39–47, 2015 (DOI: 10.1016/j.aeue.2014.07.018).

[17] S. Veisee, S. Asadi, and M. Hedayati, "A novel compact defected ground structure and its application in mutual coupling reduction of a microstrip antenna", *Turkish J. of Elec. Engin. and Comp. Sci.*, vol. 24, no. 5, pp. 3664–3670, 2016 (DOI: 10.3906/elk-1404-517).

[18] P. Anitha, A. S. R. Reddy, and M. N. G. Prasad, "Design of a compact ψ-shaped micro-strip patch antenna on modified ground planes", in *Proc. of Int. Conf. on Recent Adv. in Electron. and Commun. Technol. ICRAECT 2017*, Bangalore, India, 2017, pp. 96–99 (DOI: 10.1109/ICRAECT.2017.60).

[19] C. Sharma and V. D. Kumar, "Koch fractalized compact spiral antenna based on Fibonacci sequence", in *Proc. of IEEE 5th Asia-Pacific Conf. on Antenn. and Propag. APCAP 2016*, Kaohsiung, Taiwan, 2016, pp. 157–158 (DOI: 10.1109/APCAP.2016.7843146).

[20] F. Wang, F. Bin, Q. Sun, J. Fan, and H. Ye, "A compact UHF antenna based on complementary fractal technique", *IEEE Access*, vol. 5, pp. 21118–21125, 2017 (DOI: 10.1109/ACCESS.2017.2756672).

[21] Y. K. Choukiker and J. C. Mudiganti, "Compact hybrid fractal antenna for wideband wireless applications", *Int. J. of Microw. and Wirel. Technol.*, vol. 9, no. 5, pp. 1191–1196, 2017 (DOI: 10.1017/S1759078716001318).

[22] C. Deng, Y. Li, Z. Feng, W. Li, and D. Zhang, "A dual-band circularly-polarized stacked patch antenna with a small frequency-ratio", in *Proc. IEEE Antenn. and Propag. Soc. Int. Symp. (APSURSI 2013*, Orlando, FL, USA, 2013, pp. 942–943 (DOI: 10.1109/APS.2013.6711130).

[23] A. Foroozesh, D. Psychoudakis, J. Homer, and I. Kim, "Compact wideband stacked microstrip patch antenna for a medical application", in *2016 IEEE Int. Symp. on Antenn. and Propag. APSURSI 2016*, Fajardo, Puerto Rico, 2016, pp. 283–284 (DOI: 10.1109/APS.2016.7695850).

[24] A. Katyal and A. Basu, "Compact and broadband stacked microstrip patch antenna for target scanning applications", *IEEE Antenn. and Wirel. Propag. Lett.*, vol. 16, pp. 381–384, 2017 (DOI: 10.1109/LAWP.2016.2578723).

[25] Y. Gou, S. Yang, Q. Zhu, and Z. Nie, "A compact dual-polarized double E-shaped patch antenna with high isolation", *IEEE Trans. on Antenn. and Propag.*, vol. 61, no. 8, pp. 4349–4353, 2013 (DOI: 10.1109/TAP.2013.2262664).

[26] C. K. Hsu and S. J. Chung, "Compact antenna with U-shaped open-end slot structure for multi-band handset applications", *IEEE Trans. on Antenn. and Propag.*, vol. 62, no. 2, pp. 929–932, 2014 (DOI: 10.1109/TAP.2013.2289996).

[27] M. Tarbouch, A. E. Amri, and H. Terchoune, "Compact CPW-fed microstrip octagonal patch antenna with h slot for WLAN and WiMAX applications", in *Proc. of Int. Conf. on Wirel. Technol., Embedd. and Intell. Syst. WITS 2017*, Fez, Morocco, 2017 (DOI: 10.1109/WITS.2017.7934638).

[28] A. Roy, S. Bhunia, D. C. Sarkar, and P. P. Sarkar, "Slot loaded compact microstrip patch antenna for dual band operation", *Progress in Electromag. Res. C*, vol. 73, pp. 145–156, 2017 (DOI: 10.2528/PIERC17020903).

[29] M. T. Islam, M. Samsuzzaman, and M. Z. Mahmud, "A compact slotted patch antenna for breast tumor detection", *Microw. and Opt. Technol. Lett*., vol. 60, no. 7, pp. 1600–1608, 2018 (DOI: 10.1002/mop.31215).

[30] Y. Li, Q. Liu, Y. Chen, C. Li, Z. Mo, and F. Li, "A compact triple wideband-notched UWB antenna", in *Proc. Int. Worksh. on Antenna Technol. iWAT 2018*, Nanjing, China, 2018 (DOI: 10.1109/IWAT.2018.8379142).

[31] B. Choudhury, S. Thomas, and R. M. Jha, "Implementation of soft computing optimization techniques in antenna engineering [Antenna applications corner]", *IEEE Antenn. and Propag. Mag.*, vol. 57, no. 6, pp. 122–131, 2015 (DOI: 10.1109/MAP.2015.2439612).

[32] Z. Cai, S. Zeng, Y. Y. Li, and L. Kang, "Automated antenna design using normalized steady state genetic algorithm", in *Proc. of NASA/ESA Conf. on Adapt. Hardware and Syst.*, Noordwijk, Netherlands, 2008 (DOI: 10.1109/AHS.2008.26).

[33] S. Ledesma, J. Ruiz-Pinales, G. Cerda-Villafaña, and M. G. Garcia-Hernandez, "A hybrid method to design wire antennas: Design and optimization of antennas using artificial intelligence", *IEEE Antenn. and Propag. Mag.*, vol. 57, no. 4, pp. 23–31, 2015 (DOI: 10.1109/MAP.2015.2453912).

[34] L. Feruglio, "Artificial intelligence for small satellites mission autonomy", Doctoral Dissertation, Politecnico di Torino, 2017 (DOI: 10.6092/polito/porto/2694565).

[35] L. Y. Xiao, W. Shao, F. L. Jin, and B. Z. Wang, "Multiparameter modeling with ANN for antenna design", *IEEE Trans. on Antenn. and Propag.*, vol. 66, no. 7, pp. 3718–3723, 2018 (DOI: 10.1109/TAP.2018.2823775).

[36] A. Landi, P. Piaggi, M. Laurino, and D. Menicucci, "Artificial neural networks for nonlinear regression and classification", in *Proc. of 10th In. Conf. on Intell. Syst. Design and Appl.*, Cairo, Egypt, 2010, pp. 115–120 (DOI: 10.1109/ISDA.2010.5687280).

[37] S. Lek *et al*., "Application of neural networks to modelling nonlinear relationships in ecology", *Ecological Modell*., vol. 90, no. 1, 1996, pp. 39–52 (DOI: 10.1016/0304-3800(95)00142-5).

[38] C. G. Christodoulou and A. Patnaik, "Neural networks for antennas", in *Morden Antenna Handbook*, C. A. Balanis, Ed. Wiley, 2008, pp. 1625–1657 (ISBN: 9780470036341).

[39] M. Kayri, "Predictive abilities of Bayesian regularization and Levenberg-Marquardt algorithms in artificial neural networks: A comparative empirical study on social data", *Mathem. and Comput. Appl.*, vol. 21, no. 2, 2016 (DOI: 10.3390/mca21020020).

[40] V. S. Kushwah and G. S. Tomar, "Design and analysis of microstrip patch antennas using artificial neural network", in *Trends in Research on Microstrip Antennas*, S. Chattopadhyay, Ed. Rijeka: IntechOpen, 2017 (DOI: 10.5772/intechopen.69522).



Kanhaiya Sharma received his M.Tech. in Computer Science and Engineering from Jawahar Lal Nehru Technological University (JNTU) Hyderabad, Andhra Pradesh, India in 2011, and currently pursuing a Ph.D. in Computer Science and Engineering from the School of Technology, Pandit Deendayal Petroleum University since 2017. He worked as an Assistant Professor at Sri Sarathi Institute of Engineering and Technology, Nuzvid, Krishna District, Andhra Pradesh, India from October 2006 to May 2010, and at Siddhant group of institution, Maharashtra, India from 2010 to 2017. His current research interests include design and development of low-cost solutions for wireless applications, machine learning, wireless communications, microstrip antenna design, filter design, and artificial intelligence.
E-mail: sharmakanhaiya@gmail.com
Department of Computer Science and Engineering
School of Technology
Pandit Deendayal Petroleum University
Gandhinagar, Gujarat 382007, India



Ganga Prasad Pandey received his B.Tech. in Electronics and Communication Engineering from KNIT, Sultanpur, UP, an M.E. from Delhi College of Engineering (now Delhi Technological University), Delhi and a Ph.D. from Uttarakhand Technical University, Dehradun. He worked as an Assistant Professor at the Maharaja Agasen Institute of Technology, Delhi, from 2002 to 2016. He joined Pandit Deendayal Petroleum University, Gandhinagar in 2016, working at the ICT Department. His current research interests include machine learning in antennas, energy harvesting, ME-dipoles, active, reconfigurable, frequency-agile microstrip antennas, as well as microwave/millimeter wave integrated circuits and devices.
E-mail: Gangaprasad.Pandey@sot.pdpu.ac.in
Department of Information and Communication Technology
School of Technology
Pandit Deendayal Petroleum University
Gandhinagar, Gujarat 382007, India

2020 IEEE International Conference for Innovation in Technology (INOCON)
Bengaluru, India. Nov 6-8, 2020

Detection of Lung Cancer using Machine Learning Techniques Based on Routine Blood Indices

Puneet
M. Tech Scholar,
Department of Information Technology
Delhi Technological University
Delhi-110042, India
puneetbisarwal@gmail.com

Anamika Chauhan
Assistant Professor,
Department of Information Technology
Delhi Technological University
Delhi-110042, India
anamika@dce.ac.in

***Abstract*— Cancer occurs due to the development of a large number of abnormal cells that causes uncontrolled growth and division of abnormal cells. The purpose of this research is to predict lung cancer using machine learning techniques based on routine blood indices and have better predictive results. We have used machine learning techniques such as *XGBoost*, *Logistic Regression, SVM, Gaussian Naive Bayes, Decision Tree* and *KNN* with *GridSearchCV*. Various *scikit-learn* algorithms have been used for feature selection, and only those features are selected that are important for our model. We found that *XGBoost* using *GridSearchCV* classification model is most suited for this work.**

Index Terms*— Cancer, Abnormal Cells, Classification Techniques, Lung Cancer Classification, Machine Learning, Machine Learning Techniques

I. INTRODUCTION

During the past few decades, there has been steady progress in Cancer Research [1]. Every year around 15 million new patients are diagnosed with different cancer subtypes. Prediction of cancer subtypes in the early stages can help the doctors to give proper medication and treatment to the patients. If cancer stages are accurately predicted early, then there will be chances to cure the patients. Developing countries and underdeveloped countries are the worst-hit countries due to the lack of proper medical facilities because of which they have high mortality rates as compared to developed countries.

Ignorance of the symptoms of the cancer disease and not seeking proper medical consultation in the early stages or seeking medical consultation late leads to a higher mortality rate. Also, the lack of doctors and medical equipment's in rural areas adds up to the problem in diagnosing the cancer disease in early-stage, leading to a higher mortality rate.

Machine Learning methods are currently being used in medical fields for identification and classification of various diseases. Recently Machine Learning methods have been employed in this area in predicting the cancer in early stages, which can result in better chances of survival of patients.

Doctors and Scientists have applied various approaches, like screening, identification and classification, to detect various forms of cancer at early stage even before they cause any symptoms. Also, scientists developed various new methods for predicting the outcome of treating the cancer in early stage [2]. With the emergence of modern medical technologies, a large amount of cancer data were collected and released for medical research. However, predicting a disease outcome precisely is one of the most challenging tasks for doctors.

This has resulted in Machine Learning (ML) methods becoming a tool best in medical research. These ML methods can be used in identifying patterns and connections for complex datasets and can also help in predicting future results of different types of cancer.

In the past, lung cancer has been detected, classified and predicted using various ML techniques. Wu, Jiangpeng et al. [5] identified lung cancer using Random Forest, Integrity of cell-free DNA's had been used by Leng et al. [3] for identification of lung cancer patients, Fradkin et al. [4] had used SVM, and logistic regression, Lynch et al. [8] application of unsupervised methods to lung cancer, Dimitoglou et al. [7] Predicted lung cancer using Naïve Bayes classifier, Lynch et al. [6] used SVM, Decision Tree, Linear Regression, Ensemble, Random Forest and Generalized Boosting Machines in predicting lung cancer in patients.

In this research work various classification methods like *XGBoost*, *Logistic Regression, Support Vector Machine (SVM), Gaussian Naïve Bayes*, *K-Nearest Neighbors (KNN)* and *Decision Tree* using *GridSearchCV* was used, and we compared the accuracy of different approaches and analyzed the performance output.

The structure of this paper is as follows. In Section 2, the data source is discussed. From where the data is collected, how many tuples are there in the dataset, and all the important details are discussed. Section 3, presents details regarding Features Selection, how the features were selected, and the algorithms that were used to collect features are also mentioned. Section 4 presents the methodology. Section 5 presents the results and experimental details collected while doing research. In Section 6, we have discussed the conclusion.

II. DATA SOURCE

The dataset used has been taken from Wu, Jiangpeng et al. [5] which was collected from Lanzhou University in Second Hospital, China. A total, 277 patients are part of this dataset, with 49 types of different blood indices being present in the dataset for each patient.

Out of 277 patients, 183 patients were lung cancer positive. These patients were diagnosed by using tissue biopsies. The remaining 94 patients were diagnosed as without lung cancer. These patients include males and females, with age group in the range of 20 to 81 years. These patients are

978-1-7281-9744-9/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:07:06 UTC from IEEE Xplore. Restrictions apply.

found to be at different cancer stage. Smoking status of each patient was also mentioned in the dataset. The dataset contains label as one of the columns in which 1 represents patients diagnosed with Lung Cancer while -1 represents patients not diagnosed with Lung Cancer.

	Platelet distribution width	Basophil ratio	Albumin	Age	Large platelet ratio	Lymphocyte ratio	ALB/GLB	Neutrophile granulocyte ratio	White blood cell	Creatine kinase isoenzymes
count	277.000000	277.000000	277.000000	277.000000	277.000000	277.000000	277.000000	277.000000	277.000000	277.000000
mean	13.407942	0.003383	39.752708	56.862816	27.350181	0.240296	1.224296	0.671097	6.663466	20.754152
std	3.329155	0.004651	5.211571	13.243646	11.084908	0.102024	0.341425	0.113945	3.006356	32.663335
min	0.000000	0.000000	24.100000	20.000000	0.000000	0.030000	0.590000	0.270000	2.180000	3.000000
25%	10.700000	0.000000	36.100000	50.000000	19.200000	0.170000	0.980000	0.600000	4.790000	11.000000
50%	14.000000	0.000000	40.100000	59.000000	26.000000	0.230000	1.180000	0.660000	5.820000	14.000000
75%	16.200000	0.009000	43.100000	66.000000	33.800000	0.310000	1.380000	0.740000	7.740000	21.000000
max	20.400000	0.020000	51.600000	81.000000	61.200000	0.600000	2.660000	0.930000	24.730000	383.000000

Fig. 1. Dataset Description

Out of the 94 patients, 51 patients had tuberculosis and they were specifically included, because the use of CT scans to differentiate lung cancer from tuberculosis has a high false-positive rate. The other negative patients went for a routine examination and were not diagnosed with any diseases linked to lung tissues. Figure 1 shows the description of the dataset (includes count, mean, std, min, 25%, 50%, 75%, max). The dataset [5] were distributed into the Training set and Test set. The training set includes 226 patients out of which 153 were Lung Cancer patients, 37 tuberculosis patients and 36 were of other diseases. The test set includes 51 patients.

A. Exploratory Data Analysis (EDA):

EDA is understanding what dataset is using basic statistical tools, plotting, visualizing data to understand our dataset before we actually go and start machine learning.

EDA is used for analyzing the data using statistical analysis and visualization tools to understand the data and use visualization tools.

1) Univariate EDA

In Univariate Exploratory EDA, we take into consideration only one variable or feature at a time. It was be applied to the top 5 features of the dataset [5] obtained by using *SelectKBest method*. In Univariate Analysis, we graphed Histogram and Probability Density Function (PDF) plot.

A Histogram is the graphical representation of the distribution of numerical data. Probability Density Function (PDF) represents the continuous version of the histogram, where the numerical data estimate the continuous curve.

2) Bivariate EDA

In Bivariate EDA, we take into consideration only two variables or features at a time. Due to overlapping, we were not able to separate different classes in the Univariate Analysis. Therefore we applied this Bivariate EDA.

In Bivariate Analysis, we graphed Pair Plot Scatter graph. Pair Plot Scatter graph as the name suggests, it takes two variables or features in pairwise to plot the scatter graph.

In this 2-D pair-plot graph, we will be getting P (4, 2) = 12 scattered graphs as we have four dimensions here, in which six graphs will be unique. In the 2-D plot graph, we were also not able to make observations and differentiate between different classes as you can see in Figure 2 Pair Plot graph. In the Figure 2 the diagonal graphs are probability density function of the corresponding features.

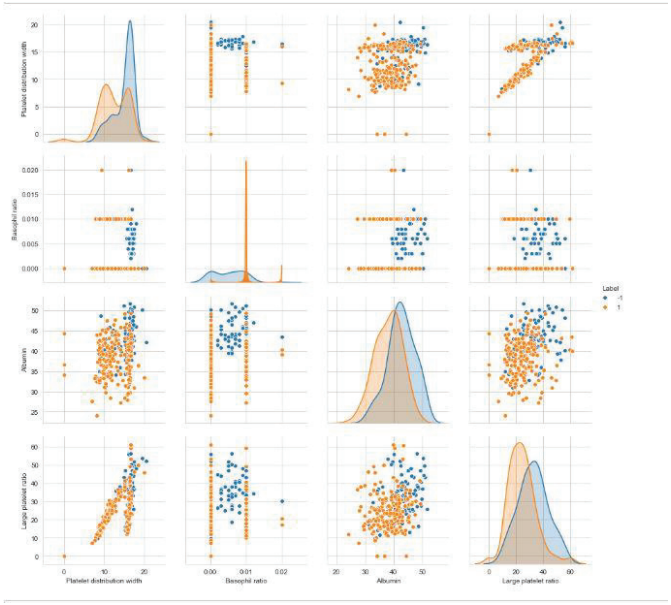


Fig. 2. Pair Plot of Platelet distribution width, Basophil ratio, Albumin, Large platelet ratio

III. FEATURE SELECTION

Features are the descriptive attribute or columns in our data-frame or dataset that helps in differentiating of one or more data element into the respective class label.

Selection of features should be carried out in a way that provides maximum information to our model for training. Features should not change with external factors if not meant to.

We have selected our features using these methods to gain maximum knowledge.

1 *SelectKBest method*: It is one of the algorithms in *Scikit-learn* library for python language. In *SelectKBest algorithm,* it selects the K features having the highest scores value in the dataset. Chi2 was used as a scoring function for *SelectKBest* for our model.

2 Using the Tree-based method: It is also one of the predominantly used algorithms in *scikit-learn* library for python language. It uses various techniques to improve the performance of the model.

Decision Tree Classifier: It is a classification algorithm which works on the concept of making decisions as per importance. It splits the attributes to get to the output-variable attribute. *Decision Trees* are used solve classification and regression problems. Overfitting and Underfitting problems also occurs in tree-based methods. Overfitting occurs when the deep tree performs perfectly in the training dataset, but it fails during validation while in underfitting, a model fails to capture essential patterns in the dataset leading to poor performance in the training dataset.

ExtraTreeClassifier: ExtraTreesClassifier is a type of ensemble learning technique. It is also known as extremely randomized trees classifier. In this classifier multiple de-correlated decision trees are created and the aggregates of their results is used as classification result

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:07:06 UTC from IEEE Xplore. Restrictions apply.

TABLE I. FEATURES RANKED FROM EXTRATREECLASSIFIER METHOD.

RANK	*INDEX*	*REFERENCE RANGE*
1	Age	20-81
2	ALB/GLB	0.590-2.660
3	Glucose	2.590-27.060
4	Aspartate aminotransferase	7.000-136.000
5	Uric Acid	86.000-821.000
6	Alanine transaminase	3.000-177.000
7	AST/ALT	0.230-6.250

TABLE II. TOP RANKING BLOOD INDICES FEATURES (WU, JIANGPENG ET AL. 2019)

RANK	*INDEX*	*REFERENCE RANGE*
1	Basophil ratio	0.00-0.01
2	Creatine kinase isoenzymes (U/L)	0.0-25.0
3	Platelet large cell ratio (%)	17.0-45.0
4	Albumin (g/L)	30.0-55.0
5	Platelet distribution width (fl)	9.0-17.0
6	Neutrophilic granulocytes (10^9/L)	2.00-7.00
7	White blood cell count (10^9/L)	4.00-10.00
8	Albumin Globulin ratio	1.10-2.50
9	Monocytes (10^9/L)	0.12-1.20
10	Monocyte ratio	0.03-0.08
11	Lymphocyte ratio	0.20-0.40
12	Neutrophile granulocyte ratio	0.50-0.70
13	Lactate dehydrogenase (U/L)	0.0-240.0
14	Carbamide (mmol/L)	1.80-8.00
15	Eosinophil cells (10^9/L)	0.02-0.50
16	Mean corpuscular volume (fl)	80.0-100.0
17	Alkaline phosphatase (U/L)	0.0-120.0
18	Mean corpuscular hemoglobin (pg)	27.0-34.0
19	Creatine kinase (U/L)	0-195

In Table 1 and Table 2, there are 26 features; these 26 features was used in our classification model. Table 1, features were ranked using ExtraTreeClassifier and Table 2, features were selected from (Wu, Jiangpeng et al. 2019); a total of 26 features selected for our final model. These selected features helped us in achieving better performance parameters as compared to (Wu, Jiangpeng et al. 2019).

IV. METHODOLOGY

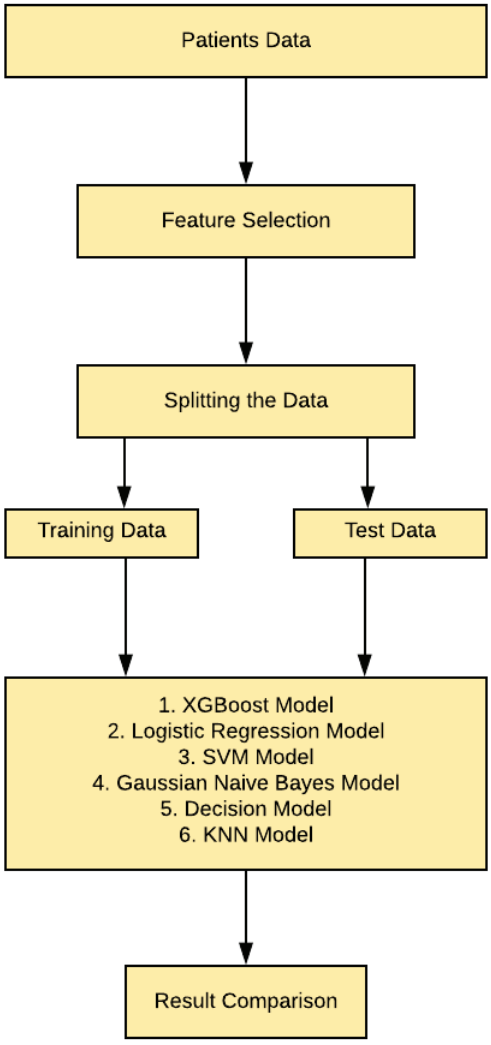


Fig. 3. Block Diagram of Methodology

We have used various classification method for this dataset such *XGBoost, Logistic Regression, SVM, Gaussian Naive Bayes, KNN* and *Decision Tree* using *GridSearchCV*. We have analyzed the results for all classification methods, and we found that *XGBoost* gave the best results.

The *XGBoost* classification method was used here for the construction of the final model of classification. The two main advantages of using *XGBoos*t method is the speed and performance of the model. *XGBoost* method used to solve problems related to ranking, regression, user-defined prediction and classification. In research work, we first applied exhaustive *XGBoost* hyperparameter tuning using *GridSearchCV*. *GridSearchCV* is a traditional way to perform hyperparameter tuning on any classification method. Using *scikit-learn GridSearchCV* algorithm, we defined our grid of parameters for our classification method and ran the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:07:06 UTC from IEEE Xplore. Restrictions apply.

grid search. After that, we applied 10-fold Cross-validation on our models.

The indicators that were used to test the performance of Improved Lung Cancer Detection (ILCD) model includes Sensitivity (True Positive Rate), Specificity (True Negative Rate), False Negative Rate (FNR) or Miss Rate, Accuracy, area under the curve (AUC) is the area under the receiver operating characteristic (ROC) curve, Matthews correlation coefficient (MCC), True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN).

- Sensitivity = TP / (TP+FN)
- Specificity = TN / (TN+FP)
- False Negative Rate = FN / (TP+FN)
- Accuracy = (TP+TN) / (TP+FP+FN+TN)
- MCC = (TP*TN – FP*FN) /√((TP + FN) ∗ (TN + FP) ∗ (TN + FN))

As per the accuracy matrix, *XGBoost* method is better. FNR for the method is low, which is good in our application domain, the weightage of FNR is quite high. The FNR is the ratio of the person who is having the disease, but the Test results say Negative. It is also known as Miss Rate. The closer its value is to 0, the better the model will be. The AUC is the area under the ROC curve, and its value can range from [0, 1]. The closer the AUC of ROC curve is to 1, the better the prediction performance of lung cancer.

Better parameters for the medical domain is FNR and Sensitivity or TPR. If we try to improve or increase Sensitivity or TPR, then it means we are decreasing FNR. FNR should be minimized because if it is high then it means the chances of classifying a lung cancer patient as not having cancer is high. This could be fatal for that patient. That's why minimizing FNR or maximizing Sensitivity, or TPR is essential.

The steps taken in the methodology are summarized below:

1. Split the dataset into training and test sets.
2. Select the features as discussed in Section III (feature selection).
3. Apply the class weight in the classification models wherever applicable as the dataset is unbalanced and do the hyperparameter tuning using *GridSearchCV*.
4. Create model using the best parameters.
5. Apply 10-fold Cross Validation on the model.
6. Find the accuracy and confusion matrix. From the confusion matrix, we will be getting TP, TN, FP and FN.

V. Result

The performance of various classification method on the lung cancer dataset have been compared in Table 4. The performance of the ILCD model using XGBoost achieved better results as compared to the routine blood indices model for lung cancer (RBLC) Model [5].

The performance matrix scores were all greater than 85% and accuracy, and AUC of ROC curve score was greater than 90%, while the FNR score was only less than 4%. The performance matrix scores are given below:

Sensitivity = 96.67%, Specificity = 85.71%, False Negative Rate = 3.33%, Accuracy = 92.16%, MCC = 83.86%, AUC = 95%.

TABLE III. False Negative Rate (FNR) of Models

MACHINE LEARNING TECHNIQUES	FALSE NEGATIVE RATE (FNR)
RBLC MODEL [5]	14.29%
ILCD MODEL	3.33%

TABLE IV. Performance of Models

MACHINE LEARNING TECHNIQUES	TRAIN ACCURACY	TEST ACCURACY
RBLC MODEL [5]	-	88.24%
ILCD MODEL	99.01%	92.16%
LOGISTIC REGRESSION MODEL	87.71%	86.27%
SVM MODEL	88.67%	84.31%
GAUSSIAN NAÏVE BAYES MODEL	76.60%	72.55%
DECISION TREE MODEL	82.55%	76.47%
KNN MODEL	100%	74.51%

TABLE V. AUC-ROC Score

MACHINE LEARNING TECHNIQUES	AUC-ROC SCORE
RBLC MODEL [5]	0.9016
ILCD MODEL	0.95
LOGISTIC REGRESSION MODEL	0.91
SVM MODEL	0.91
GAUSSIAN NAÏVE BAYES MODEL	0.81
DECISION TREE MODEL	0.80
KNN MODEL	0.77

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:07:06 UTC from IEEE Xplore. Restrictions apply.

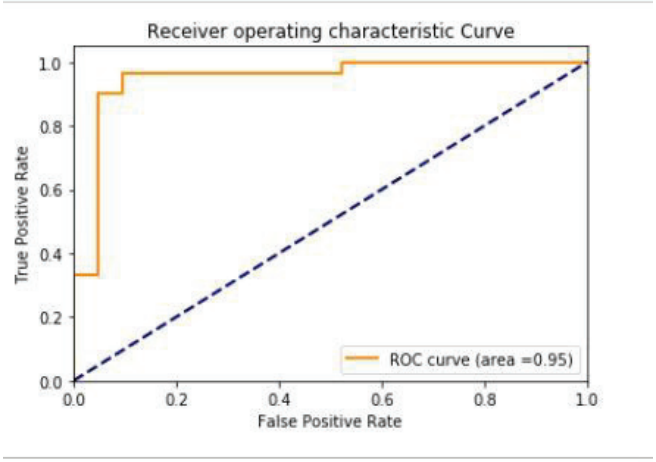


Fig. 4. ROC curve

In Table 4 performance comparison of RBLC model [5] with all the models that were applied is shown. The FNR of XGBoost Model is one order of magnitude less than RBLC model [5] as seen in Table 3. The AUC-ROC Curve score of RBLC model [5] with other models is compared in Table 5. In the Figure 4, AUC of ROC curve is 95% for the test data. It is much better for XGBoost model as compared to the RBLC model [5] whose AUC of ROC curve was 90%.

VI. CONCLUSION

Of all the methods that were applied, *XGBoost* proved to be the best in terms of Accuracy, Sensitivity, FNR, MCC, AUC of ROC curve. All the above results show that Machine Learning techniques can be helpful in identifying lung cancer. These techniques may help doctors to detect lung cancer and may be confirmed by going through diagnostic tests which can help in saving the precious life of a human being.

REFERENCES

[1] Hanahan, Douglas, and Robert A. Weinberg. "Hallmarks of cancer: the next generation." cell 144, no. 5 (2011): 646-674.

[2] Kourou, Konstantina, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, and Dimitrios I. Fotiadis. "Machine learning applications in cancer prognosis and prediction." Computational and structural biotechnology journal 13 (2015): 8-17.

[3] Leng, Shaoyi, Jianjun Zheng, Yinhua Jin, Hongbin Zhang, Ya Zhu, Jing Wu, Yan Xu, and Puhong Zhang. "Plasma cell-free DNA level and its integrity as biomarkers to distinguish non-small cell lung cancer from tuberculosis." Clinica Chimica Acta 477 (2018): 160-165.

[4] Fradkin, Dmitriy, Dona Schneider, and Ilya Muchnik. "Machine learning methods in the analysis of lung cancer survival data." DIMACS Technical Report 2005–35 (2006).

[5] Wu, Jiangpeng, Xiangyi Zan, Liping Gao, Jianhong Zhao, Jing Fan, Hengxue Shi, Yixin Wan, E. Yu, Shuyan Li, and Xiaodong Xie. "A Machine Learning Method for Identifying Lung Cancer Based on Routine Blood Indices: Qualitative Feasibility Study." JMIR medical informatics 7, no. 3 (2019): e13476.

[6] Lynch, Chip M., Behnaz Abdollahi, Joshua D. Fuqua, R. Alexandra, James A. Bartholomai, Rayeanne N. Balgemann, Victor H. van Berkel, and Hermann B. Frieboes. "Prediction of lung cancer patient survival via supervised machine learning classification techniques." International journal of medical informatics 108 (2017): 1-8.

[7] Dimitoglou, George, James A. Adams, and Carol M. Jim. "Comparison of the C4. 5 and a Naïve Bayes classifier for the prediction of lung cancer survivability." arXiv preprint arXiv:1206.1121 (2012).

[8] Lynch, Chip M., Victor H. van Berkel, and Hermann B. Frieboes. "Application of unsupervised analysis techniques to lung cancer patient data." PloS one 12, no. 9 (2017).

[9] Agrawal, Ankit, Sanchit Misra, Ramanathan Narayanan, Lalith Polepeddi, and Alok Choudhary. "A lung cancer outcome calculator using ensemble data mining on SEER data." In Proceedings of the tenth international workshop on data mining in bioinformatics, pp. 1-9. 2011.

[10] Szpechcinski, A., J. Chorostowska-Wynimko, R. Struniawski, W. Kupis, P. Rudzinski, R. Langfort, E. Puscinska, P. Bielen, P. Sliwinski, and T. Orlowski. "Cell-free DNA levels in plasma of patients with non-small-cell lung cancer and inflammatory lung disease." British journal of cancer 113, no. 3 (2015): 476-483.

[11] Dheda, K., and C. E. Barry 3rd. "3rd, Maartens G." Lancet 387 (2016): 1211-26.

[12] Zumla, Alimuddin, Andrew George, Virendra Sharma, Rt Hon Nick Herbert, Aaron Oxley, and Matt Oliver. "The WHO 2014 global tuberculosis report—further to go." The Lancet Global Health 3, no. 1 (2015): e10-e12.

[13] Travis, William D., Elisabeth Brambilla, Andrew G. Nicholson, Yasushi Yatabe, John HM Austin, Mary Beth Beasley, Lucian R. Chirieac et al. "The 2015 World Health Organization classification of lung tumors: impact of genetic, clinical and radiologic advances since the 2004 classification." Journal of thoracic oncology 10, no. 9 (2015): 1243-1260.

[14] Singh, Vikas Kumar, Subhash Chandra, Sachin Kumar, Ghanshyam Pangtey, Anant Mohan, and Randeep Guleria. "A common medical error: lung cancer misdiagnosed as sputum negative tuberculosis." Asian Pac J Cancer Prev 10, no. 3 (2009): 335-8.

[15] Esposito, Angela, Carmen Criscitiello, Dario Trapani, and Giuseppe Curigliano. "The emerging role of "Liquid Biopsies," circulating tumor cells, and circulating cell-free tumor dna in lung cancer diagnosis and identification of resistance mutations." Current oncology reports 19, no. 1 (2017): 1.

[16] Paci, Massimiliano, Sally Maramotti, Enrica Bellesia, Debora Formisano, Laura Albertazzi, Tommaso Ricchetti, Guglielmo Ferrari et al. "Circulating plasma DNA as diagnostic biomarker in non-small cell lung cancer." Lung cancer 64, no. 1 (2009): 92-97.

[17] Dimitoglou, George, James A. Adams, and Carol M. Jim. "Comparison of the C4. 5 and a Naïve Bayes classifier for the prediction of lung cancer survivability." arXiv preprint arXiv:1206.1121 (2012).

[18] Yoon, Kyong-Ah, Sohee Park, Sang Hee Lee, Jin Hee Kim, and Jin Soo Lee. "Comparison of circulating plasma DNA levels between lung cancer patients and healthy controls." The Journal of Molecular Diagnostics 11, no. 3 (2009): 182-185.

[19] Lum, Caroline, and Muhammad Alamgeer. "Technological and Therapeutic Advances in Advanced Small Cell Lung Cancer." Cancers 11, no. 10 (2019): 1570.

[20] van der Drift, Miep A., Bernard EA Hol, Corné HW Klaassen, Clemens FM Prinsen, Yvonne AWG van Aarssen, Rogier Donders, Jos WJ van der Stappen, PN Richard Dekhuijzen, Henricus FM van der Heijden, and Frederik BJM Thunnissen. "Circulating DNA is a non-invasive prognostic factor for survival in non-small cell lung cancer." Lung Cancer 68, no. 2 (2010): 283-287.

[21] Jiang, Tao, Changyun Zhai, Chunxia Su, Shengxiang Ren, and Caicun Zhou. "The diagnostic value of circulating cell free DNA quantification in non-small cell lung cancer: A systematic review with meta-analysis." Lung Cancer 100 (2016): 63-70.

[22] Gautschi, Oliver, Colette Bigosch, Barbara Huegli, Monika Jermann, Arthur Marx, Eveline Chassé, Daniel Ratschiller et al. "Circulating deoxyribonucleic acid as prognostic marker in non–small-cell lung cancer patients undergoing chemotherapy." Journal of Clinical Oncology 22, no. 20 (2004): 4157-4164.

[23] Kumar, Sachin, Randeep Guleria, Vikas Singh, Alok C. Bharti, Anant Mohan, and Bhudev C. Das. "Efficacy of circulating plasma DNA as a diagnostic tool for advanced non-small cell lung cancer and its predictive utility for survival and response to chemotherapy." Lung Cancer 70, no. 2 (2010): 211-217.

[24] Chiappetta, Caterina, Marco Anile, Martina Leopizzi, Federico Venuta, and Carlo Della Rocca. "Use of a new generation of capillary electrophoresis to quantify circulating free DNA in non-small cell lung cancer." Clinica Chimica Acta 425 (2013): 93-96.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:07:06 UTC from IEEE Xplore. Restrictions apply.

[25] Wang, Brant G., Han-Yao Huang, Yu-Chi Chen, Robert E. Bristow, Keyanunoosh Kassauei, Chih-Chien Cheng, Richard Roden, Lori J. Sokoll, Daniel W. Chan, and Ie-Ming Shih. "Increased plasma DNA integrity in cancer patients." Cancer research 63, no. 14 (2003): 3966-3968.

[26] Shen, Shu Yi, Rajat Singhania, Gordon Fehringer, Ankur Chakravarthy, Michael HA Roehrl, Dianne Chadwick, Philip C. Zuzarte et al. "Sensitive tumour detection and classification using plasma cell-free DNA methylomes." Nature 563, no. 7732 (2018): 579-583.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:07:06 UTC from IEEE Xplore. Restrictions apply.

Proceedings of the Third International Conference on Intelligent Sustainable Systems [ICISS 2020]
IEEE Xplore Part Number: CFP20M19-ART; ISBN: 978-1-7281-7089-3

Dog Breed Classifier for Facial Recognition using Convolutional Neural Networks

Bickey Kumar shah
Department of CSE
Delhi Technological University
Delhi, India
bickeyshah721@gmail.com

Aman Kumar
Department of CSE
Delhi Technological University
Delhi, India
amandeepnha@gmail.com

Amrit Kumar
Department of CSE
Delhi Technological University
Delhi, India
amritx8@gmail.com

2020 3rd International Conference on Intelligent Sustainable Systems (ICISS) | 978-1-7281-7089-3/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICISS49785.2020.9315871

***Abstract*— This paper dealt with the breed classification of dogs. To classify dog breed is a challenging part under a deep convolutional neural network. A set of sample images of a breed of dogs and humans are used to classify and learn the features of the breed. The images are converted to a single label of dimension with image processing. The images of human beings and dogs are considered for breed classification to find the existing percentage of features in humans of dogs and dogs of human. This research work has used principal component analysis to shorten the most similar features into one group to make an easy study of the features into the deep neural networks. And, the facial features are stored in a vector form. This prepared vector will be compared with each feature of the dog into the database and will give the most efficient result. In the proposed experiment, 13233 human images and 8351 dog images are taken into consideration. The images under test are classified as a breed with the minimum weight between test and train images. This paper is based on research work that classifies different dogs breed using CNN. If the image of a dog is supplied then the algorithm will work for finding the breed of dog and features similarity in the breed and if the human image is supplied it determines the facial features existing in a dog of human and vice-versa.**

Keywords*— *Breed, Feature, Images, Neural Network, Principal Component Analysis.

I. INTRODUCTION

Machine learning has been used widely these days. From small prediction models to creating high-end predictive models, machine learning has been used widely in fields ranging from business analytics to healthcare improvement [14-16]. Convolutional Neural networks (CNN) is one of the main achievements of the evolution of Machine learning and deep learning [2]. CNN constitutes the different neurons with different learning biases and weights. In this, every neuron receives some inputs and make some dot product or some mathematical manipulation and gives some outputs as desired. CNN architecture differs from other neural networks. The architecture of CNN helps to reduce the features and make the defined network to increase accuracy [4]. The neuron here is the single layers and different from the previous layers. CNN is carried out in three steps:

1. The convolutional layer of CNN

The images are generally in the form of a large matrix and the pixel intensity ranges from 0 to 256. The image is converted into 2D matrics whose entry has weight and it is used to recognize the features. Generally, a square matrix of size 4*4 or 3*3 will be considered.

2. The pooling layer of CNN

Pooling is a technique that is used to reduce the size of the image when the image is too large [11-12]. The parameters of the image are generally reduced to the size of the image by using the pooling method. Max pooling is the method that is frequently used and it gives the result as same as the original image. Other types of pooling are L2 norm and average pooling. The size of the output image is controlled by three Parameters:

- The number of filters
- Stride
- Zero Padding

3. The output layer of CNN

After passing through the Convolutional layer and Pooling layer the compressed image is obtained with a lot of features in it [13]. The size of the output image is calculated using the formula ((W-F+2P)/S) + 1.

Where W is input volume Size.
F is filter Size.
P is the number of padding applied
S is the number of strides used.

Features are extracted with the help of the convolutional layers. The architecture of CNN can be divided for the dog breed classification as followed:

- Convolutional Layer n°1 with 96 filters
  - Max pooling
  - Relu
- Convolutional Layer n°2 with 256 filters
  - Max pooling
  - Relu
- Convolutional Layer n°3 with 384 filters
  - Max pooling
  - Relu
  - DropOut
- Flatten Layer
- Fully Connected Layer with 500 nodes
  - Relu
  - DropOut
- Fully Connected Layer with n nodes (n = number of breeds)

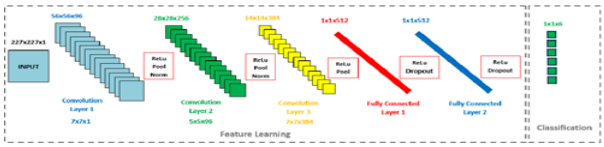


Fig.1 CNN Architecture

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 27,2021 at 10:28:03 UTC from IEEE Xplore. Restrictions apply.

We have different mathematical series and terms associated with the convolutional neural networks. For working with Convolutional Neural Networks, the Activation functions in CNN are the basic requirements for non-linear processing [1]. The activation function is used to scale the output values and if required sometimes converts the high value to probability for detection of most probable images [6]. It decides that the upcoming activation states of the neurons that it should be activated or not and it is done by the sum of the bias and weights. Without the activation function model is only the linear model and it won't bring efficiency for the non-linear calculations. the activation functions are as follows:

1. **Sigmoid Function:** The mathematical formula it uses is f(z)=1/(1+exp(-z). The value ranges from 0 to 1.

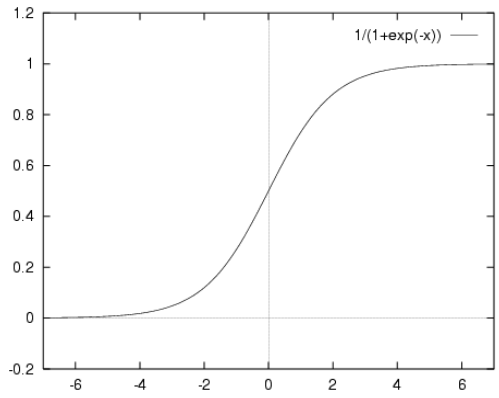


Fig.2 Sigmoid Function

2. **Tanh (Hyperbolic Function):** The value of this ranges from -1 to 1. The mathematical formula used for this function is f(z)=(1-exp(-2z))/(1+exp(-2z)).

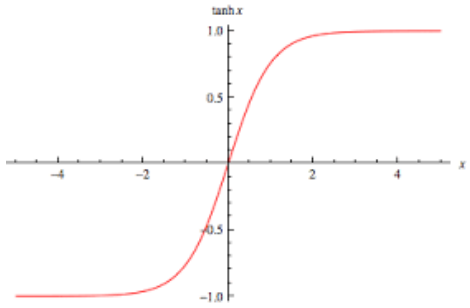


Fig.3 Tanh function

3. **ReLU**- Rectified linear Units:The mathematical formula for this is R(z)= maximum(0,z) i.e if z<0, R(z)=0 and if z>0 R(z)=a.

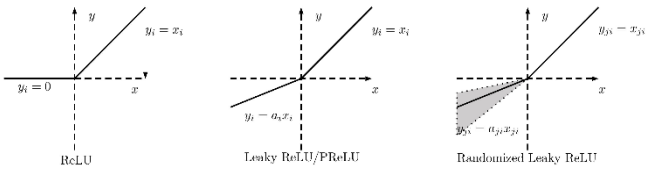


Fig.4 ReLU Function

II. METHODOLOGY

Computer vision is used for the image processing work to convert the images into black and white images [3]. In this we revolve around the CNN algorithm. The data is preprocessed and after that classified into the group of training data, testing data, and validation data. Training data is used for finding patterns of the images using deep neural networks. for the adjustment of error in the weights, Feedforward and backpropagation work are carried out over the networks.

A. *Transfer Learning*

It is a popular approach to deep learning. Pre-trained models are used at the starting points and in these cases Transfer learning is considered to be best [7-9]. To solve any type of problem we generally need a model of a similar type and transfer learning plays important role in it. To make a fine model through the transfer learning we modify the learning rates, go through the features extraction, use the architecture of the pre-trained model, and train some layers and freeze others [8]. It is an optimization technique. The learning rates from the pre-trained model should be kept in mind while performing transfer learning.

B. *Import Datasets*

We need to load the required datasets and all the required library that is important for the computation work. We will import all the required libraries and then perform the processing parts using TensorFlow. We will use the TensorFlow framework of version 2.3.0.

C. *Resize and Normalize the data*

All the images do not have the same shapes and for preparing the model we need to resize all the images into one dimension of square shape. We normalize our dataset by dividing all the pixels by 255.

D. *Preprocessings Images, CNN and Models*

We perform some more data pre-processing and features the reduction of large data using the principal components analysis and then proceed for the efficient model prediction works.

```
# Downloading the pre-trained face detector
import os
if not os.path.isdir('haarcascades'):
    os.mkdir('haarcascades')
!wget -cq -O haarcascades/haarcascade_frontalfa
clear_output()
print("Downloaded Successfully")
```

```
Downloaded Successfully
```

Fig. 5. Pre-trained face detector algorithm

We download the pre-trained face detector with the computational transfer learning algorithms taking the learning rates into considerations. It helps to detect the face that whether it belongs to the human or Dog and then after it again find the similarities to predict the breed.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 27,2021 at 10:28:03 UTC from IEEE Xplore. Restrictions apply.

```
human_files_short = human_files[:100]
dog_files_short = dog_files[:100]

homosapiens, canis = 0, 0
out = display(progress(0, 100), display_id=True)
for il, (human, dog) in enumerate(zip(human_files_short, dog_files_short)):
  out.update(progress(il+1, 100))
  if face_detector(human):
    homosapiens += 1
  if face_detector(dog):
    canis += 1

print("Human detection accuracy: {}%".format(100.* float(homosapiens) / len(human_files_short)))
print("Dogs detected as human: {}%".format(100. * float(canis) / len(dog_files_short)))
```

```
Human detection accuracy: 99.0%
Dogs detected as human: 11.0%
```

Fig. 6 Algorithm to find Human face.

This algorithm finds the number of the face detected i.e. it specifies whether the algorithm can detect or not that the facial accuracy of the human and dogs. This is for human face detection and it gives the accuracy of 99% when input images of humans are given and for the same when dog image is input it gives 11%. Since the algorithm is designed for humans it implies that our accuracy is correct. And, similarly, we have the algorithm for dog face detection which gives the accuracy of 100% when a dog face is an input and 0% when the face of a human is input. The same algorithm over different computations helps to classify breeds and further we give the image input of humans that finds the prediction of the dog breed facial match of a human face. We transform the training, validation, and testing set. We define the CNN architecture and then convolutional layers with increasing numbers of filters in each layer and max-pooling after 3 layers of value 2 [10]. We perform the non-linear task and dropout value of the layer to be 2. We give the feedforward algorithm by adding the sequence of a convolutional and max-pooling layer. We perform the ReLU activation function over the first and second hidden layers. We predict the loss function and make the system more optimized with the minimum occurrence of loss. We initialize the variable for training and testing loss predictor and also trackers for the minimum loss and preparation of the models efficiently. We train the model and clear the gradients of all optimized variables. We then work on validating the model. We update the average validation loss and calculate the batch loss. After this, we train the model and transform the training, validation, and testing sets. Further computations are made with the learning rates and image classifications.

III. EXPERIMENTAL RESULT

We performed the Convolutional Neural network experiment with a lot of images of humans and dogs. We come across the result that the total number of the dog images sample was classified into 103 dog breed images. We were able to classify the breed of the dog and finally, we detected the probability of features of dog breed that match the human features of a particular sample. The experiment was performed over a lot of images and data, distinctly some of the obtained results are: -

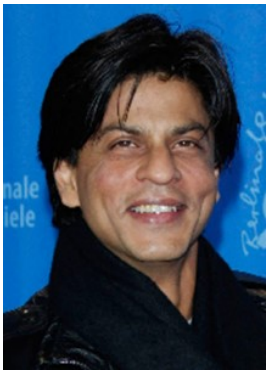


Fig. 7 sample Image1

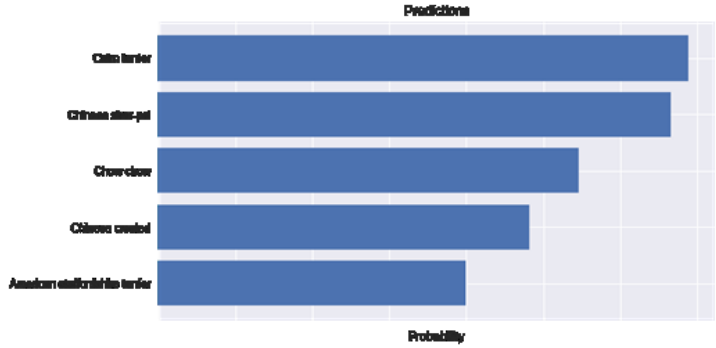


Fig. 7.1 sample Image1 result

In fig. 7 It is shown that when a human sample is given as an input, the output result shows that the human sample image has a shadow of five different dog breeds in it. The maximum probability that the human seems like is 'Cairn terrier' breed, after that the probability goes max to 'Chinese shar-pei' and so on. It signifies that the facial expression of the human in the sample matches maximum with the cairn terrier and least with the American Staffordshire terrier.



Fig. 8 sample image2

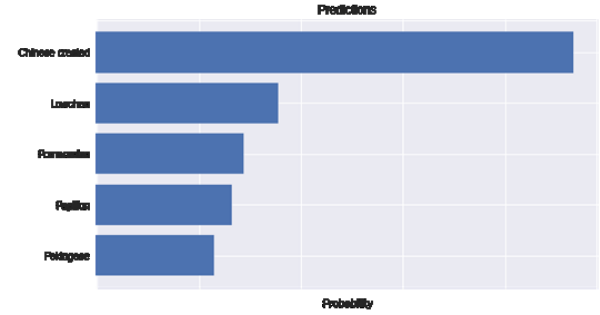


Fig.8.1 sample image2 result

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 27,2021 at 10:28:03 UTC from IEEE Xplore. Restrictions apply.

In fig. 8, It is similar to that of Fig. 7, that when a human sample is given as an input, the output result shows that the human sample image has a shadow of the top five different dog breeds in it. The maximum probability that the human seems like is a Chinese crested dog breed, after that the probability goes max to Lowchen and so on. It signifies that the facial expression of the human in the sample matches maximum with the Chinese crested and least with the Pekingese.



Fig. 9 sample image3

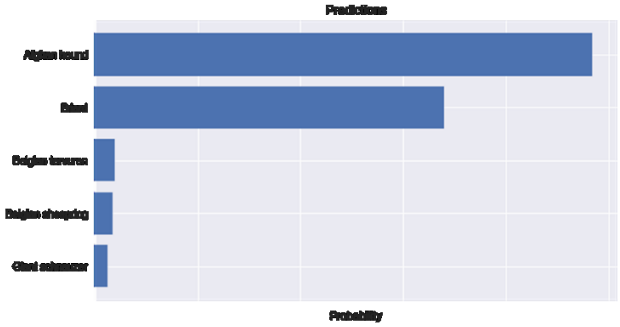


Fig. 9.1 sample image3 result

In fig. 9 It is similar to that of Fig. 7, that when a human sample is given as an input, the output result shows that the human sample image has a shadow of the top five different dog breeds in it. The maximum probability that the human seems like is Afgan hound dog breed, after that the probability goes max to briard and so on. The facial expression of human has negligible matching with Belgian Tervuren and Giants Schnauzer. It signifies that the facial expression of the human in the sample matches maximum with the Chinese crested and least with the Giants Schnauzer.



Fig. 10 sample image4

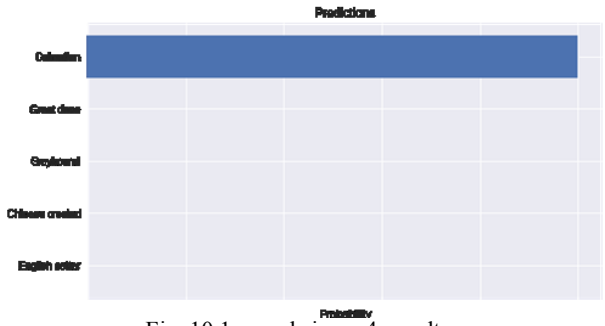


Fig. 10.1 sample image4 result

In Fig. 10 the sample images of the dog give full probability prediction with Dalmation breed, the sample dog is really of dalmation breed that is why it has shown full prediction accuracy, this sample verifies that the working model is almost accurate. This sample has not given any errors or features match with other dog breed samples.



Fig. 11 sample image5

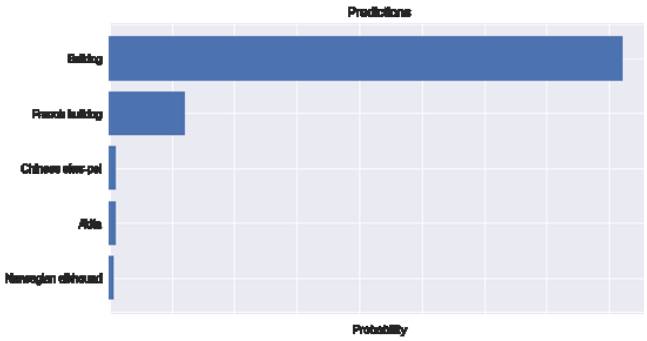


Fig. 11.1 sample image5 result



Fig. 12. Sample image6

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 27,2021 at 10:28:03 UTC from IEEE Xplore. Restrictions apply.

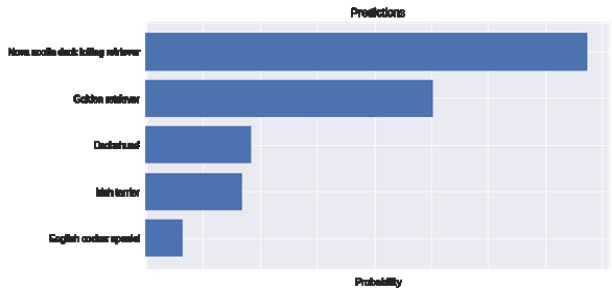


Fig. 12.1 Sample image6 result

We have similar cases in Fig.11 and Fig.12 as that of fig.10. This is the breed identifying for the given dog image. In fig. 9 it is a real Bulldog image and it has shown almost the greatest accuracy with bulldog breed and almost negligible match with the rest of the breeds. In Fig.10 the image is of a toy dog and It has shown a similar breed property as humans in the above image. All of the samples have a different result and almost accurate results on practical ideas and theory. It signifies that the models are nearly accurate. The results of the data truly depend on the type of data that we use. For the Deep learning analysis, we need millions of data for getting more and more accurate results. We have obtained the following model accuracy with the epoch.

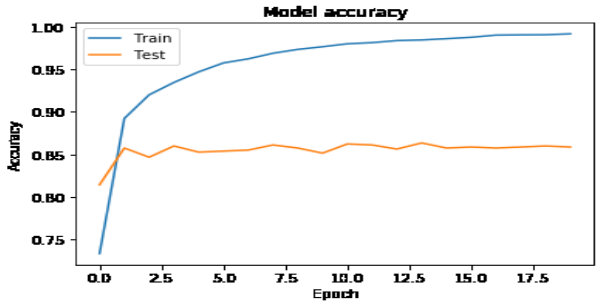


Fig13. Accuracy Vs Epoch

We have also bear some model loss that were significantly optimized with the learning rate and loss variable detection over every epoch. A snap of model loss and epoch is as:

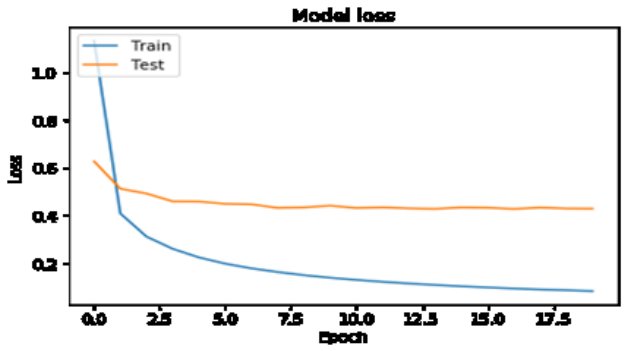


Fig. 14 Model loss VS Epoch

This paper presents that CNN is a good approach for the dog breed classification work. The experiment shows that the CNN algorithm has an accuracy of 95.5%. This signifies that CNN can be used for dog breed classification.

IV. CONCLUSION AND FUTURE SCOPE

Convolutional Neural network seems to be efficient for the dog breed classification. This learning has shown that multiple algorithms can be processed with different parameters for efficient learning and classification of dog breed. Newer Technology and better computation power are developed and more study work can be done during the course time for better classification. The algorithm can be made further efficient with the deep learning models, and the breed classification can be performed further with deep neural networks by structuring similar algorithms and better computation methods.

ACKNOWLEDGMENT

The Authors are indebted to Central Library, Delhi Technological University, New Delhi for providing the access to all the materials that were needed for the study.

REFERENCES

[1] P. Road, "Face Recognition Based Dog Breed Classification Using Coarse-to-Fine Concept and PCA Massinee Chanvichitkul , Pinit Kumhom , and Kosin Chamnongthai Department of Electronics and Telecommunication Engineering Faculty of Engineering King Mongkut's Univers," *System*, pp. 5–9, 2007.

[2] A. Ghimire, S. Thapa, A. K. Jha, A. Kumar, A. Kumar, and S. Adhikari, "AI and IoT Solutions for Tackling COVID-19 Pandemic", in 2020 International Conference on Electronics, Communication and Aerospace Technology, 2020: IEEE

[3] P. Borwarnginn, K. Thongkanchorn, S. Kanchanapreechakorn, and W. Kusakunniran, "Breakthrough Conventional Based Approach for Dog Breed Classification Using CNN with Transfer Learning," *2019 11th Int. Conf. Inf. Technol. Electr. Eng. ICITEE 2019*, vol. 7, pp. 3–7, 2019, doi: 10.1109/ICITEED.2019.8929955.

[4] P. Prasong and K. Chamnongthai, "Face-recognition-based dog-breed classification using size and position of each local part, and PCA," *2012 9th Int. Conf. Electr. Eng. Comput. Telecommun. Inf. Technol. ECTI-CON 2012*, pp. 3–7, 2012, doi: 10.1109/ECTICon.2012.6254212.

[5] A. K. Jha, S. Adhikari, S. Thapa, A. Kumar, A. Kumar, and S. Mishra, "Evaluation of Factors Affecting Compressive Strength of Concrete using Machine Learning" in 2020 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA), 2020: IEEE, pp. 70-74, doi: 10.1109/ACCTHPA49271.2020.9213199.

[6] J. Krause, T. Gebru, J. Deng, L. J. Li, and F. F. Li, "Learning features and parts for fine-grained recognition," *Proc. - Int. Conf. Pattern Recognit.*, pp. 26–33, 2014, doi: 10.1109/ICPR.2014.15.

[7] A. Ghimire, S. Thapa, A. K. Jha, S. Adhikari, and A. Kumar, "Accelerating Business Growth with Big Data and Artificial Intelligence", in 2020 Fourth International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2020: IEEE, pp. 441-448, doi: 10.1109/I-SMAC49090.2020.9243318.

[8] C. D. Casuat, A. S. Alon, J. L. Dioses, R. M. Dellosa, and R. A. Diamante, "RFID controlled 'GG' pieces ranking detection with watch-dog enable," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 1 Special Issue 3, pp. 274–279, 2020, doi: 10.30534/ijatcse/2020/4191.32020.

[9] R. O. Sinnott, F. Wu, and W. Chen, "A Mobile Application for Dog Breed Detection and Recognition Based on Deep Learning," *Proc. - 5th IEEE/ACM Int. Conf. Big Data Comput. Appl. Technol. BDCAT 2018*, pp. 87–96, 2019, doi: 10.1109/BDCAT.2018.00019.

[10] J. J. Cooper *et al.*, "Necrotizing meningoencephalitis in atypical dog breeds: A case series and literature review," *J. Vet. Intern.*

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 27,2021 at 10:28:03 UTC from IEEE Xplore. Restrictions apply.

Med., vol. 28, no. 1, pp. 198–203, 2014, doi: 10.1111/jvim.12233.

[11] C. Marín Navas, F. J. Navas González, V. Castillo López, L. Payeras Capellà, M. Gómez Fernández, and J. V. Delgado Bermejo, "Impact of breeding for coat and spotting patterns on the population structure and genetic diversity of an islander endangered dog breed," *Res. Vet. Sci.*, vol. 131, no. April, pp. 117–130, 2020, doi: 10.1016/j.rvsc.2020.04.003.

[12] R. Kumar, M. Sharma, K. Dhawale, and G. Singal, "Identification of Dog Breeds Using Deep Learning," *Proc. 2019 IEEE 9th Int. Conf. Adv. Comput. IACC 2019*, pp. 193–198, 2019, doi: 10.1109/IACC48062.2019.8971604.

[13] M. V. S. Rishita and T. A. Harris, "Dog breed classifier using convolutional neural networks," 2018 Int. Conf. Networking, Embed. Wirel. Syst. ICNEWS 2018 - Proc., 2018, doi: 10.1109/ICNEWS.2018.8903980.

[14] S. Thapa, P. Singh, D. K. Jain, N. Bharill, A. Gupta, and M. Prasad, "Data-Driven Approach based on Feature Selection Technique for Early Diagnosis of Alzheimer's Disease", in 2020 International Joint Conference on Neural Networks (IJCNN), 2020: IEEE, pp. 1-8, doi: 10.1109/IJCNN48605.2020.9207359.

[15] S. Adhikari, S. Thapa, and B. K. Shah, "Oversampling based Classifiers for Categorization of Radar Returns from the Ionosphere," in 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020: IEEE, pp. 975-978, doi: 10.1109/ICESC48915.2020.9155833.

[16] S. Thapa, S. Adhikari, A. Ghimire, and A. Aditya, "Feature Selection Based Twin-Support Vector Machine for the diagnosis of Parkinson's Disease", in 2020 8th R10 Humanitarian Technology Conference (R10-HTC), 2020: IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 27,2021 at 10:28:03 UTC from IEEE Xplore. Restrictions apply.

Silicon
https://doi.org/10.1007/s12633-020-00921-4

ORIGINAL PAPER

Dual Metal Gate Dielectric Engineered Dopant Segregated Schottky Barrier MOSFET With Reduction in Ambipolar Current

Sumit Kale [1] · Madduri Sai Chandu [2]

Received: 19 October 2020 / Revised: 19 October 2020 / Accepted: 22 December 2020
© Springer Nature B.V. 2021

Abstract
In this paper, to solve the problem of higher ambipolar leakage current ($I_{ambipolar}$) of Dielectric Engineered (DE) Dopant Segregated (DG) Schottky Barrier (SB) MOSFET (DE DS SBMOS), we have incorporated dual metal gate (DMG) in place of single metal gate for the DE DS SBMOS. The proposed device is named as DMG DE DS SBMOS. In a proposed device, the gate is consisting of dual metal having different work functions. Therefore, the gate is divided in two parts named as auxiliary gate (AG) and tunneling gate (TG). AG and TG employed near the source and drain (S/D) ends, respectively. AG is mainly used to control the on-state performance and TG for the off-state performance by modulating the SB height and width at S/D end, respectively. Simulation results show that $I_{ambipolar}$ of the device has been successfully suppressed by choosing appropriate work functions for AG and TG without affecting the on-state current (I_{on}) of the device. The proposed device combines the advantage of dual metal gate and dielectric engineering to achieve improvement in the overall performance of the device. The proposed device exhibits high I_{on}, low $I_{ambipolar}$ and low subthreshold swing (SS). The simulations study verifies the use of the proposed device for low power applications.

Keywords Schottky Barrier · Dielectric Engineered · Ambipolar current · Dual metal gate · Dopant Segregation

1 Introduction

Over the past two decades, SBMOS is under consideration by several device research communities as an alternative to replace conventional MOSFET in applications such as mixed mode, radio frequency (RF), and low power applications [1–12]. The interest towards SBMOS technology spread due to its inherent advantages over their conventional counterparts. SBMOS main features includes atomically abrupt junction, low source/drain parasitic resistances, simpler fabrication, and, immunity to short channel effects [11–13]. In past, several works reported on the SBMOS analysis, investigation and performance improvement for DC and high frequency. The performance of the device has been improved in terms of high on-state current (I_{on}), low off-state current (I_{off}), and, low subthreshold swing (SS). In addition, enhancement in the device performance parameters for analog/RF, low power, and mixed mode applications. Our group also reported few works based on the SBMOS design, investigation and performance improvement for the applications like analog/RF, low power and biosensor [4–7, 14–19]. In our recent reported work [19], we have proposed and investigated a DE DS SBMOS with NiSi Source/Drain. The device reported in the above work achieved improvement in the DC and analog/RF performance parameters such as I_{on}, on to off current ratio (I_{ON}/I_{OFF}), SS, transconductance (g_m), transconductance generation factor (g_m/I_D), cut-off frequency (f_T), and gain-bandwidth product (f_A). However, irrespective of their several advantages, DE DS SBMOS suffers from large $I_{ambipolar}$ due to SB tunneling and thermionic emission mainly at the drain side. The large $I_{ambipolar}$ of DE DS SBMOS restricts the applications of the device for mixed mode and low power applications.

To overcome the aforementioned problem, in this work, we have suppressed the $I_{ambipolar}$ of the DE DS SBMOS without affecting the on-state current of the device. For this purpose,

✉ Sumit Kale
sumitkale785@gmail.com

[1] Department of Electronics and Communication Engineering Delhi Technological University Delhi 110042, India

[2] Department of Electronics and Communication Engineering Madanapalle Institute of Technology & Science Madanapalle, AP 517325, India

Published online: 06 January 2021

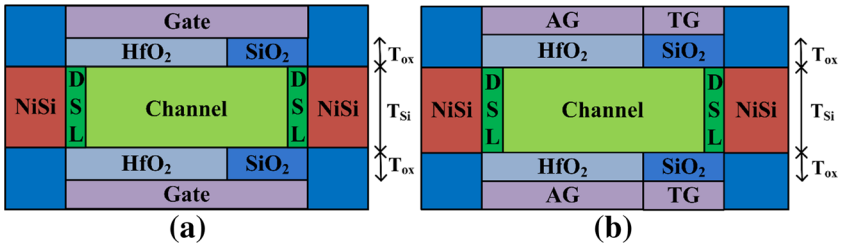


Fig. 1 The cross-sectional diagram of (**a**) DE DS SBMOS and, (**b**) the proposed DMG DE DS SBMOS

we have employed dual metal gate in place of single metal gate in DE DS SBMOS. The dual metal gates are named as Auxiliary Gate (AG) and Tunneling Gate (TG). The AG is employed near the source end and TG is employed near the drain end. The AG gate is used to control the on-state performance and TG gate is used to control the off-state performance of the device. In this study, we have investigated the on-state and off-state performance of the device using appropriate work function of metal to form AG and TG. The proposed device achieves significant reduction in $I_{ambipolar}$ as compared to DE DS SBMOS.

2 Device Structure and Simulation Setup

Figure 1 shows the cross-sectional diagram of DE DS SBMOS and the proposed DMG DE DS SBMOS. The common device design parameters are: Si channel thickness (T_{si}) = 10 nm, gate oxide thickness (T_{ox}) = 2 nm, gate length = 40 nm and channel doping concentration (N_{ch}) (p-type) = 1×10^{15} cm^{-3}, Dopant Segregation Layer (DSL) length = 2 nm and DSL doping concentration (n-type) = 1×10^{20} cm^{-3}. In addition, for the S/D contacts formation, NiSi having electron SBH of 0.65 eV and hole SBH of 0.47 eV is employed [20]. Moreover, the gate dielectric is divided in to two parts for both the devices. The high-k gate dielectric (HfO_2) is used at the source side and low-k gate dielectric (SiO_2) at the drain side. The dielectric permittivity of HfO_2 is 22 and SiO_2 is 3.9. In addition, to form dual metal gate in place of single metal gate of the proposed device, the gate is divided in two parts named as auxiliary gate (AG) and tunneling gate (TG) having different metal work function. In proposed device, all the dimensions, channel material, and S/D silicide, and gate oxides materials are kept same as employed in DE DS SBMOS.

All the simulations have been performed using SILVACO ATLAS two-dimensional device simulator [21]. For the calculation of Schottky tunneling current, universal Schottky tunneling (UST) model has been included [19, 22–24]. Also, for thermionic emission current, thermionic emission model has been included. For inclusion of barrier lowering effect, image force barrier lowering model included [25–28]. In addition, BGN, SRH, and CVT models have been also included. The effective masses for electron and hole are chosen as 0.32 m_0 and 0.54 m_0, respectively. Simultaneously, the effective Richardson's constants for electrons and holes are taken as 30 and 110 A/Kcm^2. The calibration of simulation model has been performed by reported work in [1] to verify the simulations performed for the device as presented in [19].

3 Result and Discussion

In this work, we have analyzed and investigated the effect of dual metal gate on the DC performance of the DE DS

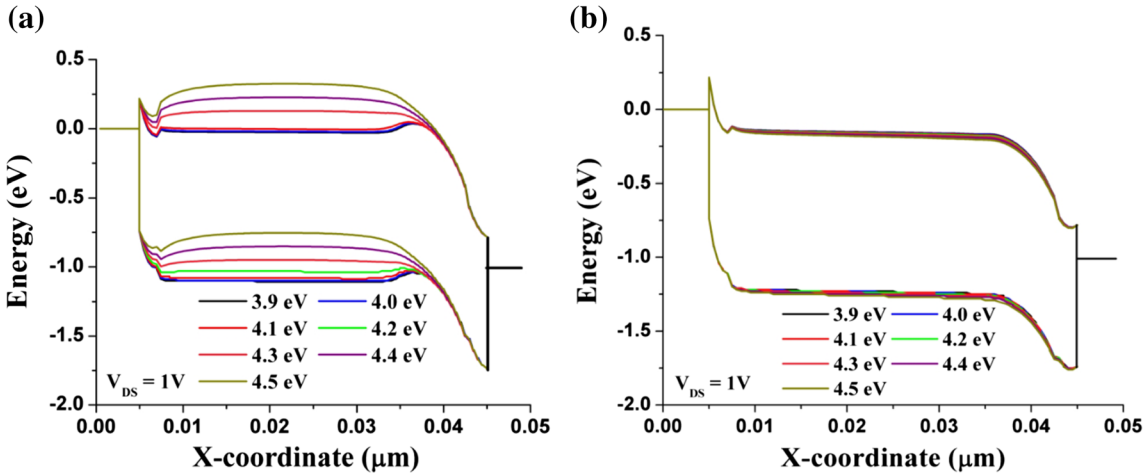


Fig. 2 Energy band diagram of the proposed device for different AG work function with gate length of 30 nm and fixed value of TG work function of 3.9 eV having gate length of 10 nm in (**a**) Off-state condition, (**b**) On-state condition

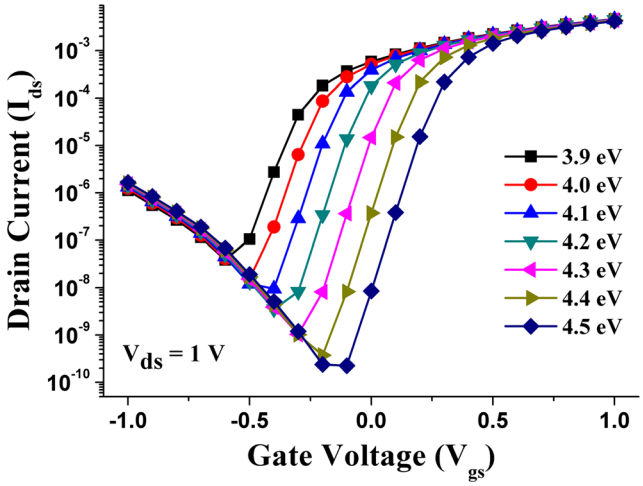


Fig. 3 Transfer characteristics of the device for different AG work function with gate length of 30 nm and fixed value of TG work function of 3.9 eV having gate length of 10 nm

SBMOS. In proposed device, by varying the work function of dual metal gate the on and off-state performance of the device has been investigated. To find the effect of dual metal gate on the device performance, extensive simulation has been performed. In addition, to optimize the performance of the device appropriate length and work function has been chosen for the metals used for dual metal gate formation. In this study, to understand the behavior of the device by employment of dual metal gate in place of single metal gate, first, we have to analyze the energy band diagram of the proposed device in the on-state and off-state condition for different work function of AG and TG.

First, to observe the effect of dual metal gate on DC performance of the device, we have employed the metals having different work function for formation of AG and metal having fixed work function for TG. Therefore, the energy band diagram in the off-state condition of the device for metals having different work function with gate length of 30 nm for AG, and metal having fixed work function of 3.9 eV with gate length of 10 nm for TG is shown in Fig. 2a. It can be observed from the figure that for increasing value of AG work function, the conduction band edge of the energy band diagram shifts in upward direction. As a result, the increase in potential barrier occurs for electron flowing from source to channel. Consequently, the thermionic emission of electron reduces due to large barrier present between source and channel. This result in decrease in the I_{off} is obtained for the device. In addition, the subthrshold swing of the device is also improved for increasing work function of AG. On the other hand, there is no change observed in SB height and width at the drain end for different work function of AG. Therefore, AG doesn't have control over the drain end. The major component of $I_{off,}$ due to flow of electron from source to channel caused by thermionic emission is controlled by AG.

Figure 2b shows the on-state energy band diagram of the device for different metal work function for AG and fixed metal work function for TG of 3.9 eV with gate length of 30 nm and 10 nm, respectively. It is clear from the figure that increase in the AG metal work function of the proposed device, band bending is obseved at the source and drain end. However, significant change in tunneling width of SB present between source and channel is not obtained for the different work function of the AG. In addition, conduction band edge of the energy band diagram slightly shifts in upward direction for increasing work function of AG. Therefore, we can conclude that I_{on} of the device will not affect considerably by the variation in the AG work function. In addition, at the drain end, negligible change is observed in SB height and width for the proposed device.

Figure 3 shows the transfer charcteristics of the proposed device for differnt AG work function and fixed TG gate work function of 3.9 eV with fixed gate lengths of 30 nm and 10 nm, respectively. It is clear that for the increasing value of AG work function low I_{off} is observed. In addition, the negligible change in I_{on} of the device is observed for the variation in AG gate work function. However, subthreshold slope

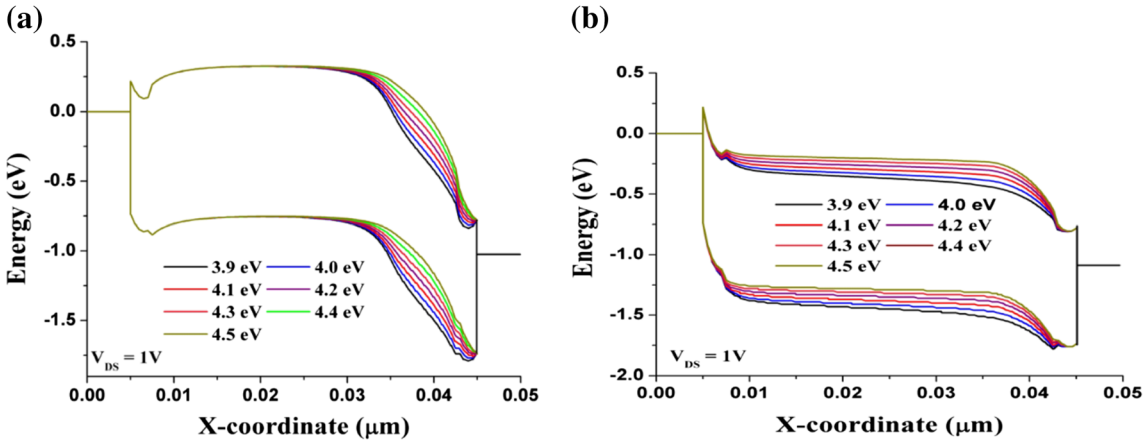


Fig. 4 Energy band diagram of the proposed device for different TG work function with gate length of 10 nm and fixed value of AG work function of 4.5 eV having gate length of 30 nm in (**a**) Off-state condition, (**b**) On-state condition

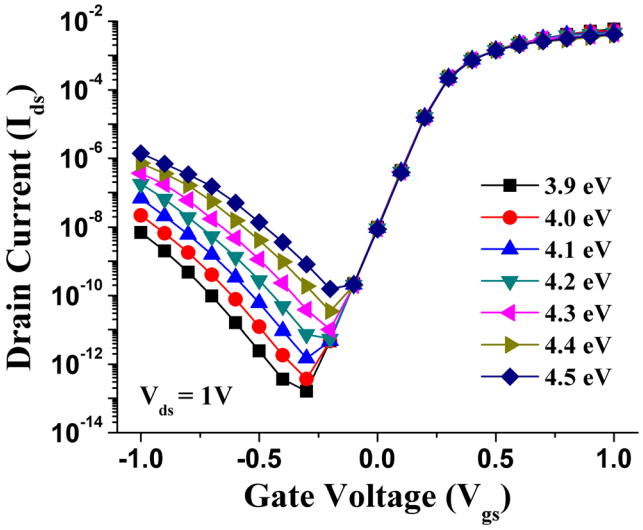


Fig. 5 Transfer characteristics of the device for different TG work function with gate length of 10 nm and fixed value of AG work function having gate length of 30 nm

of the device is degraded for the low value of AG metal work function. Moreover, the $I_{ambipolar}$ is also not significantly changes with the increase in AG metal work function.

Next, we will analyze the effect of different value of TG work function on the I_{off} of the device. Figure 4a shows the off-state energy band diagram of the proposed device for different TG work function and fixed AG work function of 4.5 eV with fixed gate length of 30 nm and 10 nm, respectively. It can be observed that for increasing value of TG work function, the SB width at the drain end reduces for higher value of TG work function. Therefore, SB width reduction increases the tunneling probability of electron between drain and channel. Consequently, the large I_{off} can be obtained for the device. However, it can also be observed that change in TG work function will not affect the source side SB width and height. As a result, I_{on} of the device will be not affected. In conclusion, for low value of TG, I_{off} reduces due to increase in tunneling probability between drain and channel.

Figure 4b shows the energy band diagram in the on-state condition of the proposed device for different TG work function and fixed AG gate work function of 4.5 eV with fixed gate length of 30 nm and 10 nm, respectively. It is clear from the figure that for increasing value of TG work function the conduction and valence band edge shifts upward direction. It is also observed that the SB width at the source side will not

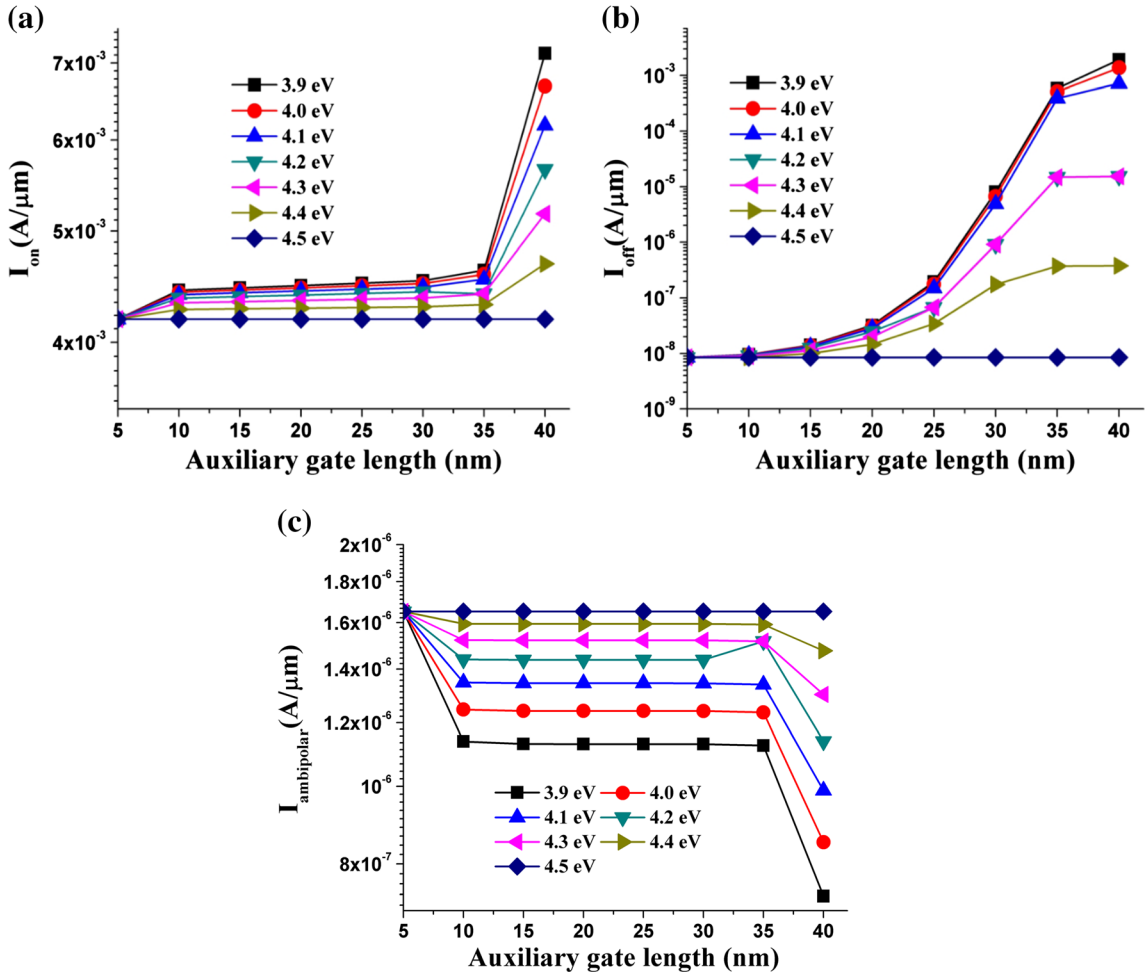


Fig. 6 The effect of AG length and work function variation on (**a**) I_{on}, (**b**) I_{off} (**c**) $I_{ambipolar}$

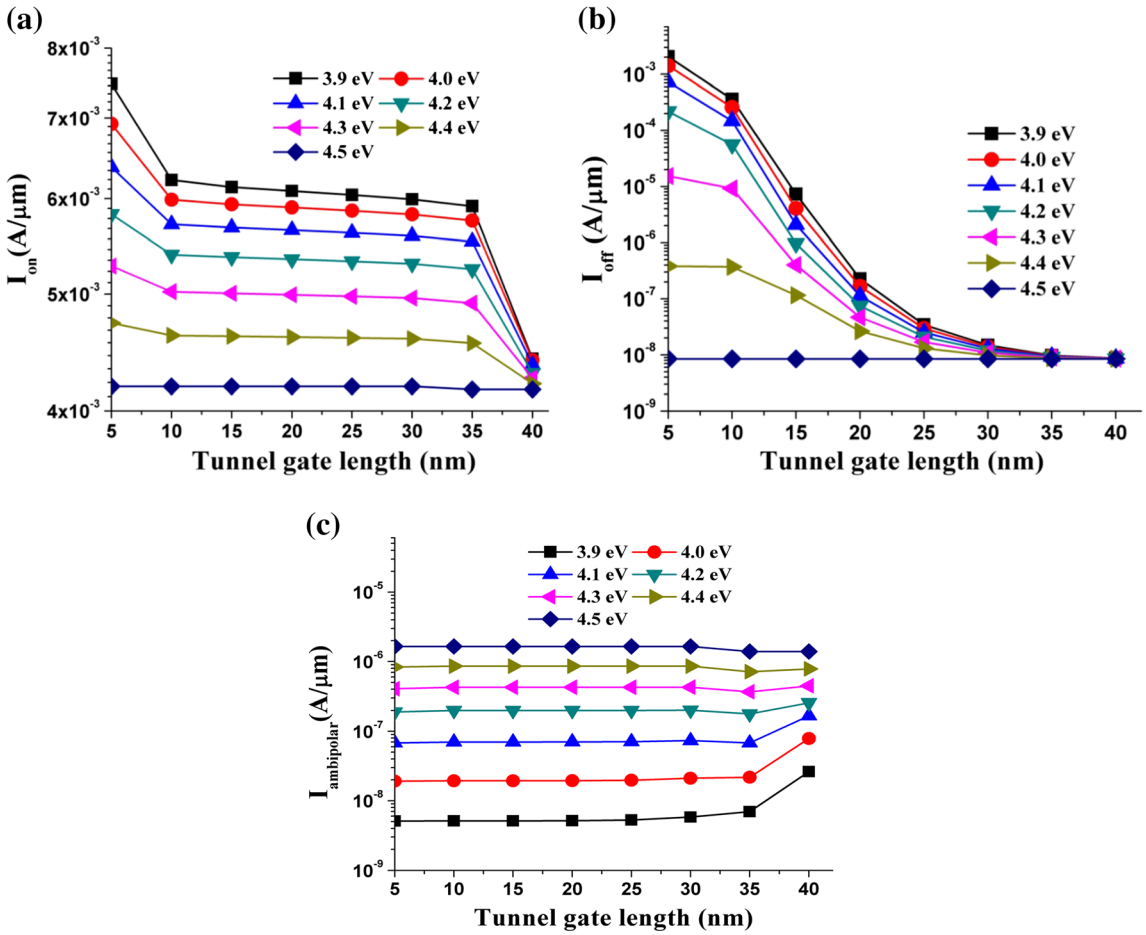


Fig. 7 The effect of TG length and work function variation on (**a**) I_{on}, (**b**) I_{off} (**c**) $I_{ambipolar}$

alter significantly; however, increase in SB width of the drain end is obtained for the device. Therefore, the on-state of the device will not be affected by the change in TG work function as can be seen in transfer characteristics of the device. However, significant decrease in the $I_{ambipolar}$ is observed for the low work function of TG. This can be also verified by the transfer characteristics of the proposed device for different TG work function and fixed AG work function as shown in Fig. 5. It is clear that increase in TG work function enhances the I_{off} of the device. Therefore, it concludes that TG having low metal work function is suitable for the proposed device.

Figure 6a, b and c shows the effect of different length and work function of AG in $I_{on,}$ I_{off} and $I_{ambipolar,}$ respectively. It is clear from the Fig. 6a that for the increasing length and work function of AG, small change in the I_{on} of the device is observed. For gate length of 40 nm and work function of 3.9 eV, high I_{on} is obtained as compared to lower gate length and higher work function. Therefore, it concludes that for varying AG length and work function, impact on I_{on} of the device is not significant. Similarly, Fig. 6b shows that for increasing length and work function of AG of the device, increase in I_{off} is observed. It is clear that for gate length of 40 nm and work function of 3.9 eV, large I_{off} is obtained. This is due to the fact that for increasing length and work function SB width at the drain end reduces. As a result, increase in tunneling probability enhances the I_{off} of the device. In Fig. 6c, it can be observed that for increasing gate length and lower work function, high $I_{ambipolar}$ is obtained. For gate length of 40 nm and work function of 3.9 eV, low $I_{ambipolar}$ is obtained.

Figure 7a, b and c shows the effect of different TG length and work function on $I_{on,}$ I_{off} and $I_{ambipolar,}$ respectively. It can be observed from the Fig. 7a that for lower work function and shorter gate length, slightly higher I_{on} is obtained. Therefore, for TG having work function of 3.9 eV and gate length of 5 nm, high I_{on} is obtained. On the other hand, for lower work function and short gate length of TG, high I_{off} is observed as shown in Fig. 7b. For gate length of 5 nm and work function of 3.9 eV, high I_{off} is obtained. Figure 7c shows that for lower work function and short TG gate length, low $I_{ambipolar}$ is obtained. Therefore, for work function of 3.9 eV and gate length of 10 nm, low $I_{ambipolar}$ is obtained. This is due to the fact that for lower work function of TG, tunneling width for hole injection increased. Therefore, tunneling probability of hole injection reduced, as a result, reduction in $I_{ambipolar}$ is obtained.

The presented simulation results show that the proposed device proposed device achieves improved performance for optimized value of AG and TG length of 30 nm and 10 nm, in addition, metal work function of 4.5 eV and 3.9 eV, respectively.

4 Conclusions

In this manuscript, we have successfully suppressed the $I_{ambipolar}$ of the DE DS SBMOS. In proposed device, by employing dual metal gate in place of single metal gate of DE DS SBMOS, significant reduction in the $I_{ambipolar}$ is obtained. The on-state and off-state performance of the device is effectively controlled by modulating the SB width and height at the source and drain end. The proposed device exhibits low $I_{ambipolar}$ without affecting the I_{on} of the device. The proposed device achieves more than two-decade reduction in $I_{ambipolar}$ as compared to DE DS SBMOS. In addition, proposed device achieves low I_{off} and improved SS. The proposed device has optimized value of AG and TG length of 30 nm and 10 nm, in addition, metal work function of 4.5 eV and 3.9 eV, respectively. Presented results verify the credibility of device for low power applications and provides path for further research and investigation.

References

1. Jhaveri R, Nagavarapu V, Woo JCS (2009) Asymmetric Schottky tunneling source SOI MOSFET design for mixed-mode applications. IEEE Trans Electron Devices 56(1):93–99. https://doi.org/10.1109/TED.2008.2008161
2. Schwarz M, Calvet LE, Snyder JP, Krauss T, Schwalke U, Kloes A (2017) On the physical behavior of cryogenic IV and IIIV Schottky Barrier MOSFET devices. IEEE Trans Electron Devices 64(09): 3808–3815
3. Martín-Martínez MJ, Couso C, Pascual E, Rengel R (2014) Monte Carlo study of dopant-segregated Schottky barrier SoI MOSFETs: Enhancement of the RF performance. IEEE Trans Electron Devices 61(12):3955–3961. https://doi.org/10.1109/TED.2014.2360468
4. Kale S (2020) Performance improvement and analysis of PtSi Schottky barrier p-MOSFET based on charge plasma concept for low power applications. Silicon 12:479–485
5. Hema Latha NK, Kale S (2020) Dielectric modulated Schottky barrier TFET for the application as label-free biosensor. Silicon 12:2673–2679. https://doi.org/10.1007/s12633-019-00363-7
6. Kale S, Kondekar PN (2015) Suppression of ambipolar leakage current in Schottky barrier MOSFET using gate engineering. Electron Lett 51(19):1536–1538. https://doi.org/10.1049/el.2015.0283
7. Kale S, Kondekar PN (2015) Ambipolar leakage suppression in Ge n-channel Schottky barrier MOSFET. IETE J Res 61(4):323–328. https://doi.org/10.1080/03772063.2015.1021387
8. Valentin R, Dubois E, Larrieu G, Member J-P, Raskin G, Dambrine N, Breil F, Danneville (2009) Optimization of RF performance of metallic source/drain SOI MOSFETs using dopant segregation at the Schottky interface. IEEE Electron Device Lett 30(11):1197–1199
9. Kumar P, Bhowmick B (2017) 2-D analytical modelling for electrostatic potential and threshold voltage of a dual work function gate Schottky barrier MOSFET. J Comput Electron 16(3):658–665
10. Kumar P, Bhowmick B (2018) Scaling of dopant segregation Schottky barrier using metal strip buried oxide MOSFET and its comparison with conventional device. Silicon 10(3):811–820
11. Guin S, Chattopadhyay A, Karmakar A, Mallik A (2014) Impact of a pocket doping on the device performance of a Schottky tunneling field effect transistor. IEEE Trans Electron Dev 61(7):2515–2522. https://doi.org/10.1109/TED.2014.2325068
12. Bashir F, Loan SA, Rafat M, Abdul Rahman M, Alamoud SA, Abbasi (2015) A high-performance source engineered charge plasma-based Schottky MOSFET on SOI. IEEE Trans Electron Devices 62(10):3357–3364
13. Larson JM, Snyder JP (2006) Overview and status of metal S/D Schottky-barrier MOSFET technology. IEEE Trans Electron Devices 53(5):1048–1058. https://doi.org/10.1109/TED.2006.871842
14. Kale S, Kondekar PN (2015) Design and investigation of double gate Schottky barrier MOSFET using gate engineering. IET Micro Nano Lett 10(12):707–711. https://doi.org/10.1049/mnl.2015.0046
15. Kale S, Kondekar PN (2016) Ferroelectric Schottky barrier tunnel FET with gate-drain underlap: Proposal and investigation. Superlattice Microst 89:225–230. https://doi.org/10.1016/j.spmi.2015.11.019
16. Kale S, Kondekar PN (2018) Charge plasma-based source/drain engineered Schottky Barrier MOSFET: Ambipolar suppression and improvement of the RF performance. Superlattice Microstruct 113:799–809
17. Kale S, Banchhor S, Kondekar PN (2015) Performance study of high-k gate & spacer dielectric Dopant Segregated Schottky Barrier SOI MOSFET. 2nd International Conference on Electronics and Communication Systems (ICECS), Coimbatore, pp 1142–1145. https://doi.org/10.1109/ECS.2015.7124762
18. Kale S. Investigation of dual metal gate Schottky barrier MOSFET for suppression of ambipolar current. IETE J Res. https://doi.org/10.1080/03772063.2020.1823250
19. Kale S, Kondekar PN (2017) Design and investigation of dielectric engineered dopant segregated Schottky barrier MOSFET with NiSi source/drain. IEEE Trans Electron Devices 64(11):4400–4407
20. Zhao QT, Rije E, Bruer U, Lenk S, and S. Mantl (2004) Tuning of silicide Schottky barrier heights by segregation of sulfur atoms. In: Proc. 7th Int. Conf. Solid-State Integr. Circuits Technol, pp 456–459
21. (2014) ATLAS Device Simulation Software. Silvaco, Santa Clara
22. Ieong M, Solomon PM, Laux SE, Wong H-SP, Chidambarrao D (1998) Comparison of raised and Schottky source/drain MOSFETs using a novel tunneling contact model. In: IEDM Tech. Dig., pp 733–736. https://doi.org/10.1109/IEDM.1998.746461.
23. Matsuzawa K, Uchida K, Nishiyama A (2000) A unified simulation of Schottky and ohmic contacts. IEEE Trans Electron Devices 47(1):103–108. https://doi.org/10.1109/16.817574.
24. Vega RA (2006) Comparison study of tunneling models for Schottky field effect transistors and the effect of Schottky barrier lowering. IEEE Trans Electron Devices 53(7):1593–1600. https://doi.org/10.1109/TED.2006.876261.
25. Neamen DA (2003) Semiconductor physics and devices, 3rd ed. McGraw-Hill, New York
26. Sze SM, Ng KK (2007) Physics of semiconductor devices. Wiley, Hoboken
27. Banchhor S, Kale S, Kondekar PN (2015) Influence of underlap gate length on analog/RF performance of pocket doped Schottky Barrier MOSFET. 2015 2nd International Conference on

Electronics and Communication Systems (ICECS), Coimbatore, pp 1152–1155. https://doi.org/10.1109/ECS.2015.7124764.

28. Kale S, Banchhor S, Kondekar PN (2016) Impact of underlap channel on analog/RF performance of dopant segregated schottky barrier MOSFET on ultra-thin body SOI. 2016 International Conference on Emerging Trends in Engineering, Technology and Science(ICETETS), Pudukkottai, pp 1–6. https://doi.org/10.1109/ICETETS.2016.7603033.

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

Silicon
https://doi.org/10.1007/s12633-020-00840-4

ORIGINAL PAPER

Effect of SNF-Based Superplasticizer on Physical, Mechanical and Thermal Properties of the Geopolymer Concrete

Manvendra Verma [1] · Nirendra Dev [1]

Received: 24 September 2020 / Accepted: 12 November 2020
© Springer Nature B.V. 2021

Abstract

In the experimental investigation, to analyse the effect of the superplasticiser on the GPC mechanical properties; it includes the compressive strength, flexural strength, splitting strength and MOE of the GPC. Superplasticizers used for reducing the water content from the mix design of the concrete, reduction in water content enlarges the strength of the concrete. GPC is an innovative, eco-friendly, cementless and durable concrete in which cement content replace by the pozzolanic content that is full of silica and alumina mineral constituent. GPC indirectly reduces the carbon footprints by the whole replacement of the cement because, in the production of cement, much amount of carbon dioxide emitted in the environment. After the comprehensive survey, the references conclude that the best superplasticiser for the GPC are SNF based because the comprehensive survey authors define the PCE based superplasticisers increases the workability of the GPC but the reduction of the strength at a very high level. So, in the experimental study, to analyse the different dosage of the SNF based superplasticiser in the GPC mix designs for the physical, mechanical properties, microstructural properties and thermal analysis of the specimens. The M2 mix got the optimum point of strength at 1% dosage of superplasticiser in the GPC mix design. It's increasing the workability with the enhancement of the strength in the mixes of GPC. The mix design specimens show higher thermal stability at the elevated temperature.

Keywords Geopolymer concrete · Flyash · GGBFS · Mechanical properties · Curing conditions · XRD analysis · TGA

Abbreviations

GPC	Geopolymer concrete
GGBFS	Ground granulated blast furnace slag
XRD	X-Ray diffraction
SEM	Scanning Electronic Microscope
EDS	Energy-dispersive X-ray spectroscopy
PCE	Poly Carboxylate ether
SNF	Sulphonated Naphthalene Formaldehyde
ITZ	Interfacial Transition Zone
SCGC	Self Compacting Geopolymer Concrete
ASTM	American Standard Testing Materials

✉ Manvendra Verma
manvendraverma_2k16phdce05@dtu.ac.in

Nirendra Dev
nirendradev@dce.ac.in

[1] Department of Civil Engineering, Delhi Technological University, Delhi 110042, India

1 Introduction

GPC is an innovative concrete in the upcoming world, and it reduces the carbon footprints directly by the replacement of cement from the concrete [1]. Superplasticizer used for increasing strength by reducing the water content from the mix design of concrete. The addition of superplasticiser up to 2% approximately of binder mass, enhance the workability of the fresh GPC with significantly less effect on the strength of hardened GPC [2]. The superplasticiser usage is not beneficial for the high-temperature performance of the GPC [3]. The naphthalene based superplasticiser very useful by increasing the workability of a slump of around 136% but in the PCE based superplasticiser increase the workability larger than SNF based around 145% but its effects on the strength of the design mix specimens by reducing the strength around 29%. In some case, SNF based superplasticiser is not showing adverse effects on the strength of the concrete [4]. The citric acid and sucrose addition in the mix as a perfect alternative chemical admixture to flyash-based GPC for enhancing the rheological properties of paste with increasing strength.

Published online: 06 January 2021

Sucrose work as a retarder in the GPC mix, while citric acid works like the accelerator in the mix for hardening. The addition of sucrose in the mix, control the compressive strength, but it affects the porosity characteristics of the matrix. Sucrose added samples have a relationship between the compressive strength and porosity, where porosity directly attach to compressive strength. Sucrose work as chemical admixture as a retarder in the mix of GPC [5]. The PCE-based superplasticiser defines the retarding effect on the flyash-slag based GPC without effect on the heat of hydration of the paste and enhance the workability larger than the SNF based superplasticiser. PCE-based superplasticiser dosage increases beyond 2%, directly affect the development of strength before the 7 days, but after it could be adverse effects on the strength of the GPC [6].

The use of superplasticiser not only enhance the workability and mechanical strength of SCGC but also increased the development of microstructure of bonding between the paste and aggregate at ITZ of GPC. The addition of superplasticiser variation in the GPC mix affects the thickness of ITZ in microstructure level, and also affected the compressive strength of the mix. The increase in the superplasticiser utilisation enhances the compressive strength of SCGC, as well as built quality microstructure by decreasing ITZ thickness. Superplasticiser content over 2% found insufficient to generate the desirable workability with the resistant to segregation. In comparison, superplasticiser content of 6% and 7% gives the required workability properties within the EFNARC limits and 7% superplasticiser dosage produced the most significant strength at all ages and also increased the microstructural properties [7].

[8] stated that the most significant drop in yield stress in alkali-activated slag cement found when a naphthalene-derivative superplasticiser added to NaOH-activated slag pastes and mortars because of its inherent stability in alkaline media. The other admixtures incorporated in alkali-activated slag pastes and mortars, however, do not significantly modify their rheological parameters. The Conventional superplasticiser generally used in the OPC concrete as an additive to the binder for increasing strength but in the GPC mixes, it deteriorates the strength of the hardened specimens. The use of superplasticiser at the elevated temperature shows the negative effect on the GPC mixes specimens [3]. SNF-based superplasticiser got a negligible effect on the strength at the higher temperature. SNF-based superplasticiser is also useful for the high molar NaOH content used in the GPC mix design [9].

2 Materials

Locally available raw materials used in the GPC, it contains maximum 20 mm size coarse aggregate and M-sand as fine aggregate on the replacement of sand. Flyash and GGBFS used as the binder in the GPC and alkaline solution for activating the pozzolanic binder.

Stone dust sand of Delhi used for the sample and tests in the laboratory. The weight of the sample taken is 1000 g for the gradation of sand/stone-dust and calculate the weight retained or passing through various sieves [10]. The fineness modulus is an empirical factor obtained by adding the cumulative percentages of aggregate retained on each of the standard sieves ranging from 80 mm to 150 μ and dividing this by an arbitrary number 100. Hence Fineness modulus = (10.7 + 40.5 + 61.9 + 74.9 + 87.6)/100 = 2.756. As fineness modulus lies between 2.6 and 2.9, the sand is medium. As the percentage passing 600 μ sieve is between 35 and 59, the sand belongs to gradation II. From the gradation curve, we find, $D_{10} = 0.1$, $D_{30} = 0.4$, $D_{60} = 1.2$ and calculate the coefficient of uniformity is 12 is more than 6 and coefficient of curvature is 1.33. Thus the sand is well-graded. The specific gravity and water content of the samples tested by the pycnometer apparatus and its value calculated as 2.62 and 1.21% respectively [11]. The silt content and bulk density areas 6% and 1610 kg/m^3, respectively.

Locally available material used for the tests and 20 mm aggregate used as maximum aggregate size in the mix. The 20 mm and 10 mm aggregate used as the ratio 60/40 by mass in the GPC mix design. The sieve analysis of the aggregates is done for particles size and gradation of the coarse aggregates. The fineness modulus of samples calculated (13.2 + 41.85 + 78.45 + 95.15 + 99.6 + 5X100)/100 = 7.29. The properties of the coarse aggregates in which the fineness modulus, specific gravity, water absorption, crushing value, impact value, abrasion value, flakiness index and elongation index tested in the laboratory [10–14]. All the properties of coarse aggregate desirable for the construction purpose as per the Indian Standard codes. The specific gravity, water absorption, crushing value, impact value, abrasion value, flakiness index and elongation index of the aggregates samples are 2.79, 0.2%, 23%, 22%, 8%, 24%, and 30%, respectively.

Conplast SP 430 is based on SNF used as a superplasticiser in the design mix of GPC. It appearance look like brown liquid and have air entrainment up to 2% with the specific gravity of 1.18 at 25 °C.

2.1 Binding Materials

Sodium hydroxide is used as a chemical solution component for activating the pozzolanic material like flyash and GGBFS. Sodium hydroxide uses in the project made by the Central Drug House (P) Ltd. in New Delhi-110,002 (India). Its appearance like a white deliquescent flake and have a solubility in 10% solution in water is evident with the 96% minimum assay in acidimetric form.

Sodium silicate also a component of the alkaline solution is used in GPC mix design for activating the Flyash and GGBFS. Sodium Silicate solution (water glass) uses in the project made by the Central Drug House (P) Ltd. in New Delhi-110,002 (India). It is thick, tack/slightly cloudy liquid Miscible With Water and have the minimum assay of Na_2O titrimetric is 10% with the specific gravity of 1.42.

Flyash is produced in the thermal power stations by the combustion of coals for the electricity generation. Flyash collects from the Guru Nanak dev Thermal Plant Bhatinda, Punjab in India. Flyash is the type of Class F by the classification as per the ASTM C618–33512 [15]. Table 1 presents the chemical compositions present in the sample materials and Figs. 1, 2 and 3 shows the scanning electron microscope picture in which that shows the porous spherical particles of flyash. Figure 4 denotes the elemental component graph of the material tested by EDS. Figure 1 shows the XRD graph of samples.

GGBFS generated through the isolation of molten slag with the help of high-pressure water jets. Separation of the molten slag prevents crystallisation, so the end products found in the form of glassy and granular aggregates. The end product crushed, pulverised and separated for the multipurpose of the applications. GGBFS is used in the project taken from the Bhilai Steel Plant, Durg, Chhattisgarh in India [16]. Table 1 presents the chemical compositions present in the sample materials, and Fig. 3 shows the scanning electron microscope picture in which that shows the porous undefined shape of particles of GGBFS. Figure 5 denotes the elemental component graph of the material tested by EDS.

3 Tests Setup

3.1 Physical Test

The slump test is used for the workability of the fresh GPC mix in which the slump apparatus is conical in shape or size of top diameter, bottom diameter and height are 100 mm, 200 mm and 300 mm respectively. It is a physical properties test of GPC mix [17, 18] (Table 2).

Table 1 Composition of flyash and GGBFS

Characteristics	SiO_2	Al_2O_3	CaO	Fe_2O_3	MgO	SO_3	LOI
Flyash (%)	45.8	21.4	13.7	12.6	1.3	1.9	.1
GGBFS (%)	34.52	20.66	32.43	.57	10.09	.77	.3

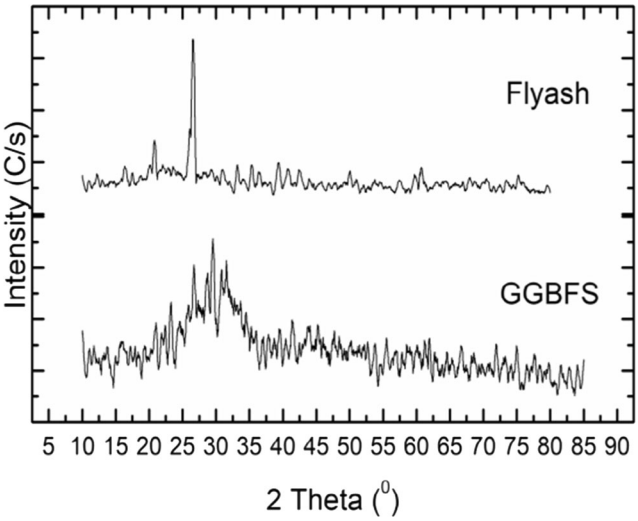


Fig. 1 XRD Graph

3.2 Density

The GPC mix design specimens are cast in the forms of the cube, cylinder and prism shape of different sizes of the moulds and weight after the demould and the time of testing for examining the density of the GPC mix samples [19].

3.3 Mechanical Properties

The GPC mix is cast in the mould of size 150 mm*150 mm*150 mm. The GPC mix specimens examine for compressive strength under axial loading in Universal Testing Machine with 5.3kN/s rate of loading [18, 20–27]. The cylindrical samples of size 150 mm*300 mm diameter*hieght are used for the testing of splitting tensile

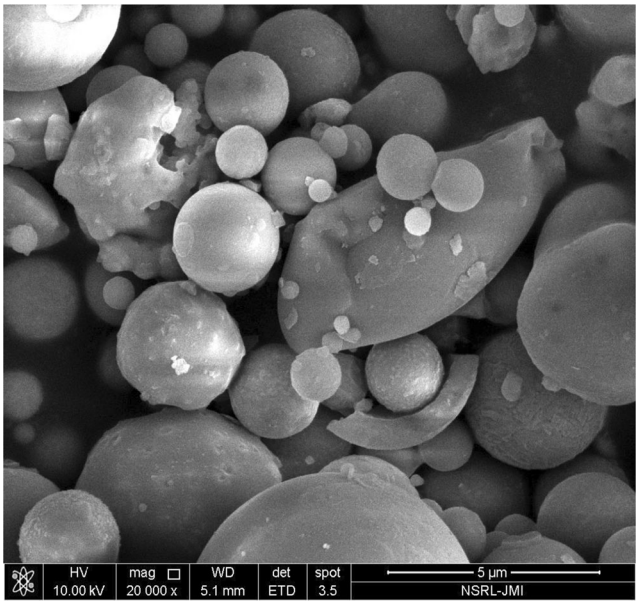


Fig. 2 SEM image of Flyash

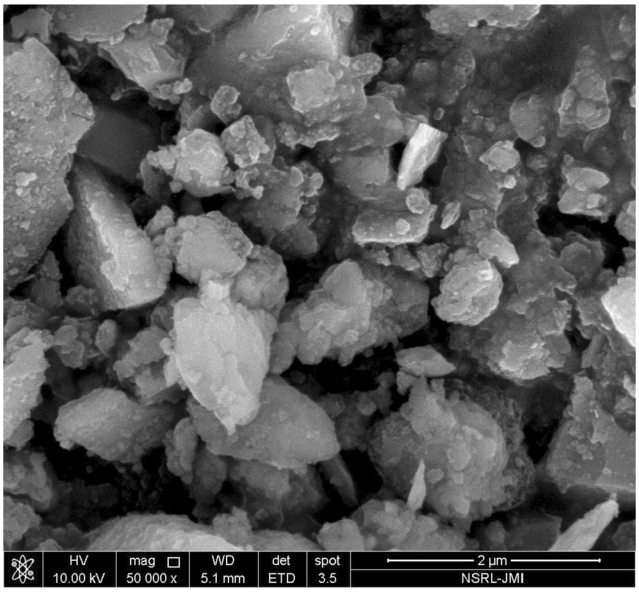


Fig. 3 SEM image of GGBFS

strength of concrete mix design. The transverse load is applying on the cylinder with 4.5kN/s rate of loading to examine the splitting tensile strength of the GPC cylinder specimen in the Universal testing machine [28]. The beam samples of moulds size width, height and length are 100 mm, 100 mm and 500 mm, respectively are used for the testing of the flexural tensile strength of the GPC mix design specimen. The two-point load applied on the beam specimen to examine the flexural tensile strength of the GPC mix specimen [22, 24].

The cylinder samples of size diameter*length are 150 mm*300 mm, respectively, for the testing of modulus of elasticity and poisons ratio of GPC mix design. The axial load is applying on the cylinder to examine the vertical and horizontal strain of the GPC cylinder specimen in the universal testing machine to calculate the modulus of elasticity and poisons ratio [29].

3.4 Non-Destructive Test

Rebound hammer is the name of an apparatus that examines the strength of the GPC hardened sample without destruction

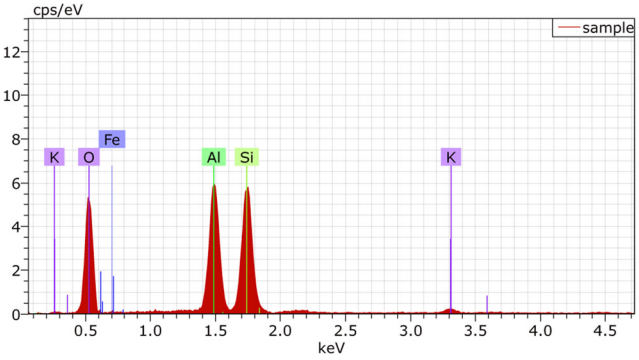


Fig. 4 EDS graph of Flyash

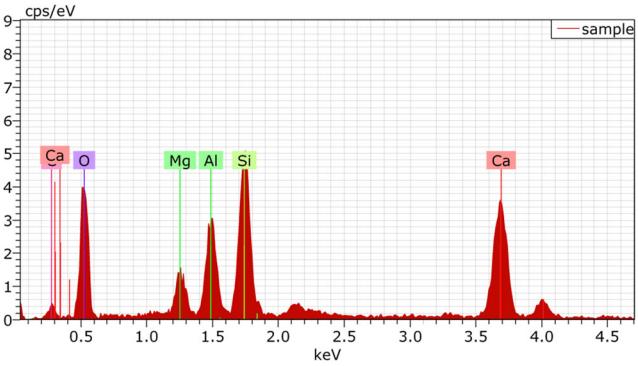


Fig. 5 EDS graph of GGBFS

of the specimens. In which both type of samples cube and cylinder tests for the strength of the GPC mix [30, 31]. The UPV test also a non-destructive test for the concrete specimens. It passes the ultrasonic wave through the specimen and read the time passage of the ultrasonic pulse wave by the apparatus. The time passage would decide the strength quality of the specimen of GPC if the time passage is lesser shows the better quality or dense matrix of the specimen compared to the higher time passage is taken specimens [32].

3.5 Microstructural Analysis

The TGA test does the thermal analysis of the GPC matrix, in which check the weight retained or loss with the increasing temperature 10 °C per minute up to the 800 °C [33]. The bond formation or chemical reaction end products phase analysis does through the XRD test in the laboratory in which check the crystalline or amorphous phase formation with oxides [34].

4 Results and Discussion

GPC mix designs samples cast in the cube, cylinder and beam form, cured under both conditions ambient-curing and oven-curing samples tests at 7, 14, 28, 42 and 56 days of all mixes.

4.1 Workability

The slump test is used for the workability of the GPC fresh mix. It is value varies with the change of the superplasticiser dosage in the GPC mix. Figure 6 describes the slump value changes with an increment of the superplasticiser dosage in the mix. The slump value increases with the increment of superplasticiser dosage from 0.5% to 1.0% of the binder in the mix, but beyond 1.0% dosage decreases the slump value slightly.

Table 2 Mix design

Mix Designs	Fly ash (kg/m^3)	GGBFS (kg/m^3)	Coarse Aggregate (kg/m^3)	Fine Aggregate (kg/m^3)	NaOH Solution (kg/m^3)	Sodium Silicate (kg/m^3)	Super Plasticizer (kg/m^3)	Extra water (kg/m^3)
M1	303.75	101.25	1269	683	81.0 (13 M)	81.0	2.025	20.25
M2	303.75	101.25	1269	683	81.0 (13 M)	81.0	4.05	20.25
M3	303.75	101.25	1269	683	81.0 (13 M)	81.0	6.075	20.25

4.2 Density

The density of the mix is measured through the weight of specimens before the destructive test of cube samples. The ambient-cured specimens got higher density compared to the oven-cured samples. The density of the specimens increases from 0.5% to 1% then after decreases gradually with increases the superplasticiser content in the GPC mix design. The M2 mix design got the optimum point of density in both cured samples. The M2 mix of ambient cured got highest 2476 kg/m^3 in all mixes of GPC. Figure 7 describe the relationship between the density of the mix and superplasticiser dosage in the mixes for both curing conditions.

4.3 Compressive Strength

The compressive strength of the mixes results found by the test of cube samples in the UTM machine at 7, 14, 28, 42 and 56 days after the casting. The compressive strength of the oven-cured samples is higher compared to the ambient-cured samples of the same mix of GPC. In the GPC mixes, Superplasticisers 1% dosage in the mix got a higher strength compared to the other mixes with different percentage dosage. Figure 8 shows the compressive strength of different mixes in both curing conditions; M2 got optimum strength. The superplasticiser dosage varies from 0.5%–2.0% in the GPC mix design, to check the effect dosage variation in the compressive strength of the mix. The compressive strength increases with the increment of superplasticiser dosage from 0.5% to 2.0% but after the increment of dosage reduces the compressive strength of GPC mix specimens in both curing conditions. The maximum compressive strength of ambient-cured and oven-cured specimens are 24.4 MPa and 33.2 MPa, respectively at 56 days after the casting. The compressive strength of mixes got 95% strength at 28 days in both curing conditions, but the initial seven days strength of the oven-cured specimens is higher compared to ambient-cured specimens of the same mix.

4.4 Splitting Tensile Strength

The splitting tensile strength measures through the destructive test of the cylinder specimens in the UTM machine by giving statical load along the transverse direction on the samples. The splitting strength of the oven-cured samples is higher compared to the ambient-cured samples of the same mix of GPC. In the GPC mixes, Superplasticisers 1% dosage in the mix got a higher strength compared to the other mixes with different dosages. Figure 9 shows the splitting strength of the various design mix a) shows the ambient cured specimens strength and b) shows the oven-cured specimens of

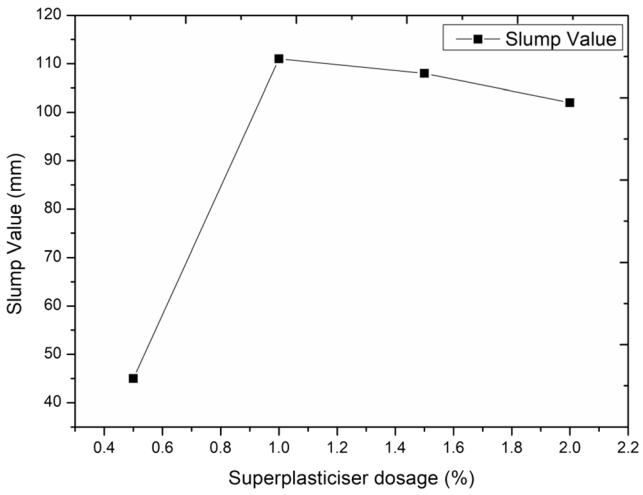


Fig. 6 Graph between the Slump value and superplasticiser dosage

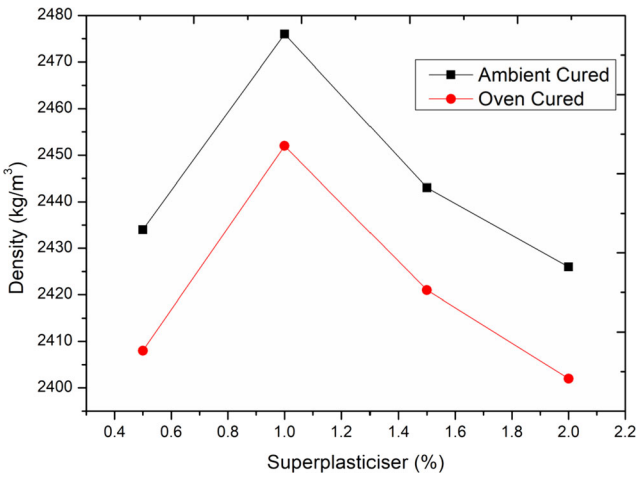


Fig. 7 The density of various mixes

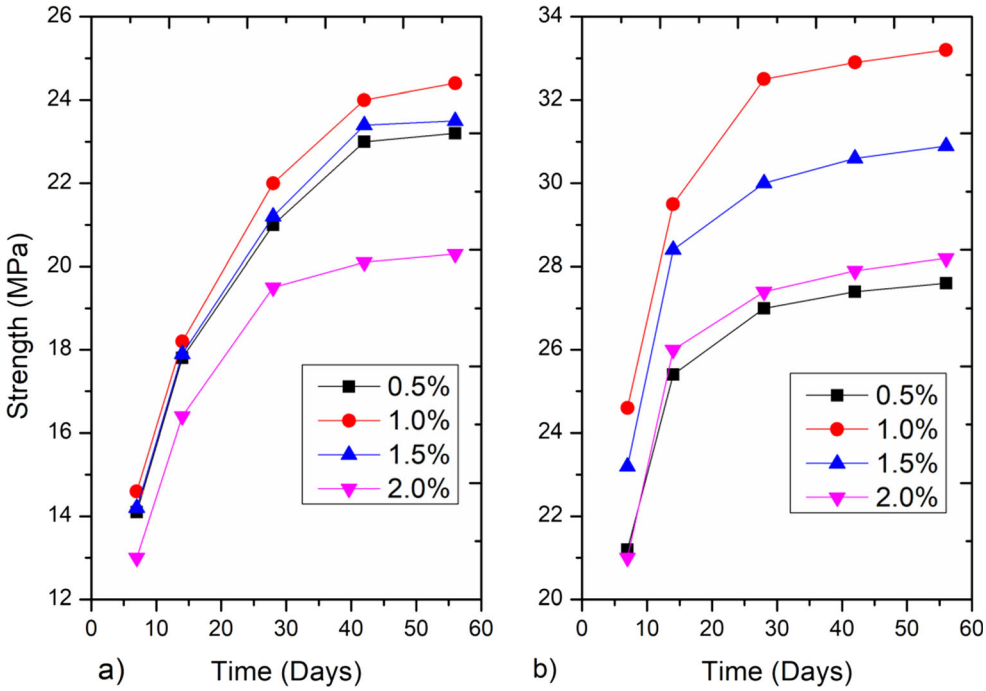


Fig. 8 Compressive strength of **a** Ambient-cured **b** Oven-Cured GPC

GPC, in both curing conditions M2 mix got optimum strength. The variation in the splitting strength is similar to the compressive strength of the mix. The maximum splitting tensile of ambient-cured and oven-cured specimens are 3.5 MPa and 5.0 MPa respectively at 56 days test after the casting. The splitting tensile of mixes of superplasticiser dosage of 0.5%, 1.0%, 1.5%, and 2.0% are 2.4 MPa, 3.5 MPa, 3.1 MPa, and 2.6 MPa respectively at 56 days test of the ambient-cured specimens, whereas the 3.5 MPa, 5.0 MPa, 4.5 MPa, and 3.9 MPa respectively of oven-cured specimens at 56 days after the casting.

4.5 Flexural Strength

The flexural strength is measure through the destructive test of the beam specimens in the flexure testing machine by giving statical load along with the two-point load on the samples. The flexural strength of the oven-cured samples is higher compared to the ambient-cured samples of the same mix of GPC. In the GPC mixes, Superplasticiser 1% dosage in the mix got a higher strength compared to the other mixes with various dosages. Figure 10 shows the flexural strength of the var-

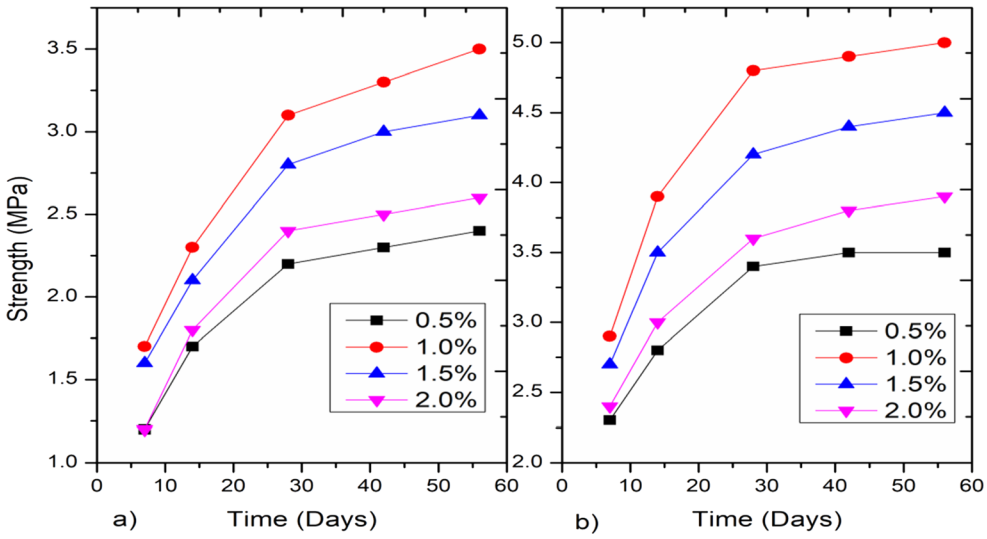


Fig. 9 Splitting Strength of **a** Ambient-cured **b** Oven-Cured GPC

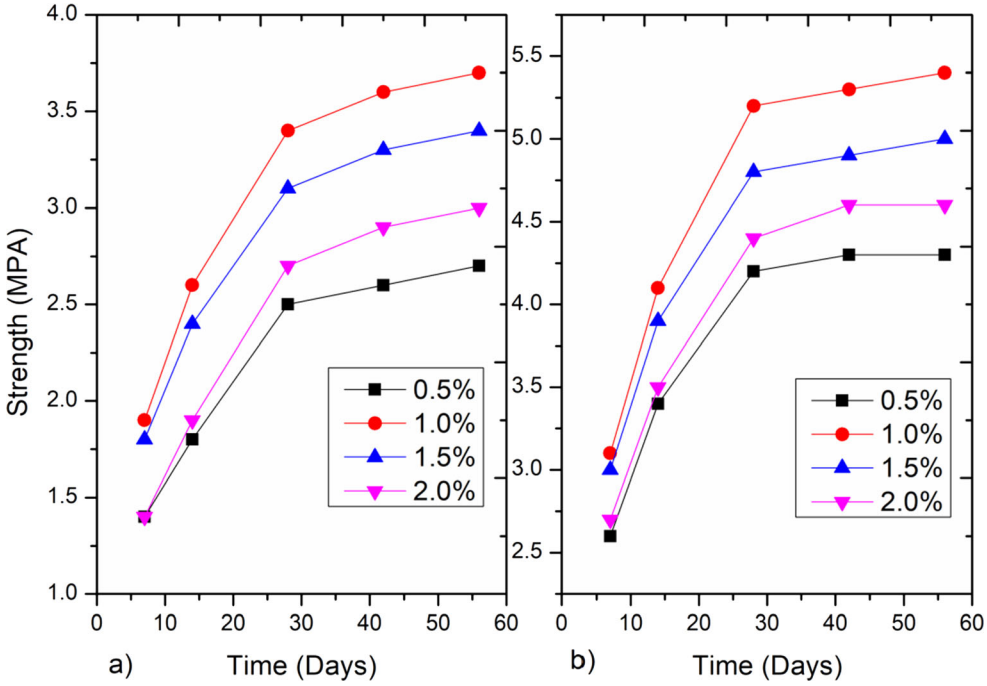


Fig. 10 Flexural Strength of **a** Ambient-cured **b** Oven-Cured GPC

ious design mix a) shows the ambient-cured specimens strength and b) shows the oven-cured specimens of the GPC, in both curing conditions M2 mix got the optimum strength. The flexural strength shows the bending properties or indirect tensile strength of the concrete mix samples. It's about 15%–20% of the compressive strength of the mix. The maximum flexural strength of mixes with superplasticiser dosage 0.5%, 1.0%, 1.5%, and 2.0% are 2.7 MPa, 3.7 MPa, 3.4 MPa, and 3.0 MPa, respectively of ambient-cured specimens whereas 4.3 MPa, 5.4 MPa, 5.0 MPa, and 4.6 MPa, respectively of oven-cured specimens at 56 days test after the casting.

4.6 Modulus of Elasticity

The modulus of elasticity of the mix design samples obtained by the static load applied on the cylinder specimen along the longitudinal axis and measured the deflection of the sample in both directions vertical and horizontal. Stress conclude by the load applied to the specimen to the surface area of the specimen. Poisson ratio also measured through this test by the ratio of lateral strain to the linear strain of the samples. The M2 mix specimens got the higher modulus of elasticity compared to the other mix designs. The oven-cured specimens got a higher MOE compared to the ambient-cured specimens. The optimum point of the MOE of the ambient-cured, and oven-cured specimens is 21.4 MPa and 21.9 MPa, respectively. Figure 11 describe the relationship between the MOE and superplasticiser dosage percentage in both curing conditions.

4.7 Rebound Strength

The rebound test is a non-destructive experiment for the specimens. Rebound strength of all mixes specimens of both cured conditions tests at 7, 14, 28, 42 and 56 days after casting and before the destructive test on the samples. Rebound strength of the oven-cured samples is higher compared to the ambient-cured samples of the same mix of GPC. In the GPC mixes, Superplasticizer 1% dosage in the mix got a higher strength compared to the other mixes with various dosages. Figure 12 shows the graph of rebound strength with time variation of both curing conditions. The M2 mix got the optimum strength in both curing conditions. The rebound strength is slightly higher compared to the destructive compressive strength but

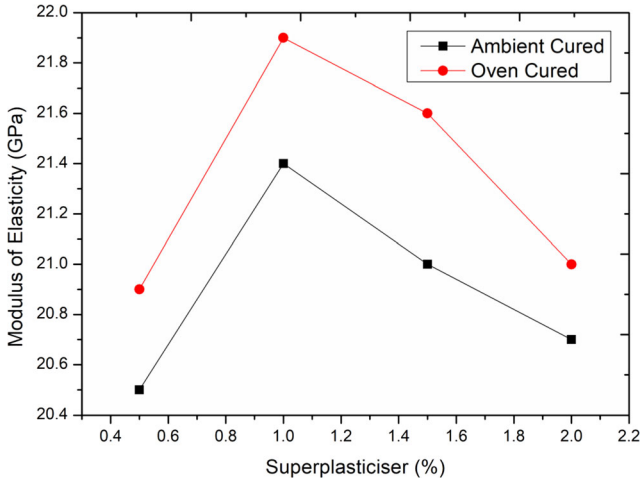


Fig. 11 Modulus of Elasticity of Various mixes

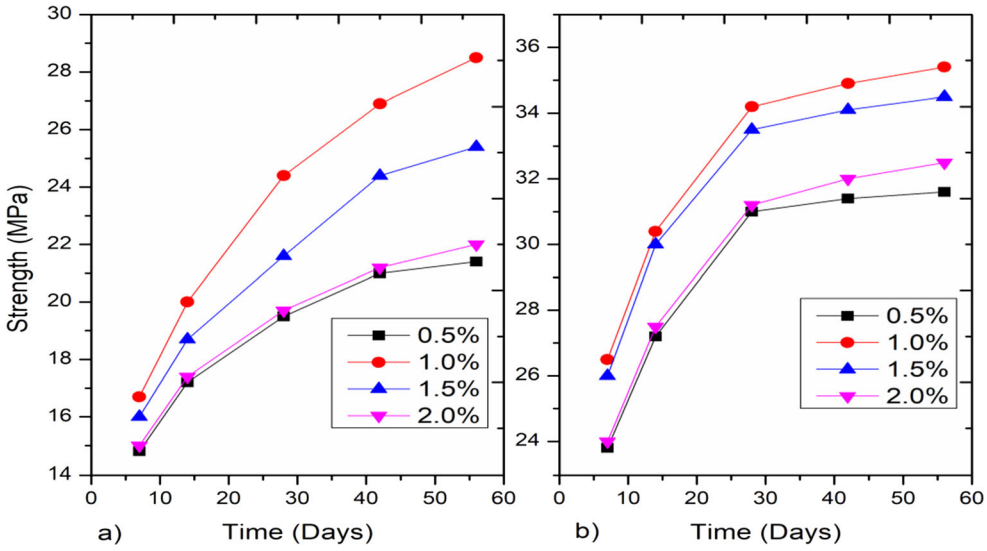


Fig. 12 Rebound Strength of **a** Ambient-cured **b** Oven-Cured GPC

not exceed the limit as per the IS code. The rebound strength is based on the surface hardness of the specimens.

4.8 UPV Test

The UPV test uses for the quality of the hard material by the quantify the time passage from the specimens. It is a non-destructive test to evaluates the concrete strength without destruction. The UPV tests are conducted on cube samples of GPC mix at 7 days, 14 days, 28 days, 42 days and 56 days after the casting; and before the destructive tests on the samples. The apparatus records the time duration of the UPV wave and calculate the velocity that defines the quality of GPC samples. When the time passage of the UPV wave is high, it shows the less strength and vice-versa [35]. The M2 mix shows a higher velocity compared to the other mixes, and it shows the superplasticiser dosage of 1% of binder gives the optimum point strength in both curing conditions. Figure 13 describes the graph of UPV of various mixes in the same graph in which a) shows the ambient-cured results;) shows the oven-cured GPC samples. The maximum velocity of oven-cured specimens and ambient-cured specimens are 4.27 km/s and 2.41 km/s respectively at 56 days test.

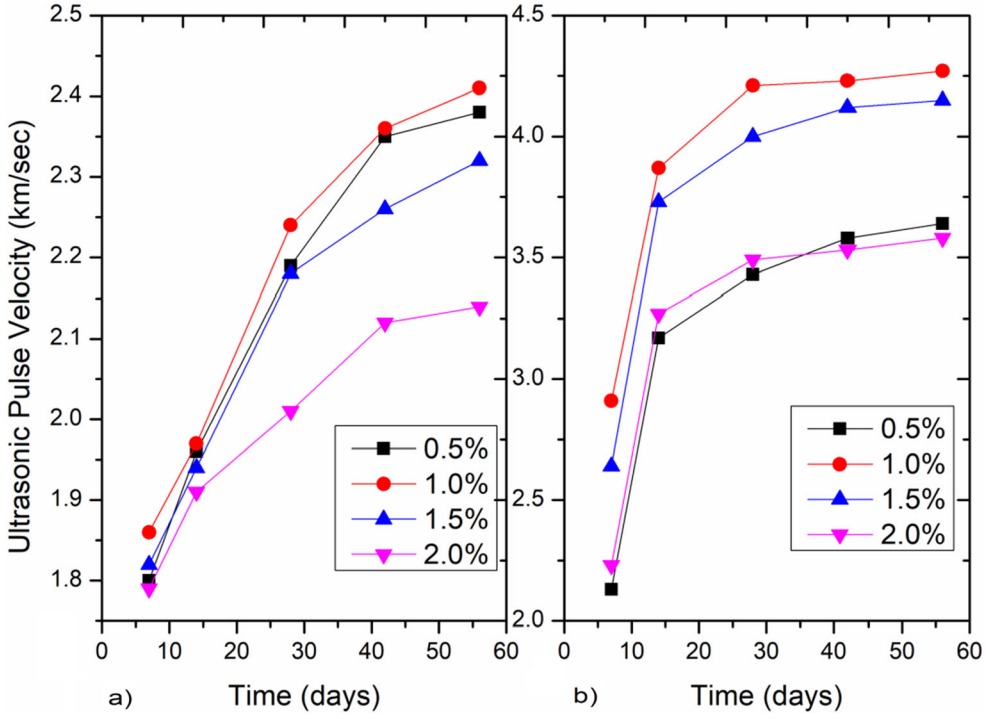


Fig. 13 Graph between the UPV and Time **a** ambient-cured **b** oven-cured GPC

4.9 XRD Analysis

The XRD analysis of the materials defines the intensity of the various oxides present in the material. The XRD test is conducted on the fine particles in powder form below 150-μm size. It defines the number of counts of that mineral oxides in the sample [34]. Figure 14 describes the XRD graph of GPC samples in which; A) shows the ambient-cured, B) shows the oven-cured GPC specimens. It shows that the intensity of the oven-cured sample is low compared to the ambient-cured, due to the oxides minerals component reduction during oven-curing. The pattern of the intensity of both cured samples shows a similar pattern. The sample is taken from the destructive test conducted on the cube after the 28 days and select the optimum strength mix for the XRD test. The XRD test results describe the crystal form by the mineral oxides with their intensities in the proper pattern. In both cured condition, the quartz(silica) shows the highest peak in the pattern.

4.10 Thermal Analysis

TGA-DTG test does the thermal analysis of material in which evaluates the weight with the increment of temperature at rate 10 °C per minutes up to 800 °C. The sample of TGA-DTG test is prepared in the lab by grinding and sieving through 150-μm of the materials. Figure 15 describes the graph of TGA-DTG of; A) ambient-cured B) oven-cured specimens of the GPC mix. It shows that the weight retained a percentage of oven-cured samples is higher compared to the ambient-cured up to 800 °C. The weight retained of ambient-cured and oven-cured samples are 92% and 94% respectively because the oven-cured specimens some water already evaporates during the curing period. The stability of the sample materials is very higher compared to the other concrete materials like OPC concrete. It reduces the weight up to 8% at 800 °C, and it shows a similar derivative pattern.

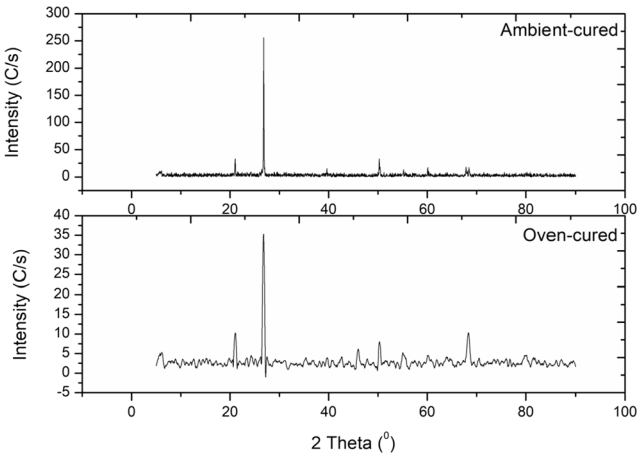


Fig. 14 Xrd graph of **a** ambient-cured **b** oven-cured GPC samples

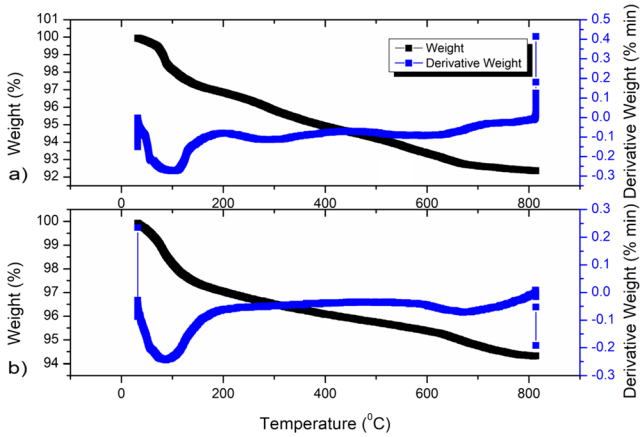


Fig. 15 TGA-DTG graph of **a** ambient-cured **b** oven-cured GPC

5 Conclusion

The conclusion of the work infiltrates through the experimental investigation of the specimens of the various mix designs testing by destructive and Non-destructive test of the samples cured in both conditions. The specimens cast in the cube, cylindrical and beam form for the experimental investigation and cured under ambient and oven conditions of all mixes specimens. After the curing of specimens tests at 7, 14, 28, 42 and 56 days for Compressive, Splitting and Flexural of the design mix of GPC.

- The slump value increases with the increment of superplasticiser dosage from 0.5% to 1.0% of the binder in the mix, but beyond 1.0% dosage decreases the slump value slightly. The M2 mix design got the optimum point of density in both cured samples. The M2 mix of ambient cured got highest 2476 kg/m^3 in all mixes of GPC.
- Compressive strength of the oven-cured samples is higher compared to the ambient-cured samples of the same mix of GPC. In the GPC mixes, Superplasticizer is 1% in mix got a higher strength compared to the other mixes with different percentage of superplasticiser. Similarly, Splitting tensile and flexural strength of the oven-cured samples is higher compared to the ambient-cured samples of the same mix of GPC. In the GPC mixes, Superplasticizer is 1% in mix got a higher strength compared to the other mixes with different percentage of superplasticiser.
- M2 mixes specimens got the higher Modulus of Elasticity compared to the other mix designs. Oven-cured specimens got a higher MOE compared to the ambient-cured specimens. The optimum point of the MOE of the ambient-cured specimen is 21.4 MPa and MOE of the oven-cured specimens is 21.9 MPa.
- The M2 mix got the optimum strength in both curing conditions. The rebound strength is slightly higher compared to the destructive compressive strength but not

exceed the limit as per the IS code. The maximum velocity of oven-cured specimens and ambient-cured specimens are 4.27 km/s and 2.41 km/s respectively at 56 days test.

- The XRD analysis shows that the intensity of the oven-cured sample is low compared to the ambient-cured, due to the oxides minerals component reduction during oven-curing. The pattern of the intensity of both cured samples shows a similar pattern. The stability of the sample materials is very higher compared to the other concrete materials like OPC concrete. It reduces the weight up to 8% at 800 °C, and it shows a similar derivative pattern.

Acknowledgements This work supported by the Civil Engineering Department, Delhi Technological University. Moreover, the help of Sandeep Mishra for XRD analysis and Ankesh Kumar and Jawed Alam for the another in the work that is related to the central instrumentation facility of the Delhi Technological University.

Availability of Data and Material The available data had been used and discussed in the manuscript.

Author Contribution All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for valuable intellectual content; and (c) approval of the final version.

Funding This study was funded by the Delhi Technological University, Delhi, India.

Compliance with Ethical Standard

This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.

Conflict of Interest There are no conflicts of interest or competing interests in this article.

Consent to Participate As a corresponding author or on behalf of all authors of the research paper, I consent to participate.

Consent to Publication All author of the research paper is consent to the publication.

References

1. Verma M, Dev N (2018) Geopolymer concrete : a way of sustainable construction. Int J Recent Res Asp 5(1):201–205
2. Hardjito D, Wallah SE (2004) On the development of Fly ash-based geopolymer concrete. ACI Mater J 467–72
3. Kong DLY, Sanjayan JG (2010) Cement and concrete research effect of elevated temperatures on geopolymer paste, mortar and concrete. Cem Concr Res 40(2):334–339
4. Nematollahi B, Sanjayan J (2014) Effect of different superplasticisers and activator combinations on workability and strength of fly ash-based geopolymer. Mater Des 57:667–672
5. Kusbiantoro A, Ibrahim MS, Muthusamy K, Alias A (2013) The 3 rd international conference on sustainable future for human security development of sucrose and citric acid as the natural-based admixture for fly ash-based geopolymer. Procedia Environ Sci 17:596–602
6. Jang JG, Lee NK, Lee HK (2014) Fresh and hardened properties of alkali-activated fly ash/slag pastes with superplasticisers. Constr Build Mater 50:169–176
7. Demie S, Nuruddin MF, Shafiq N (2013) Effects of microstructure characteristics of interfacial transition zone on the compressive strength of self-compacting geopolymer concrete. Constr Build Mater 41:91–98
8. Palacios M, Banfill PFG, Puertas F (2008) Rheology and setting of alkali-activated slag pastes and mortars: effect if organic admixture. ACI Mater J
9. Albitar M, Visintin P, Ali MSM, Drechsler M, Mohamed Ali MS, Drechsler M (2015) Assessing behaviour of fresh and hardened Geopolymer concrete mixed with class-F Fly ash. KSCE J Civ Eng 19:1445–1455
10. Bureau of Indian Standard (1997) Coarse and fine aggregate from natural sources for concrete. *IS 383–1970*, no. 1997, pp. 1–20
11. Bureau of Indian Standard (1997) Specific gravity, density, voids, absorption and bulking. *IS 2386 (Part III)*, vol. 2386, no. October 1963
12. Bureau of Indian Standard (1997) Mechanical properties. *IS 2386(Part IV)*, vol. 2386, no. March 1964
13. Bureau of Indian Standard (1997) Particle size and shape. *Is 2386(Part I)*, vol. 2386
14. Bureau of Indian Standard (1997) Soundness. *IS 2386*
15. ASTM International (2019) Standard specification for coal fly ash and raw or calcined natural pozzolan for use. *C618–19*, pp. 1–5
16. Metals P-II (2016) Yearbook 2015. 2015(0712):1–10
17. ASTM International (2019) Standard test method for slump of hydraulic-cement concrete 1. *C143/C143M − 15a*, pp. 6–9
18. Bureau of Indian Standard (1959) Methods of sampling and analysis of concrete. *IS 1199–1959*, no. 2004
19. ASTM International (2019) Standard test method for density (Unit Weight), yield, and air content (Gravimetric) of concrete. *C138/C138M − 17a*, pp. 1–6
20. Bureau of Indian Standards (1999) Concrete admixtures-specification. *IS 91031999*, vol. April 1999
21. Bureau of Indian Standard (2009) Concrete mix proportioning – guidelines. *IS 10262 2009*
22. Bureau of Indian Standard (1979) Specification for apparatus for flexural testing of concrete. *lS9399–1979*, no. 2004
23. Bureau of Indian Standards (1996) Specification for concrete vibrating tables. *IS 2514*–1963, no. November 1963
24. Bureau of Indian Standard (2004) Methods of test for strength of concrete. *IS 516–1959*
25. Bureau of Indian Standard (1999) General requirement for pan mixtures for concrete. *IS 12119–1987*, no. 1999
26. Bureau of Indian Standard (1999) Specification for apparatus for determining constituents of fresh concrete. *IS 7325–1974*, pp. 1–12
27. Bureau of Indian Standard (2008) Specification for moulds for use in tests of cement and concrete. *IS 10086–1982*, no. 1952
28. Bureau of Indian Standard (1999) Splitting tensile strength of concrete - method of test. *IS 5816 1999*
29. ASTM International (2019) Standard test method for static modulus of elasticity and poisson's ratio of concrete. *C469/C469M − 14*, pp. 1–5
30. Bureau of Indian Standards (1992) Method of non-destructive testing of concrete-methods of test Part 2: Rebound hammer. *Is 13311–2*, no. 2004
31. ASTM International (2002) Rebound Number of Hardened Concrete. *C805/C805M − 18*, no. 2019, pp. 2–4
32. Bureau of Indian Standard (1992) Non-destructive testing of concrete methods of test Part 1: Ultrasonic pulse velocity. *IS 13311 (Part1) 1992*, no. 2004

33. ASTM International (1868) Standard test method for loss-on-drying by thermogravimetry. *E2550–17*, vol. 11, no. 2019, pp. 1–6
34. ASTM International (2014) Identification of crystalline compounds in water-formed deposits by X-Ray diffraction. *D934–13*, vol. 11, no. February, pp. 1–10
35. ASTM International (2016) Standard test method for pulse velocity through concrete ASTM C 597. *C597–16*, p. 4

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

ISSN(Online) : 2456-8910
International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE)
Volume 4, Issue 7, DOI:10.29027/IJIRASE.v4.i7.2021.836-840,January 2021

Electronic Voting Machine Using a Microcontroller

Payal Yadav[1], Ansh Gupta[2]
[1,2]Department of Electrical Engineering,
Delhi Technological University, Delhi, India
Email: payaa27@gmail.com, gupta.ansh256@gmail.com

Abstract- Electronic voting machine is an efficient and fast solution that allows people to choose their representative in a systematic manner. It is a simple device that has replaced the conventional ballot papers used to record votes. Also, in large populations, the ballot system failed, whereas EVMs proved to be successful and reliable. This project describes a real-time voting machine that is designed using the microcontroller AT89C51 and the software program used is written in assembly language. It is a real-time operating device that produces lesser errors and is easy to operate. Once the votes are cast, the result can be produced in no time just by pressing the result button.

Index Terms *- Electronic Voting Machine, Liquid Crystal Display, Light Emitting Delettevice, Microcontroller.*

I. INTRODUCTION

India is the-largest democratic-country with a total of 850 million recorded voters. The election commission has taken the creative project of recording, storing and counting votes across the nation transparently and securely that is backed by proper legal support. The Election commission has been successfully using EVMs for performing elections for the last 23 years.

The electronic Voting Machine maintains all the traditional ballot paper voting system characteristics while making voting a lot more expedient. It also saves a lot of time and manpower being fast and absolutely reliable.

In addition to this, EVM helps maintain the voting secrecy and the result is 100% tamperproof. EVMs have been of great use to India since they were first introduced in 1982 and subsequently became very popular as well. It is highly user friendly as it provides the result with just a button and also, the voters have only to press one button to cast their vote. The AT89C51 microcontroller used is a high performance, low power 8bit microcomputer with 4K Bytes of programmable flash memory.

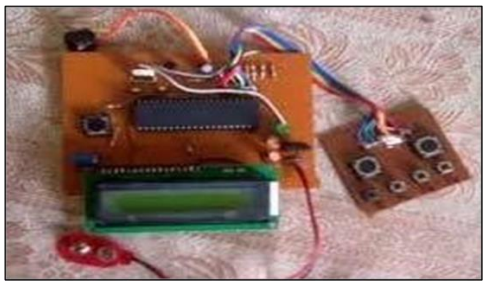
EEPROM memory is used to store the voting sum. The advantage of this memory is that it keeps data in memory even after power is turned off. The microcontroller provides the following features: 128 bytes RAM, -32 I/O lines, two 16-bit counters/timers, full-duplex UART serial channel and an on-chip oscillator and clock circuit. The Idle Mode halts the CPU while allowing the RAM, serial port, timers, and interrupt system to continue operating.

II. OUTLINE OF E-VOTING TECHNOLOGIES

Document Ballot Voting System - This system was used to keep a count of all the votes marked on a ballot sheet electronically. Direct Recording Electronic voting system - In this system, ballot display was provided using a touch display or an electro-optical component. Public NetworkkDRE Voting System – This system uses an electronic ballot to record the entire data throughout the voting process and sends it through a public network to a central location. The Indian EVMs – These are designed and developed by two manufacturing units that are owned by the government, i.e. Bharat Electronics Limited (BEL) and Electronics Corporation of India (ECIL).

The Indian EVMs – These are designed and developed by two manufacturing units that are owned by the government, i.e. Bharat Electronics Limited (BEL) and Electronics Corporation of India (ECIL).

Figure 1. Components of an EVM



ISSN(Online) : 2456-8910
International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE)
Volume 4, Issue 7, DOI:10.29027/IJIRASE.v4.i7.2021.836-840,January 2021

III. BLOCK DIAGRAM OF EVM

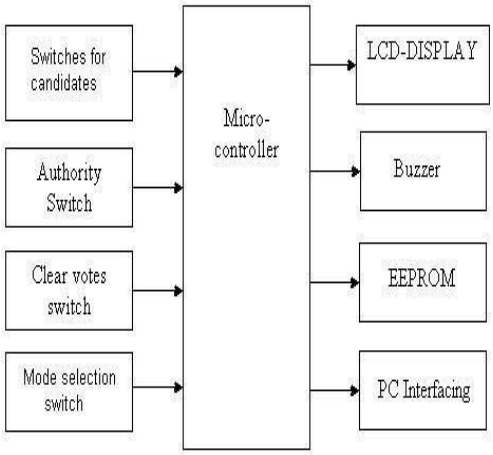
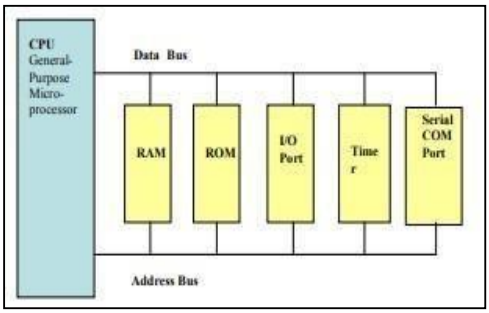


Figure 2. Block Diagram of EVM

IV. BLOCK DESCRIPTION

1. Microcontroller- Based on the signal received from the switches, the mode of operation is decided. Its job is to increment the data when the switches are pressed against the name of each candidate in the voting model. Also in counting mode, it fetches the data from memory and sends it to the display devices.

Figure 3. Microcontroller internal structure



2. LCD - An alphanumeric display is used instead of a seven-segment display because it can display both alphabets and numbers and special symbols that make LCD more user-friendly.
3. LED - Light-emitting diodes are semiconductor devices made out of silicon. These LEDs emit photons when current passes through them, which makes them appropriate for indicator lamps.
4. CONTROL SWITCHES - These are - clear votes, a controller switch, and total votes.

V. CIRCUIT DIAGRAM

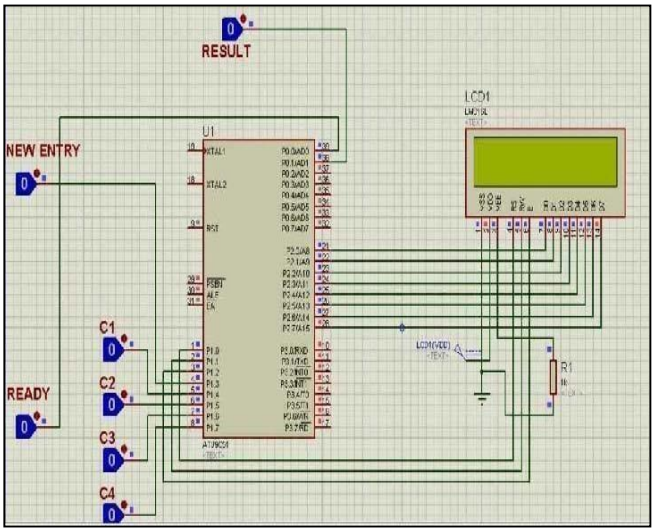


FIGURE 4. PROTEUS MODEL OF THE EVM

VI. WORKING OF EVM

Power on- The supply when turned 'ON' is indicated by a red light.Mode Selection- It has two modes, namely, voting mode, i.e., when the toggle switch is on VCC. In this mode, the message "Please Vote" is displayed on the screen. After a vote is cast, the controller switch is pressed by the voting authority and the next mode is Counting Mode, i.e. when the toggle is on GROUND. In this mode, once the respective key is pressed, each candidate's total number of votes is displayed on the screen.

1. Clear Mode - This switch ensures that all of the preceding data has been cleared and the EVM is ready for another election as it clears all the entries. Usually, this switch is pressed before the voting procedure begins.

2. Buzzer Signal - Pressing of a key when a vote is a buzzer soundController Switch specifies cast - This switch is always under the control of the voting authority as it is provided for enabling the keypad in the voting mode to let a voter cast his vote.

ISSN(Online) : 2456-8910
International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE)
Volume 4, Issue 7, DOI:10.29027/IJIRASE.v4.i7.2021.836-840,January 2021

VII. MAIN UNITS OF EVM

1. Control Unit - The most important part of an EVM is a control unit consisting of a power supply, switch, LCD, microcontroller, and result key. This unit is used to keep track of the input buttons. It increments the content of the memory locations reserved for each candidate whenever a-key is pressed.

Once a vote is a cast, it disables the input switches for a calculated interval of time to avoid tampering in the form of multiple ballots cast by one voter. Results can also be displayed using the 'Result' switch.

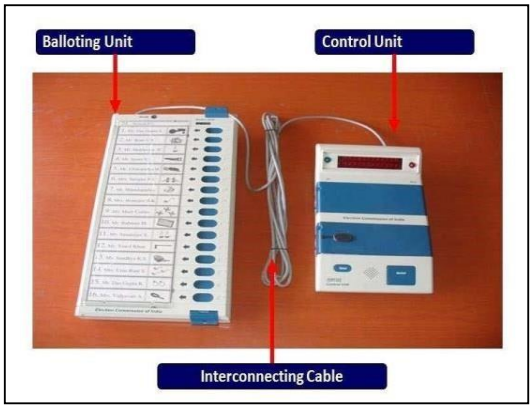


Figure 5. Control Unit and Ballot Unit

2. Ballot Unit - It is the ballot box that displays the list of all the candidates. It works as a facility to display various party names and symbols. A single Ballot Unit can contain up to 16 candidates and therefore, by connecting 4 ballot units, a total of 64-candidates can be accommodated in a single election.
Also, it is user-friendly and demands no training for the same.

VIII. ISSUES OF EVM

Electoral officials are examining around the world and trying to address various challenges such as system, and ease of preservation of the system along with an error-free operation and tamperproof results acceptability by common people in remote areas, system operation as close to the conventional ballot.

1. Bias - Many times, the candidates feel that the order in which their names are displayed on the EVM affects their vote counts because they believe that terms displayed first will be preferred more.

2. Verifiability - Earlier, before introducing the EVMs people, were confident that their votes were secured. The votes cast on paper or on punch cards were held accountable until the introduction of EVMs, after which the government had no oversight into the production of such systems handled by private companies beyond the choice of using them or not.

Candidates can know how many people have voted for them from any polling station, which results in misconduct with the voters due to lesser votes in any area. That is why EVMs are now created to display only the overall total results from an assembly or Sabha.

IX. NEED FOR FURTHER DEVELOPMENT

Even though the EVMs are designed for the electoral system of a country, it needs some alterations. The authentication of voters has to be stretched to another level, i.e. beyond the VOTER ID. Either the fingerprint could be scanned or the face detection system could be used to avoid multiple votes by a single voter.
Since now the EVMs are equipped with networking capabilities allowing voters to vote from anywhere in the world, there must be a more secure system and also, those votes must be brought in as soon as possible so that the result is not delayed.

X. FUTURE SCOPE

As we know that currently the total number of votes that can be catered by an EVM is 3840. So, to be able to provide EVMs to a larger audience in a single day such that the total duration of an electoral process is reduced, the memory and internal architecture also need to be altered. The main focus now is to increase the internal memory size of an electronic voting machine. The idea is to now incorporate some other method for authentic voting, such as iris detection or fingerprint scanner.

XI. RESULTS

The voting machine can be moderated as per the system and number of candidates. Also, post-2013 the M3 EVMs can serve a maximum of 384 candidates by connecting 24 ballot units together at a time. The control switches help in an easy and fast calculation of results and provide complete information about the total votes cast, displayed on the LCD screen. The clear button allows the machine to be used multiple times, thereby reducing costs.

ISSN(Online) : 2456-8910
International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE)
Volume 4, Issue 7, DOI:10.29027/IJIRASE.v4.i7.2021.836-840,January 2021

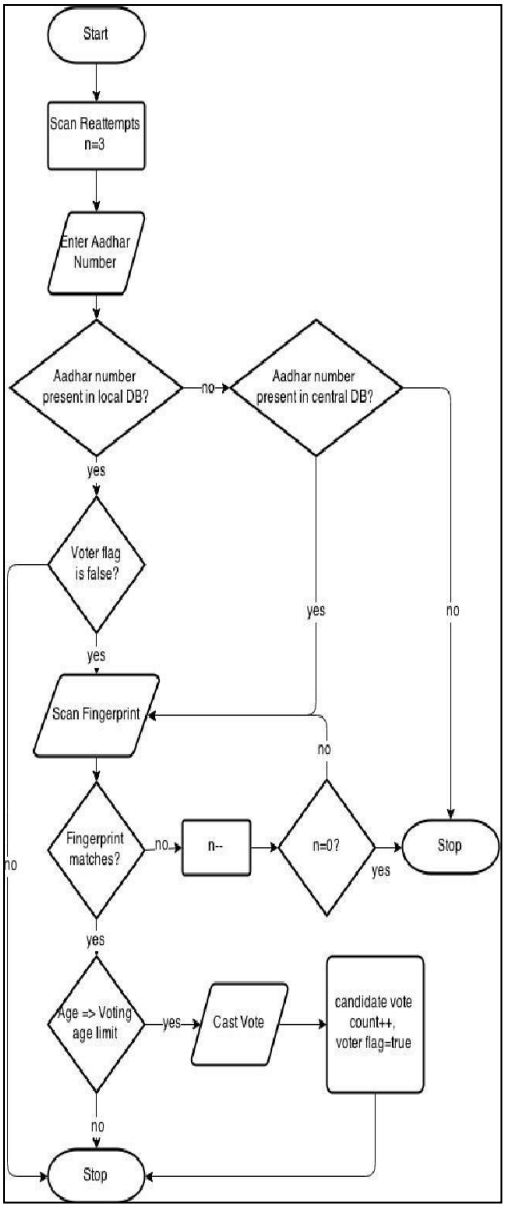


Figure 6. Flowchart of EVM Using Fingerprint Scanner for Better Authentication.

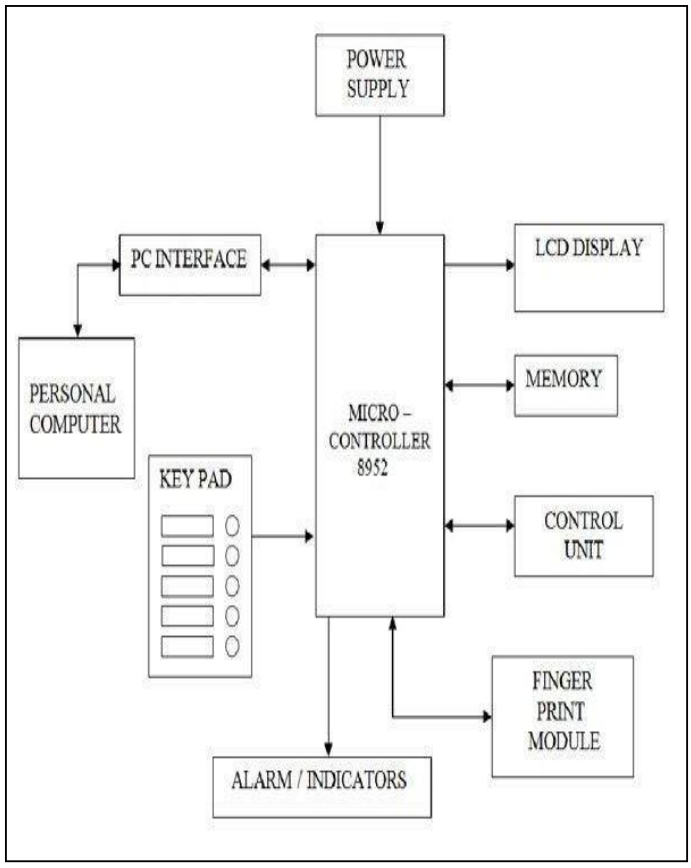


Figure 7. Block Diagram with Fingerprint Scanner

XII. CONCLUSION

The architecture and operation of a microcontroller-based Electronic Voting Machine have been outlined in this paper. This project is economical, and less manpower is required. Its compact size saves transportation cost and produces results in no time. EVMs have been implemented successfully with full legal support. They are fast and help in fair conduction of elections.

ACKNOWLEDGMENT

The authors gratefully acknowledge the contributions of Alfred J. Gillespie for the invention of EVMs and the Election Commission of India in alliance with Bharat Electronics Limited (BEL)-and Electronics Corporation of India Limited (ECIL)-for the origination of EVMs in India. Also, the authors would like to thank their mentor Professor Narendra Kumar for his constant guidance which made the completion of this paper possible.

REFERENCES

[1] Kraft, S., & Wall, E. (1986). Experimental microprocessor-based adaptive control system. *IEEE Control Systems Magazine*, *6*(1), 35-40.

[2]. Yung, R., Rusu, S., & Shoemaker, K. (2002, September). Future trend of microprocessor design. In *Proceedings of the 28th European Solid-State Circuits Conference* (pp. 43-46). IEEE..

[3]. Weaver, A. C. (1978, January). A real-time, multi-task programming language for microprocessor-based industrial process control. In *Proceedings of the 1978 annual conference-Volume 2* (pp. 522-525).

[4]. Chen, C. J., Cheng, W. M., Tsai, H. Y., & Wu, J. C. (2009). A quasi-delay-insensitive microprocessor core Implementation for Microcontrollers. *Journal of information science and engineering*, *25*(2), 543-557.

[5]. Lee, J. H., Lee, W. C., & Cho, K. R. (2002, August). A novel asynchronous pipeline architecture for CISC type embedded controller, A8051. In *The 2002 45th Midwest Symposium on Circuits and Systems, 2002. MWSCAS-2002.* (Vol. 2, pp. II-II). IEEE.

[6] Staunstrup, J., & Wolf, W. (2013). *Hardware/software co-design: principles and practice*. Springer Science & Business Media.

[6] Mazidi, M. A., Mazidi, J. G., & McKinlay, R. D. (2005). *The 8051 microcontroller and embedded systems using assembly and C*. Dept. of Computer Science and Information Engineering National Cheng Kung University, TAIWAN.

[7] Ali, M. M. (2007). *The 8051 Microcontroller and Embedded Systems: Using Assembly and C*. Pearson Education India.

[8] Calcutt, D., Cowan, F., & Parchizadeh, H. (2003). *8051 Microcontroller: An Applications Based Introduction*. Elsevier.

[9] Schultz, T. W. (2004). *C and the 8051*. PageFree Publishing, Inc..

Experimental Techniques
https://doi.org/10.1007/s40799-020-00411-7

RESEARCH PAPER

Estimation and Experimental Validation of Mean-Field Homogenised Effective Properties of Composite

A. Sharma[1] · V. Rastogi[1] · A.K. Agrawal[1]

Received: 27 December 2019 / Accepted: 26 October 2020
© The Society for Experimental Mechanics, Inc 2020

Abstract
For composite materials, the system response changes abruptly with a change in the properties of the material. Therefore, attaining significant knowledge about the effect of the material composition on the material properties is crucial. The researchers are looking for new computational methods which can predict these alterations so that the effort in experimental testing can be reduced. In this direction, this paper presents a robust and novel methodology of validating the estimation of the composite's effective through a multi-scale approach by a set of standardized experimentation. These effective properties are estimated through the mean-field homogenization technique whose parameters are driven from the image analysis of Scanning Electron Microscopy (SEM) images. The predicted results are validated with the results obtained by the experimentation as per ASTM E1876 standard. The estimated error between the predicted properties and the experimental values increased with the increase of alumina particle fraction in the matrix. The mean-field homogenization lags behind the experiments for the parameters defined by the image analysis method than the experimental results. The upper bounds of the mean-field homogenization can be used for the composites with higher reinforcement volume fraction.

Keywords Multi-scale analysis · Mean-field homogenization · Image analysis method · ASTM E1876 · Impact hammer test

Introduction

With the advent of composite materials, industries have started focusing on the use of lighter weight materials with the same mechanical properties. Metal matrix composites have the edge over the parent metals for rotor applications, as it has a higher specific modulus, tensile strength and other mechanical properties. Aluminium /alumina MMCs have shown prominent growth in the composite material market because of their compatibility to the rotor systems. [1–3]. Many manufacturers such as Duralcan, G.K.N., Toyota, GM and Nissan have used MMC materials to develop various components with both static and dynamic applications [4].

Metal matrix composites can be manufactured by various methods from liquid and semi solid-state of matrix mixed with solid reinforcement particles. There are various methods such as powder metallurgy, diffusion bonding, infiltration, stir casting and spray forming to manufacture MMC as per requirements. Stir casting is one of the most inexpensive methods which has been used for the fabrication of metal matrix composites. For complicated and extensive size design of metal matrix composites, stir casting is very convenient and useful [5].

A computation tool that provides results validated by experimental or real-time results saves a lot of time and energy for industries. Researchers seek for more tools that can provide accurate and reliable results for metal matrix composites properties. Multi-scale computation of composites is a two-phase process, where the effective macroscopic properties are derived through micro-mechanics. Analysis of composites with discrete reinforcements with random orientation has been found in archival literature [6–9]. There are various approaches through which these effective properties can be predicted efficiently [10]. For composites with a low volume fraction of inclusions, Eshelby's method [11] with a dilute scheme can be used to evaluate the effective properties. For composites with a higher volume fraction of inclusions, aggregate model [12, 13] can be used, which utilizes the known properties of constituents. Mori-Tanaka [14] have proposed a theory by assuming that the average strain in the inclusion is

✉ A. Sharma
anujsharma_phd2k16@dtu.ac.in

[1] Department of Mechanical, Production, Industrial and Automobile Engineering, Delhi Technological University, Delhi 110042, India

Published online: 23 November 2020

related to the average strain in the matrix by a fourth-order tensor. The relation between the uniform strains in the inclusion in the matrix is provided by this fourth-order tensor. It was further extended by Benveniste [15] to investigate the stress and strain concentration tensors along with the overall elastic modulus of composites. This mean-field homogenization theory has been adopted by various researches [16–20] to determine the effective properties such as modulus, density and thermal expansion and is used in this work.

Significant increase of tensile strength, yield strength and hardness on the addition of alumina particles in the aluminium matrix is validated experimentally in archival literature [21–23]. Properties such as young's modulus, poisson's ratio and shear modulus are computationally derived successfully by various researchers. However, experimental validation of these effective properties with standard practice and technique is equally important. These properties can be experimentally determined by impact hammer test based on ASTM E1876 standard [24].

Recently, *Praveen* et al. [25] performed an impact hammer test based on ASTM E1876 to evaluate the dynamic elastic properties of composites at ambient temperatures. However, the comparison of multi-scale computation of the effective properties and the respective experimental results have not been explored in detail. Daramola et al. [26] compared the mean-field homogenization technique with the experiment and estimated an error of 4.2% between the predicted and experimental values. However, the microscopic parameters such as volume fraction were assumed to be same as the fraction of fibres added during manufacturing. For metal matrix composites fabricated by stir casting method, there is much difference between the added alumina particles and the particles mixed with the matrix. Therefore, a unified and novel approach is adopted in this paper, where the effective properties of the composite are derived computationally using parameters derived through an image analysis method. Moreover, these computationally driven values are compared with the experimental results as per ASTM E1876 standard, evaluated by the impact hammer test.

Material and Method

The composite shaft samples (800 mm length and 22 mm diameter) are fabricated using a stir casting process, as shown in Fig. 1. The aluminium 6061 metal matrix is mixed through stirring with a different weight percentage of alumina (0 wt.%, 6 wt.% and 12 wt.% of alumina) supplied by Loba Chemie Pvt. Ltd. In order to increase the wettability of the alumina particles in the aluminium matrix, the alumina particles were preheated, and magnesium was added during the stir casting process. Magnesium forms an interface of MgO and $MgAl_2O_4$ between the alumina and aluminium, resulting in better wettability and strong bonding [27]. As shown in Fig. 2, the bonding between the reinforcement and the aluminium matrix is strong for both dispersed and agglomerated particles. The Scanning Electron Microscopy (SEM) images are taken for these samples, and then these images are analyzed through an image analysis method. The average aspect ratio and average volume fraction are determined through image analysis. These parameters are used to predict the effective properties of the composite through Mori-Tanaka mean-field homogenization. These properties are then compared with the experimental response of the composite specimen.

Estimation of Effective Mechanical Properties through Mean-Field Homogenization by Using Parameters Extracted through Image Processing

Cubic samples of 5 mm length, 5 mm width and 5 mm height were processed by EDM wire cutting and Scanning Electron Microscopy (SEM) micrographs at the magnification of 500X

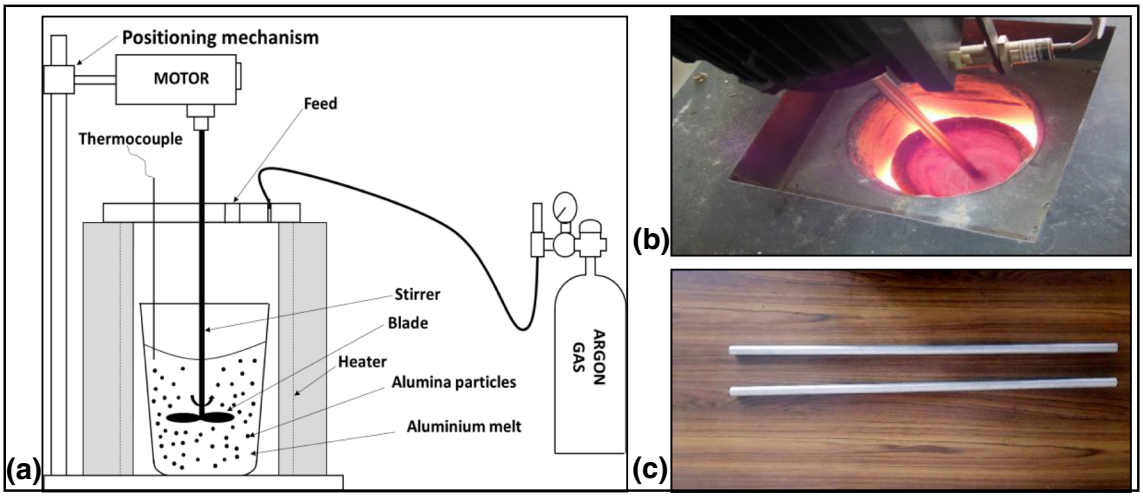


Fig. 1 (**a**) Stir casting set up (**b**) Vortex creation by a four-blade stirrer (**c**) Stir casted 800 x 22 mm shaft

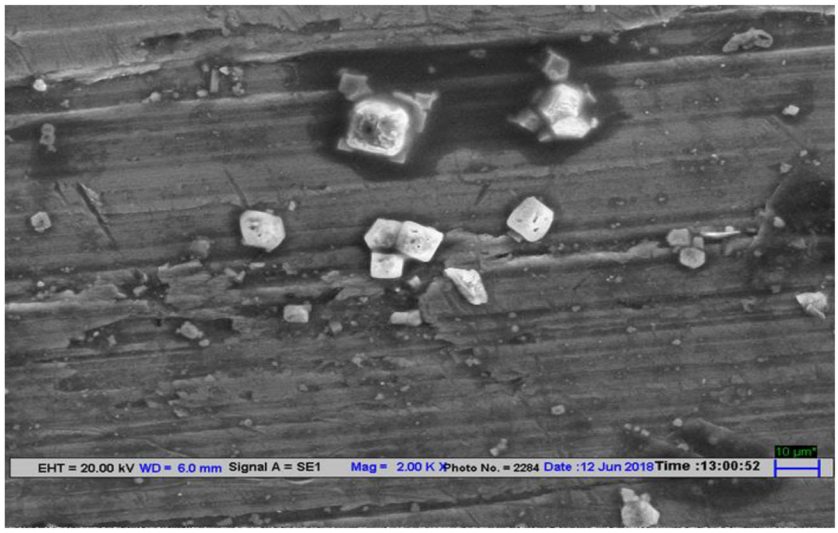


Fig. 2 2000X image showing strong bond between alumina particles and aluminium matrix

and 1000X were obtained using a Zeiss EVO 50 machine. Morphology of the micrographs depicts uniform distribution of reinforcements in the aluminium matrix. For each composition, three samples were scanned, and the fraction and size of alumina particles were evaluated by image processing and averaged for the computation of effective properties.

The open-source image processing software ImageJ was used to generate binary images of the composites with ellipsoid form, assumed for homogenization of properties. SEM micrographs were converted to binary images by using threshold of 61 foreground pixels and 255 background pixels. This transformed the different luminous pixels into white/black bifurcating with respect to the threshold values. At this value of threshold, particles were easily differentiated from the matrix material. Parameters such as the size of the reinforcement, aspect ratio and volume percentage for the composite were derived from the image analysis, as shown in Fig. 3. These parameters were incorporated in the mean-field homogenization of the metal matrix composite material where a Representative Volume Element (RVE) is assumed to be the building block of the macroscopic structure.

RVE is a statistical representation that connects microscopic structure and constituents with the macroscopic properties of composite materials. Therefore, the size of the RVE should

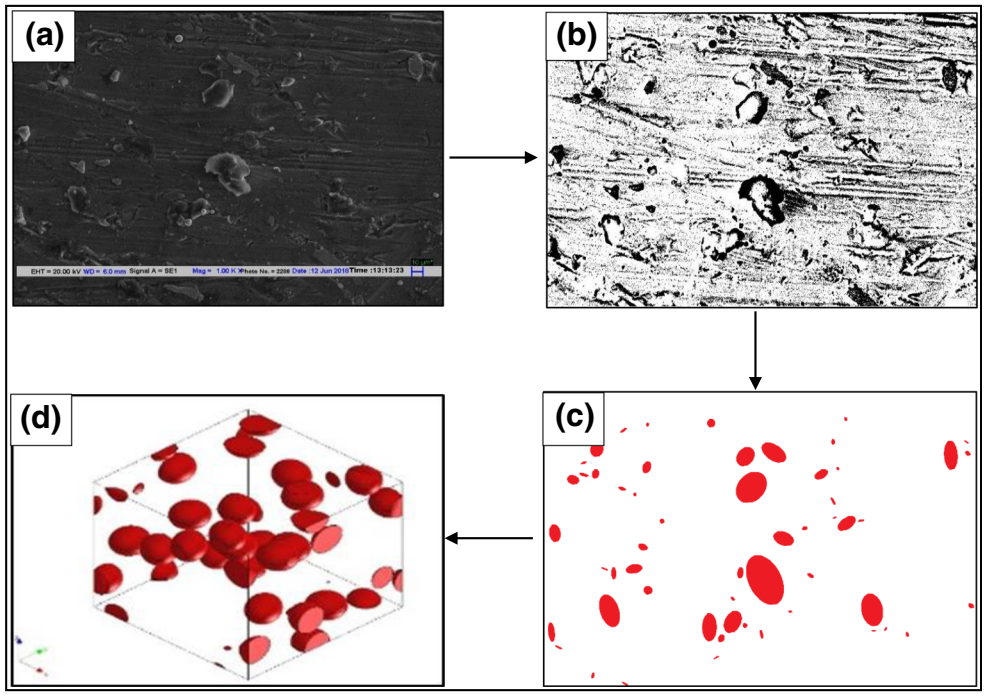


Fig. 3 Conversion of an SEM image into an RVE (**a**) SEM Image of a composite sample (**b**) Cropped SEM image converted into a binary image (**c**) Particle analysation of the binary image (**d**) RVE with mean aspect ratio, size and volume fraction

be such, that it is aptly smaller than the macroscopic structural dimensions and simultaneously large enough to contain an ample amount of information on the microstructure [28, 29].

Digimat is a composite modelling platform which has been used for the generation of RVEs with standard size and threshold. It consists of a Digimat-FE tool, which generates realistic RVEs with the finite element method by using the RVE generation algorithm. This algorithm is based on Random sequential adsorption (RSA) technique which randomizes the orientation of the particles [30]. Particles position themselves in random directions sequentially, and each particle blocks nearby locations from being occupied by another particle, therefore avoiding coagulation. This platform has been used for the development of RVEs in many types of researches. RVEs with randomly distributed spherical inclusions [31, 32], perfectly aligned fibres [33] or random fibre distributions by using a fibre randomization algorithm [34] are generated by using this technique. These RVEs can be exported as standard CAD formats. The effective properties of these RVEs were calculated and have been used for further analysis.

For mean-field homogenization, Mori-Tanaka method [14] has been adopted to determine the effective Young's modulus of the composite material. Therefore, it assumed that the metal matrix composite is comprised of N number of phases. Matrix phase is denoted by subscript m, and there are remaining N-1 inclusion phases in the control volume. The volume fractions of the matrix and the i_{th} inclusion are denoted by ϑ_m and ϑ_i. Similarly, stiffnesses of the matrix and the i_{th} inclusion are denoted by K_m and K_i. These stiffness quantities are generally represented as elasticity tensors of fourth-order which possess specific symmetrical properties. The overall elastic stiffness of the composite [35] is defined in Eq. (1).

$$\bar{\bar{K}}_{av} = \left(\vartheta_m * K_m + \sum_{i=1}^{N-1} \vartheta_i \langle K_i * D_i^{dil} \rangle\right) + \left(\vartheta_m * \boldsymbol{I} + \sum_{i=1}^{N-1} \vartheta_i \langle D_i^{dil} \rangle\right)^{-1} \tag{1}$$

Where $\boldsymbol{I}$ is the identity tensor and the dilute strain concentration factor D_i^{dil} is generally expressed in Eq. (2).

$$D_i^{dil} = \left[\boldsymbol{I} + S_i\, K_m^{-1}(K_i - K_m)\right]^{-1} \tag{2}$$

Here S_i is the Eshelby tensor [6] for the spheroidal inclusion in the infinite matrix.

Since the alumina particles are randomly oriented, orientation averaging has to be taken into account. Therefore, the terms in the angle parentheses $\langle\rangle$ denote the averaged value of the parameters over all possible orientations. This orientationally averaged tensor is a fourth-order tensor in the 3D space [35] and can be expressed as in Eq. (3).

$$\langle T_{klmn} \rangle = \frac{1}{2\pi} \int_0^{2\pi} \int_0^{2\pi} T_{klmn}(\alpha, \beta) sin\beta \, d\alpha \, d\beta \tag{3}$$

where α, β are the generalized angular orientations. For the randomization, the local coordinates should be transformed to the global coordinates. For overall random transformation, the matrix may be expressed as sown in Eq. (4).

$$a_{kl} = \begin{bmatrix} \cos\alpha & \cos\beta \sin\alpha & \sin\beta \sin\alpha \\ -\sin\alpha & \cos\beta \cos\alpha & \sin\beta \cos\alpha \\ 0 & -\sin\beta & \cos\beta \end{bmatrix} \tag{4}$$

Moreover, the transformation may be written, as shown in Eq. (5).

$$T_{klmn}(\alpha, \beta) = a_{ko} a_{lp} a_{mq} a_{nr}\, T'_{opqr} \tag{5}$$

The effective Young's Modulus(E_{efv}) [35] can be now calculated by the general equation (Eq. (6)) with shear modulus and bulk modulus.

$$E_{efv} = E_m * \frac{\kappa_{efv} * \mu_{efv}(3\kappa_m + \mu_m)}{3k_{efv} * \kappa_m + \mu_{efv} * \mu_m} \tag{6}$$

Effective Shear modulus can be derived using Eq. (7).

$$\frac{\kappa_{efv}}{\kappa_m} = \frac{1}{1 + z\, p} \tag{7}$$

Effective Bulk modulus can be derived using Eq. (8).

$$\frac{\mu_{efv}}{\mu_m} = \frac{1}{1 + z\, q} \tag{8}$$

Where E_{efv} is the effective Young's modulus, E_m is Young's modulus of the matrix, κ_{efv} and μ_{efv} are the effective bulk and shear moduli of the composite, z is the volume concentration/fraction, κ_m and μ_m are the bulk and shear modulus of the matrix. Also, p and q are the explicit expressions, and these expressions and details of the expression can be seen in Tandon and Weng [36] for spheroidal inclusions for detailed reference.

Also, effective density can be easily calculated through the rule of the mixture, as shown in Eq. (9).

$$\rho_{efv} = \rho_m \vartheta_m + \rho_i \vartheta_i \tag{9}$$

Where ρ_{efv} is the effective density, ρ_m and ϑ_m are the density and volume fraction of matrix and ρ_i and ϑ_i are the density and volume fraction of inclusions.

Experimental Evaluation of Elastic Properties

For the experimental validation of the estimated/predicted properties of the composites, ASTM E1876 impact hammer vibration analysis was employed. The analysis is performed on a standard experiment setup which conforms the prerequisites for the ASTM E1876 standard, as shown in Fig. 4.

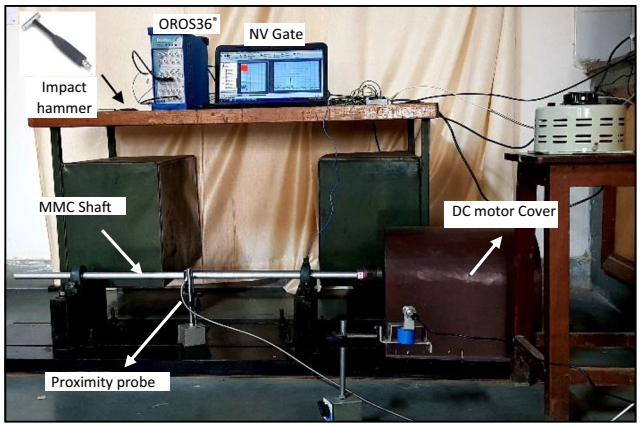


Fig. 4 Experimental set up for impact hammer test

The inductive proximity sensor (LJ12A3–4-Z /BY) having a sensitivity of 1000 V/m and a detection range of 4000 μm is used. The test specimens are energized through a force transducer with a plastic tip. Impact Hammer (Model: - PCB-086C03) with sensitivity 2250 μ(V)/(N) and range 4440 N is used for transient excitation. The response from the transducers is collected and analyzed by using a 32 channel, vibration analyzer OROS36®. The channels of these analyzers are handled in real-time: FFT, 1/3rd Octave. The fundamental natural frequencies of the specimen are observed from the peaks of the frequency response function obtained by using NV Gate® (version 10.1.1) interface.

The prime objective of the experimentation was to extract the values of natural frequencies for the first two bending (flexural) modes of the shaft. A range of 0–600 Hz was targeted as per the dimensions for experimental analysis to confirm that primarily these modes are excited. For excitation, the plastic tip of suitable hardness was used such that the input spectrum doesn't prominently excite the frequencies beyond the working range. The experimental samples were selected carefully such that the impact force input spectrum excites the required range of frequency at amplitude high enough to avoid noise interceptions. The sampling rate of 2.04 kS/s (kilo samples per second) was chosen, which was much higher than the Nyquist frequency according to the range excited by the input spectrum and the range of observation. This high rate of sampling prevented the problem of leakage and aliasing, and no anti-aliasing filters were required. The motive of the experimentation was also to observe the damping of the material relative to each other. The windowing technique would have altered the peak width and thus, the estimation of damping. Therefore, the time range was taken in such a way that windowing was not required for sample selection. The high resolution of the sampling ensured that the natural frequency is confidently determined by Gaussian peak fitting method. The position of the sensor has to be chosen in such a way that it should detect the first and second mode excitation. At the centre of the shaft, the first mode was highly excited, but the second mode was not excited (Since the antinode of the second mode is at the centre for the simply supported system). Therefore, a non-contact inductive proximity transducer was roved, and a suitable position was chosen where the excitation of first and second peaks was obtained successfully (At a distance of 50 mm from the centre as shown in Fig. 5. At this position, the amplitude of first bending natural frequency peak is the highest followed by second bending natural frequency peak.

The Young's modulus expression for the known fundamental flexural natural frequency is shown in Eq. (10).

$$E = 1.6067\left(\frac{l^3}{d^4}\right)\left(mf_f^2\right)T_1 \tag{10}$$

where E = Young's modulus, l = Length of the specimen, d = diameter of the shaft, m = mass of the shaft, f_f = fundamental

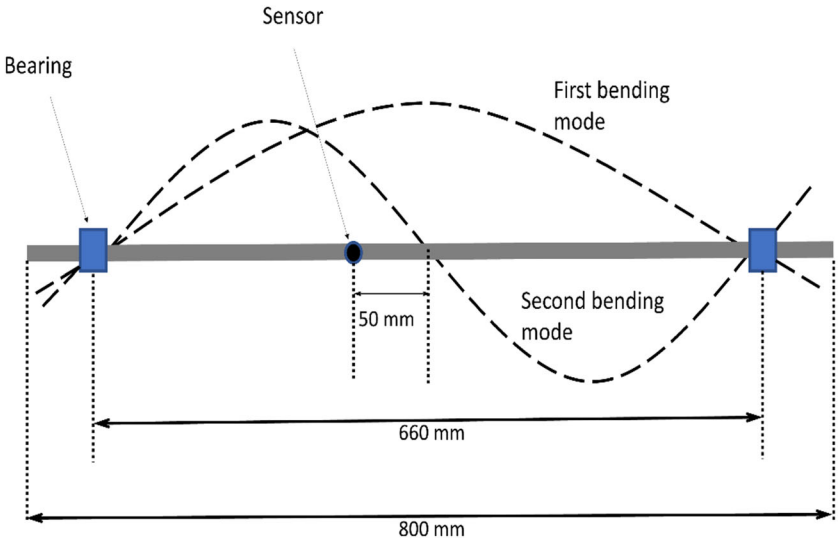


Fig. 5 First and second bending mode with sensor positioning

Table 1 Mean volume fraction, aspect ratio and particle size for various SEM images

Shaft Material		Mean volume fraction			Mean aspect ratio			Mean particle size (μm)		
		S1	S2	S3	S1	S2	S3	S1	S2	S3
Al 6061/6 wt.% Al_2O_3	Max	4.0	4.8	4.7	3.1	2.9	3.1	8.7	8.9	8.8
	Min	2.9	3.4	3.1	2.6	2.6	2.7	7.3	7.2	7.1
	Mean	3.4	4.1	3.8	2.9	2.7	3.0	8.0	8.1	8.1
Al 6061/12 wt.% Al_2O_3	Max	8.3	9.4	9.9	3.4	3.2	3.3	8.9	8.8	9.0
	Min	6.9	7.4	7.9	2.8	2.7	2.9	7.3	7.5	7.4
	Mean	7.3	8.1	9.1	3.2	3.0	3.1	8.3	8.2	8.4

flexural frequency of the shaft and T_1 = correction factor for the fundamental flexural mode to account for finite diameter of the rod, and Poisson's ratio.

Since the aspect ratio (l/d) is higher than 20, the simplified form for T_1 can be used as shown in Eq. (11).

$$T_1 = \left[1 + 4.939\left(\frac{d}{l}\right)^2\right] \tag{11}$$

The expression for Dynamic shear modulus for known fundamental torsion resonant frequency of a cylindrical is shown in Eq. (12).

$$G = 16mf_t^2\left(\frac{l}{\pi d^2}\right) \tag{12}$$

where G = Dynamic shear modulus, l = Length of the specimen, d = diameter of the shaft, m = mass of the shaft, f_t = fundamental torsional frequency of the shaft.

Correspondingly, Poisson's Ratio can be determined using Eq. (13).

$$\mu = \left(\frac{E}{2G}\right) - 1 \tag{13}$$

Where μ is the Poisson's ratio.

Results and Discussion

Since there might be inhomogeneous zones within the sample which lead to interference of the results, measures are taken to avoid these uncertainties. Three samples were taken for each material for SEM analysis. The SEM images were randomly taken for different positions within the sample and all six surfaces of the cuboid sample with 1000x magnifications. For each sample (S1, S2 and S3), maximum, minimum, and mean values of volume fraction, aspect ratio and particle size are taken for various SEM micrographs as shown in Table 1.

The aspect ratio and particle size doesn't vary much for different samples and materials. The mean volume fraction increased for the Al 6061/12 wt.% Al_2O_3 as compared to the Al 6061/6 wt.% Al_2O_3. The variation of the microstructural parameters within the sample were not resulting in interference with the microstructural parameters of different composition. For example, the mean volume fraction ranges from 2.9 to 4.8 for Al 6061/6 wt.% Al_2O_3 and 6.9 to 9 for Al 6061/12 wt.% Al_2O_3. The data is therefore categorical, and we can classify the results for different composition of materials respectively.

The modulus of elasticity is highest for the largest volume fraction, highest aspect ratio and smallest mean particle size and vice versa. Therefore, the overall highest young's modulus can be predicted by taking mean volume fraction as 4.8, aspect ratio as 3.1 and mean particle size as 7.1 μm and vice versa for Al 6061/6 wt.% Al_2O_3. Also, overall highest young's modulus can be predicted by taking mean volume fraction as 9.7, aspect ratio as 3.4 and mean particle size as 9.0 μm and vice versa for Al 6061/12 wt.% Al_2O_3. The consolidated results for the estimated microscopic properties through image analysis and mean-field homogenization are shown in Table 2.

On the addition of 6 wt.% of alumina, the overall mean volume fraction of 3.8% is observed by image analysis of

Table 2 Image analysis overall mean results

Material	Mean volume fraction	Mean aspect ratio	Mean particle size (μm)
Al 6061	–	–	–
6 wt.% Al_2O_3/Al 6061	3.8	2.9	8.1
12 wt.% Al_2O_3/Al 6061	8.166	3.1	8.3

Table 3 Mean-field homogenization results

Engineering constants	Al 6061	6 wt.% Al_2O_3/Al 6061			12 wt.% Al_2O_3/Al 6061		
		Mean	Max	Min	Mean	Max	Min
In-Plane Young's modulus (GPa)	70	73.73	74.84	72.9	78.66	80.1	77.4
Out -Plane Young's modulus (GPa)	70	73.61	74.71	72.78	78.32	79.75	77.06
In-Plane Poisson's ratio	0.328	0.326	0.325	0.326	0.322	0.321	0.323
Transverse Poisson's ratio	0.328	0.326	0.325	0.326	0.322	0.321	0.323
In-Plane shear modulus (GPa)	26.35	27.8	28.21	27.48	29.75	30.29	29.27
Transverse shear modulus (GPa)	26.35	27.76	28.17	27.44	29.2	29.73	28.73
Global Density (Kg/m^3)	2710	2748	2772	2723	2798	2838	2753

SEM images for different samples. The observed volume fraction of alumina in fabricated samples is low as compared to the volume fraction of alumina added during stir casting. The limitation of stir casting method that all alumina particles are not mixed with the matrix is the reason for this deviation. Although stir casting is flexible enough and large-sized samples used for analysis can be easily manufactured through this method. Similarly, for the addition of 12 wt.% of alumina particles, the mean volume fraction of 8.166% is observed. The mean aspect ratio and the average particle size are also in the conformance with the alumina particles manufacturer specification.

The effective Young's Modulus, Poisson's ratio and the shear modulus of in-plane and transverse plane for a particular RVE is estimated through mean-field homogenization as shown in Table 3.

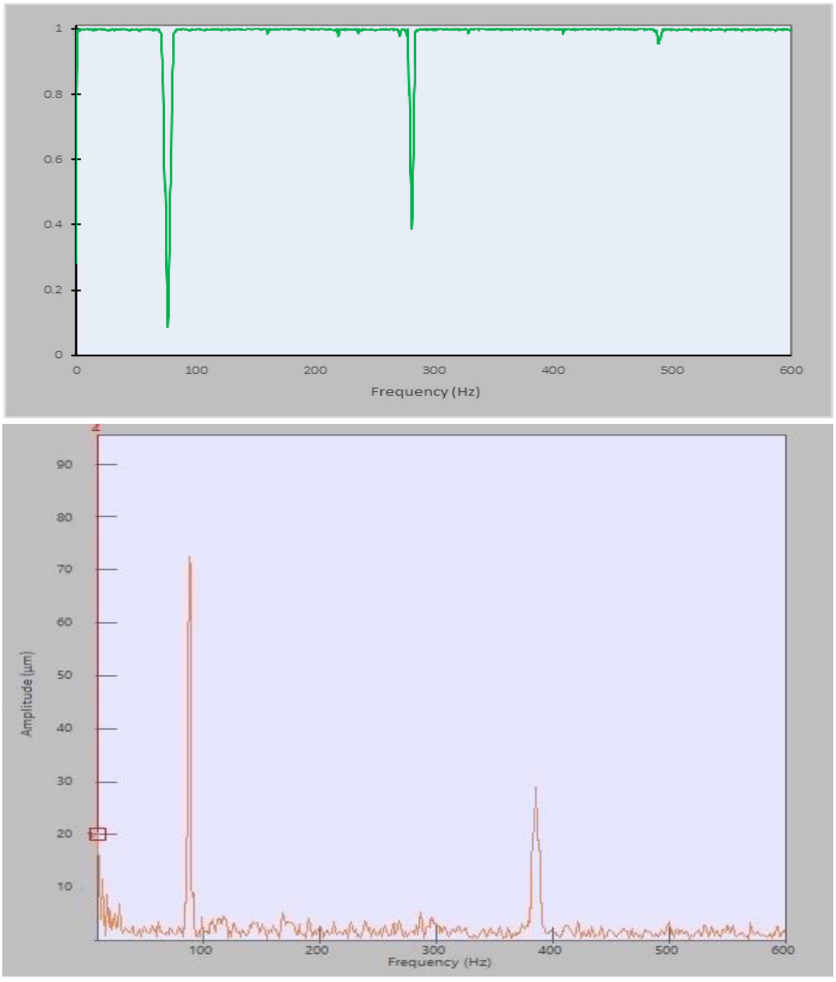


Fig. 6 FRF and coherence for Al6061

The estimated mean In-plane Young's modulus is increased by 5.32%, on the addition of 6 wt.% of alumina due to enhanced dislocation density and precipitation hardening. Since the orientation of the particles is randomized, the value of out of plane Young's modulus is approximately the same as values of In-plane Young's modulus. The global density of the composite is also estimated to be increased by the addition of the particles since the density of the alumina particles is more than the aluminium matrix. The mean In-Plane shear modulus is also increased by 5.5% with transverse shear modulus value approximately similar to the In-plane shear modulus. Moreover, the effective In-plane Young's Modulus is estimated to be increased by 12.37% on addition of 12 wt.% of alumina. The highest value of density and shear modulus is observed for the composite with 12 wt.% of alumina reinforcements.

As shown in Table 3, there is no interference in respective values of Young's modulus, Poisson's ratio and shear modulus for different materials. For example, the minimum value of In-plane young's modulus of 12 wt.% Al_2O_3/Al 6061 is 77.4 GPa, whereas the maximum value of In-plane young's modulus of 6 wt.% Al_2O_3/Al 6061 is 74.84 GPa. So, this data can be classified by the composition of the material.

The strengthening effect of reinforcement is not directionally dependent as in the case of fibre reinforced composites due to the random orientation of particles. By adding 12 wt.% of alumina, In-Plane Young's modulus is increased by 12.37%, but the global density is increased by 3.2%. It implies that the addition of reinforcement improves the specific modulus of the material and therefore improves the modal response of the material. Therefore, impact hammer test is performed to obtain the properties of composites through modal response.

The frequency response function and the coherence for Al6061, Al6061/6 wt.% Al_2O_3 and Al6061/ 12 wt.% Al_2O_3 composite in the impact hammer test is shown in Figs. 6, 7 and 8. The input and output spectrums were coherent for all shafts with the resonance value dropped at anti-nodes. The high coherence between the input and output spectrum, except at the anti-nodes confirms the replicability of the experiments. The high resolution of the frequency spectrum was obtained because of the high sampling rate. The amplitude and the natural frequency were calculated by determining the peak using the Gaussian peak fitting method, and the damping ratio is obtained by the half-power bandwidth method.

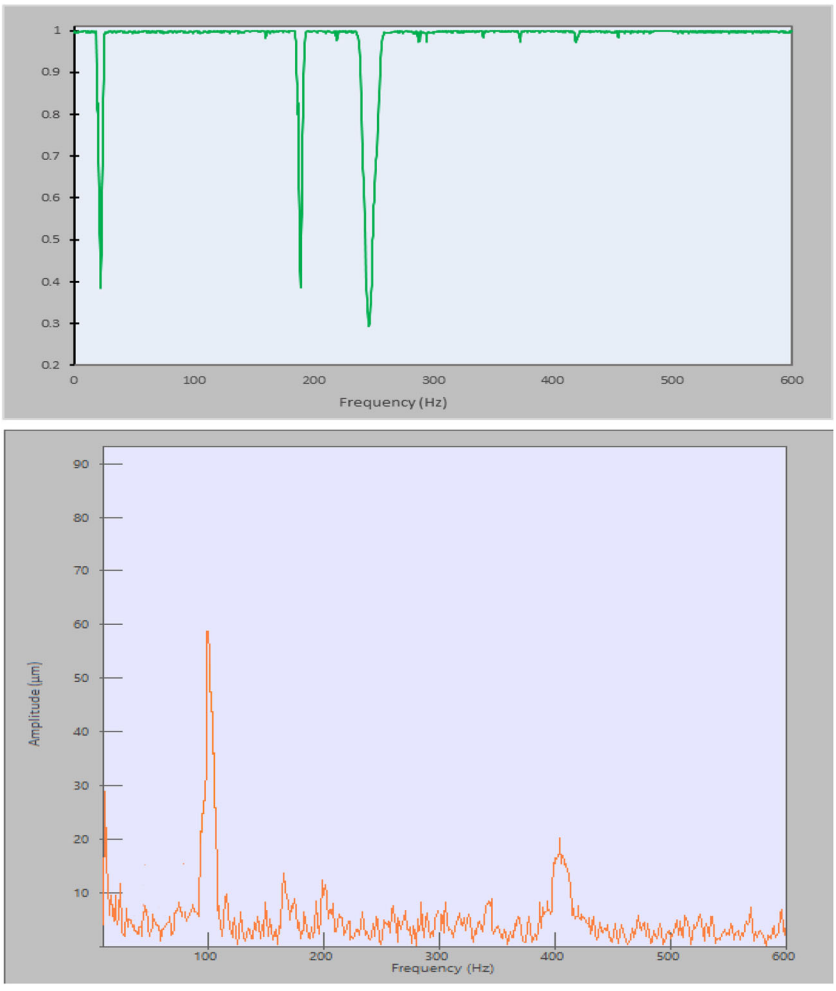


Fig. 7 FRF and coherence for Al6061/ 6 wt. % Al_2O_3

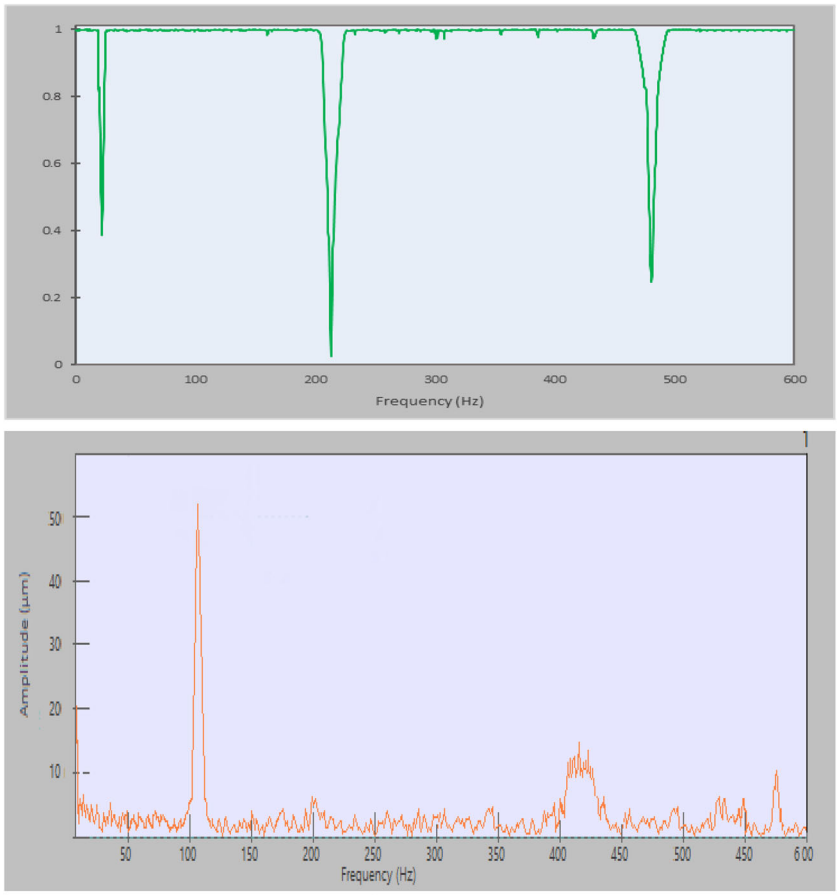


Fig. 8 FRF and coherence for Al6061/ 12 wt. % Al_2O_3

For each shaft with different composition, thirty experiments were taken for the determination of natural frequencies, damping ratio and amplitudes. The mean, standard deviation and range of the natural frequencies for different sets of experimentations are shown in Table 4.

The low standard deviation was obtained for the values of natural frequencies for both first and second mode. Also, the range of natural frequency values for different composition doesn't interfere with each other. The natural frequency increases significantly with the increase of reinforcement particles in the shafts. This is due to the reason that the addition of alumina particles in the Al6061 matrix leads to a substantial improvement in the bending stiffness as compared to the material density. The mean, standard deviation and range of the damping for different sets of experimentations are shown in Table 5.

The higher value of damping was observed for the specimen with higher wt.% of alumina particles in the aluminium matrix. The damping was significantly increased for the second mode of vibration. It suggests that the damping increases with the increase of alumina particles in the aluminium matrix. This is due to the grain refinement and formation of secondary phases which increases elastic strain energy dissipation in the materials.

The mean, standard deviation and range of the amplitude of vibrations for different sets of experimentations are shown in Table 6.

Table 4 Experimental natural frequency mean, standard deviation and range

Shaft Material	Natural frequency (first mode)			Natural frequency (second mode)		
	Mean	Standard deviation	Range	Mean	Standard deviation	Range
Al6061	98.62	0.65	97.6–99.1	388.93	1.92	386.1–391.6
Al 6061/6 wt.% Al_2O_3	101.65	0.72	100.1–102.8	408.43	2.42	406.3–411.7
Al 6061/12 wt.% Al_2O_3	108.91	1.01	107.1–110.7	420.21	3.02	416.2–425.1

Table 5 Experimental damping mean, standard deviation and range

Shaft Material	Damping (first mode)			Damping (second mode)		
	Mean	Standard deviation	Range	Mean	Standard deviation	Range
Al6061	0.0097	0.0001	0.0095–0.0099	0.021	0.0003	0.0207–0.026
Al 6061/6 wt.% Al_2O_3	0.0123	0.0004	0.0119–0.0129	0.0233	0.0005	0.0225–0.0242
Al 6061/12 wt.% Al_2O_3	0.0146	0.0006	0.0137–0.155	0.026	0.0008	0.0252–0.0262

Table 6 Experimental amplitude mean, standard deviation and range

Shaft Material	Amplitude (first mode) (μm)			Amplitude (second mode) (μm)		
	Mean	Standard deviation	Range	Mean	Standard deviation	Range
Al6061	69.1	2.11	67.1–70.9	30.21	2.01	31.2–27.1
Al 6061/6 wt.% Al_2O_3	62.7	2.03	60.7–64.2	21.62	2.42	22.5–19.1
Al 6061/12 wt.% Al_2O_3	52.1	2.81	50.2–54.8	14.29	2.65	12.31–17.01

The amplitude reduced significantly with the increase in alumina particles for both the modes. The range and the standard deviation are well within the limit that is required for the data to be categorical about the composition of the material. This is due to the increase in bending stiffness of the shafts because of the reinforcing effect of the alumina particles in the matrix.

These values are utilized to obtain the value of In-plane Young's modulus, shear modulus and Poisson's ratio experimentally, as per ASTM E1876 standard. The comparison of predicted values obtained through mean-field homogenization and the experimental values obtained is shown in Table 7.

The experimental value of In-plane Young's modulus of Al 6061 is observed to be 2.88% lower than that of initial prediction. This lower value can be attributed to the defects induced during the casting process of aluminium alloy. However, these defects might have prevailed for the case of 6 wt.% Al_2O_3/Al 6061 also, but the experimental In-plane Young's modulus has increased by 1.19% as compared to the predicted value. Similarly, experimental In-plane Young's modulus has increased by 3.75% as compared to the predicted value for the case of 12 wt.% Al_2O_3/Al 6061. From literature [26], it is observed that the addition of 6 wt.% micro-sized kaolinite particle in epoxy polymer matrix composite increased the effective elastic modulus from 3100 MPa to 3107.5 MPa. The mean-field homogenization method predicted the value 4.2% higher than the experimental value. Whereas in the present study, the mean-field homogenization predicted the value 1.19% lower than the experimental values. The reason for predicting a lower value by the homogenization process may be that the actual alumina particle volume fraction is higher than the volume fraction estimated by image analysis.

The range of the experimental and the predicted values for the In-plane Young's modulus is shown in Fig. 9. For Al 6061, the predicted value is the standard value for Young's modulus, and the experiment values are of the cast material. Therefore, the values are lower due to the introduction of some porosity. For, Al 6061/ 6 wt.% Al_2O_3 the experimental range intersects with the predicted value range. For, Al 6061/ 12 wt.% Al_2O_3 the experimental range exceeds the predicted

Table 7 Comparison of Predicted and experimental value

Engineering constants	Al 6061		6 wt.% Al_2O_3/Al 6061		12 wt.% Al_2O_3/Al 6061	
	Predicted	Experimental	Predicted	Experimental	Predicted	Experimental
In-Plane Young's modulus	70 GPa	67.98 GPa	73.73 GPa	74.61 GPa	78.66 GPa	81.61 GPa
In-Plane shear modulus	26.35 GPa	25.55 GPa	27.82 GPa	29.01 GPa	29.75 GPa	30.91 GPa
In-Plane Poisson's ratio	0.328	0.327	0.326	0.323	0.322	0.320



Fig. 9 Comparison of range of predicted and experimental value of In-plane Young's modulus

value range at a few points. Therefore, it can be concluded that the upper bounds of mean-field solutions should be used for the higher reinforcement compositions.

The similar kind of trend is observed for In-plane shear modulus, and the mean-field homogenization analysis does not deviate significantly from the experimental results.

Conclusion

The composite specimens have been successfully fabricated by stir casting method as per ASTM E1876 standards. The microstructure of the composites was taken through SEM and particles average size; aspect ratio and fraction were obtained through image analysis method. The effective properties were then evaluated by mean-field homogenization method by using these parameters. These properties were then evaluated experimentally as per ASTM E1876 standards. Following conclusions can be driven from the analysis in this paper:

- The volume fraction of the alumina particles observed by the image analysis method was significantly less as compared to the added volume fraction during fabrication. It suggests that the volume fraction added initially during fabrication should not be used directly for predicting the effective properties.
- The addition of alumina particles results in the strengthening of composite specimens due to enhanced dislocation density and precipitation hardening.
- The difference between the In-Plane and transverse effective properties is negligible since the strengthening effect of reinforcement is not directionally dependent due to the random orientation of particles.
- The addition of alumina particles in the Al6061 matrix leads to a substantial improvement in the bending stiffness as compared to the material density. Due to this reason, the natural frequency for various modes of vibration increased on the addition of particles.
- Damping ratio also increased on the addition of alumina particles due to the grain refinement and formation of secondary phases.
- The estimated error for the case of 6 wt.% Al_2O_3/Al 6061 is 1.19%, whereas the estimated error for the case of 6 wt.% Al_2O_3/ Al6061 is 3.75% for the mean values.
- The range of mean-field values lags behind the experimental values as the material of high reinforcement composition is used. It can be concluded that the upper bound of mean-field homogenization can be used for the material of high reinforcement composition and can be the future scope of work. However, mean-field homogenization predicts a low higher than experimental value while assuming that all alumina particles added for process are mixed into the matrix.
- Also, samples were taken from the end of the shaft for SEM analysis; this might be the low/high composition area of the shaft during casting. This variation can be addressed in future work.
- The variation of the experimental and the predicted values was such that categorization of the values was possible in reference to the composition of the shaft.

Author's Contribution All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Mr. Anuj Sharma, Prof. Vikas Rastogi and Prof. Atul kumar agrawal The first draft of the manuscript was written by Mr. Anuj Sharma and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they do not have any conflict of interest.

References

1. Chawla N, Shen YL (2001) Mechanical behavior of particle reinforced metal matrix composites. Adv Eng Mater 3(6):357–370. https://doi.org/10.1002/1527-2648(200106)3:6<357::AID-ADEM357>3.0.CO;2-I
2. Rawal S (2001) Metal-matrix composites for space applications. Jom 53(4):14–17. https://doi.org/10.1007/s11837-001-0139-z
3. Sarmasti AB, Yazdanirad M, Nouri Khezrabad M, Karbalaie M (2011) Effect of alumina particle size and thermal condition of casting on microstructure and mechanical properties of stir cast Al-Al2O 3 composites. Mater Sci Technol 27(11):1653–1656. https://doi.org/10.1179/1743284710Y.0000000020
4. Koczak MJ, Khatri SC, Allison JE, Bader MG (1993) Metal-Matrix Composites for Ground Vehicle, Aerospace, and Industrial Applications. Fundamentals of Metal-Matrix Composites, 297–326. https://doi.org/10.1016/b978-0-08-052371-2.50020-1
5. Singh S, Singh I, Dvivedi A Design and development of novel cost-effective casting route for the production of metal matrix composites (MMCs) (2017). Int J Cast Metals Res 6(30):356–364. https://doi.org/10.1080/13640461.2017.1323605

6. Chaturvedi SK, Tzeng GY (1991) Micromechanical modeling of material damping in discontinuous fiber three-phase polymer composites. Compos Eng 1(1):49–60. https://doi.org/10.1016/0961-9526(91)90025-N
7. Guden M, Hall IW (1998) Dynamic properties of metal matrix composites: a comparative study. Mater Sci Eng A 242(1–2): 141–152. https://doi.org/10.1016/s0921-5093(97)00516-9
8. Lesieutre GA, Yarlagadda S, Yoshikawa S, Kurtz SK, Xu QC (1993) Damped structural composite material using resistively-shunted piezoelectric ceramic short fibers. Collect Tech Papers - AIAA/ASME Structures, Structural Dynamics Mater Conf (pt 6): 3238-3243
9. Vaidya UK, Dadzie P-E, Haque A, Mahfuz H, Jeelani S (1997) Nondestructive evaluation and characterization of microfibers modified textile carbon-phenolic and carbon-carbon composites. J Reinf Plast Compos 16(11):968–1001. https://doi.org/10.1177/073168449701601101
10. Mura T (2013) Micromechanics of defects in solids. Mechanics of Elastic and Inelastic Solids. Springer Science & Business Media. https://doi.org/10.1007/978-94-009-3489-4
11. Eshelby JD (1957) The determination of the elastic field of an ellipsoidal inclusion, and related problems. Proc Royal Soc A: Math Phys Eng Sci 241(1226):376–396. https://doi.org/10.1098/rspa.1957.0133
12. Christensen RM, Waals FM (1972) Effective stiffness of randomly oriented fibre composites. J Compos Mater 6(3):518–535. https://doi.org/10.1177/002199837200600307
13. Halpin JC, Jerine K, Whitney JM (1971) The laminate analogy for 2 and 3 dimensional composite materials. J Compos Mater 5(1):36–49. https://doi.org/10.1177/002199837100500104
14. Mori T, Tanaka K (1973) Average stress in matrix and average elastic energy of materials with misfitting inclusions. Acta Metall 21(5):571–574. https://doi.org/10.1016/0001-6160(73)90064-3
15. Benveniste Y (1987) A new approach to the application of Mori-Tanaka's theory in composite materials. Mech Mater 6(2):147–157. https://doi.org/10.1016/0167-6636(87)90005-6
16. Abaimov SG, Trofimov A, Sergeichev IV, Akhatov IS (2019) Multi-step homogenization in the Mori-Tanaka-Benveniste theory. Compos Struct. https://doi.org/10.1016/j.compstruct.2019.03.073
17. Imani SM, Goudarzi AM, Rabiee SM, Dardel M (2018) The modified Mori-Tanaka scheme for the prediction of the effective elastic properties of highly porous ceramics. Ceram Inter 44(14):16489–16497. https://doi.org/10.1016/j.ceramint.2018.06.066
18. Kamarian S, Shakeri M, Yas MH (2016) Natural frequency analysis and optimal design of CNT/fiber/polymer hybrid composite plates using Mori-tanaka approach, GDQ technique, and firefly algorithm. Polym Compos 39(5):1433–1446. https://doi.org/10.1002/pc.24083
19. Tran VP, Brisard S, Guilleminot J, Sab K (2018) Mori–Tanaka estimates of the effective elastic properties of stress-gradient composites. Int J Solids Struct 146:55–68. https://doi.org/10.1016/j.ijsolstr.2018.03.020
20. Sobhaniaragh B, Batra RC, Mansur WJ, Peters FC (2017) Thermal response of ceramic matrix nanocomposite cylindrical shells using Eshelby-Mori-Tanaka homogenization scheme. Compos Part B 118:41–53. https://doi.org/10.1016/j.compositesb.2017.02.032
21. Yılmaz O, Buytoz S (2001) Abrasive wear of Al_2O_3-reinforced Aluminium-based MMCs. Compos Sci Technol 61:2381–2392
22. Kok M (2005) Production and mechanical properties of Al_2O_3 particle-reinforced 2024 Aluminium alloy composites. J Mater Process Technol 161:381–387
23. Suthar J, Patel KM (2018) Processing issues, machining, and applications of aluminium metal matrix composites. Mater Manuf Process 33(5):499–527. https://doi.org/10.1080/10426914.2017.1401713
24. Standard Test Method for Dynamic Young's Modulus, Shear Modulus, and Poisson's Ratio by Impulse Excitation of Vibration. ASTM E1876, ASTM International, United States, 1–17
25. Praveen AP, Rajamohan V, Mathew AT (2019) Material and mechanical characterization of multi-functional carbon nanotube reinforced hybrid composite materials. Exp Tech 43(3):301–314. https://doi.org/10.1007/s40799-019-00316-0
26. Daramola OO, Olajide JL, Adediran AA, Adewuyi BO, Ayodele TT, Desai DA, Sadiku ER (2020) Multiscale analysis and experimental validation of the effective elastic modulus of epoxy-dioctahedral phyllosilicate clay composite. Heliyon 6(6):e04008. https://doi.org/10.1016/j.heliyon.2020.e04008
27. Sangghaleh A, Halali M (2009) Effect of magnesium addition on the wetting of alumina by aluminium. Appl Surf Sci 255:8202–8206. https://doi.org/10.1016/j.apsusc.2009.05.044
28. Bouhfid N, Raji M, Boujmal R, Essabir H, Bensalah M-O, Bouhfid R, Qaiss AeK. (2019) Numerical modeling of hybrid composite materials. Modelling of Damage Processes in Biocomposites, Fibre-Reinforced Composites and Hybrid Composites, 57–101. https://doi.org/10.1016/b978-0-08-102289-4.00005-9
29. Gitman IM, Askes H, Sluys LJ, Valls OL (2004) The concept of representative volume for elastic, hardening and softening materials. International Summer School - IPME conference. XXXII International Summer School - IPME Conference: Advanced Problems in Mechanics. p 1-5
30. Babu KP, Mohite PM, Upadhyay CS (2018) Development of an RVE and its stiffness predictions based on mathematical homogenization theory for short fibre composites. Int J Solids Struct 130-131:80–104. https://doi.org/10.1016/j.ijsolstr.2017.10.011
31. Klusemann B, Svendsen B (2010) Homogenization methods for multi-phase elastic composites: comparisons and benchmarks. Tech Mech 30(4):374–386
32. Yi I, Wiedmaier J, Schmauder S (2015) Modeling and simulation of mechanical properties of nano particle modified polyamide 6. J Mater Sci Chem Eng 3(1):80–87
33. Ogierman W, Kokot G (2016) A study on fiber orientation influence on the mechanical response of a short fiber composite structure. Acta Mech 227(1):173–183
34. Soni G, Singh R, Mitra M, Falzon BG (2014) Modelling matrix damage and fibre matrix interfacial decohesion in composite laminates via a multi-fibre multilayer representative volume element (M2RVE). Int J Solids Struct 51(2):449–461
35. Chang CY, Chang MY, Huang JH (2004) Vibration analysis of rotating composite shafts containing randomly oriented reinforcements. Compos Struct 63(1):21–32
36. Tandon GP, Weng GJ (1986) Average stress in the matrix and effective moduli of randomly oriented composites. Compos Sci Technol 27(2):111–132. https://doi.org/10.1016/0266-3538(86)90067-9

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

Applied Water Science (2021) 11:8
https://doi.org/10.1007/s13201-020-01334-9

ORIGINAL ARTICLE

Evaluation of water quality and potential metal contamination in ecologically important Bhindawas bird sanctuary, India

Vandana Shan[1] · S. K. Singh[1] · A. K. Haritash[1]

Received: 22 November 2018 / Accepted: 26 November 2020
© The Author(s) 2021

Abstract
Considering the environmental, agricultural, and ecological significance of Bhindawas wetland, the present study is the first comprehensive investigation to assess the water quality, determine the suitability of water for aquatic life in the wetland; and its suitability for irrigation in areas around the wetland. Twenty samples of water from Bhindawas wetland were analyzed and spatial variations of dissolved oxygen (DO), dissolved phosphate, nitrate, and biological oxygen demand (BOD_5) were observed. The concentration of DO was higher in areas with shallow depth and rich growth of submerged vegetation compared to deeper areas with no vegetation. Spatial variations of phosphate, nitrate, and heavy metals correlated with nesting zone of birds, runoff from agricultural fields, and wastewater from adjoining villages, respectively. Values of heavy metal pollution index (HPI), heavy metal evaluation index (HEI), and degree of contamination (C_d) in water confirmed high level of metal contamination of the medium. Based on the water quality index (WQI), the water was unsuitable for aquatic life and use in agricultural utilization. It can be concluded that water quality of Bhindawas wetland was adversely affected by heavy metals, which is a cause of concern since this wetland is a temporary resort of migratory birds. Immediate intervention is required to improve the water quality, especially scrutiny and inspection of the added wastewater from surrounding villages, and runoff from adjoining agricultural fields.

Keywords Bhindawas wetland · Water quality index (WQI) · Heavy metal pollution index (HPI) · Heavy metal evaluation index (HEI) · Principal component analysis (PCA)

Introduction

Water is an essential natural resource for every living organism on this planet. It is the most important factor for sustaining life. Fortunately, a huge amount of water is present on earth's surface, but safe drinking water is available as a limited fraction in rivers, lakes, streams, and wetlands. The wetlands are most diverse and highly productive aquatic ecosystems and are known as natural sponges of ecosystem as they absorb excess water, store nutrients/pollutants from agricultural runoff and drainage system. Wetlands act as a buffer between land and water ecosystem and maintain sediment and nutrient balance between them (Haritash et al. 2015). These act as habitats for birds and provide a platform to native as well as migratory birds for resting, breeding, nesting, and rearing young ones. Water quality is a major factor in managing wetlands, which is affected by a number of parameters like seasons (Sharma et al. 2015), topography, type of soil, quantity of water, and vegetation present in it (Shelef et al. 2011). Increasing stress due to burgeoning population, intensified anthropogenic activities, urbanization, rapid industrialization, lack of proper legislation and lack of awareness contributed to rapid decline in water quality which ultimately will impose harmful effects on health of inhabitants (Senthil et al. 2012; Giri and Singh 2014). Many researchers assessed the water quality in order to draw out some information for management and conservation of wetlands. The water quality is an important tool, which classifies not only the type of water but also determines health status of the ecosystem. Whereas wetlands are important source of freshwater, these have also got in application in nutrient/pollutant removal and improving the water quality. Several studies have reported the role of wetlands in removal of suspended solids (Tanner et al. 1995), nutrients (Haritash

✉ A. K. Haritash
akharitash@dce.ac.in

[1] Department of Environmental Engineering, Delhi Technological University, Bawana Road, Shahbad Daulatpur, Delhi 110 042, India

Published online: 02 January 2021

et al. 2015, 2017), organic impurities (Fisher and Acreman 2004; Mohanty et al. 2014), pathogens (Rogers 1983), and even metals (Prasanna et al. 2012; Zhang et al. 2012). Realizing the ecological and environmental importance of wetlands, various conservation plans are going on in different states to conserve wetlands. India has 26 wetlands (Ramsar Sites) of international importance besides a number of other wetlands of national importance. Since the Northern states of India (J&K, Himachal Pradesh, Rajasthan, and Haryana) lie enroute of migratory birds, the wetlands of these states have gained significant environmental importance as nesting and breeding grounds for migratory birds. Haryana is an important North Indian State with two important wetlands (Sultanpur and Bhindawas) which offer temporary habitat to migratory birds. Whereas Sultanpur wetland goes dry during summer season, Bhindawas is a man-made wetland with Jawahar Lal Nehru (JLN) Canal as source of water to maintain its water level. This man-made wetland is surrounded by the agricultural fields. During monsoon, water from the water-logged fields is pumped into the wetland which adds nutrients like phosphate and nitrate which and acts as a stimulus to eutrophication (Kaur and Singh 2012; Yadav et al. 2015; Haritash et al. 2015). Bhindawas and Sultanpur Bird Sanctuaries have also been identified under National wetland conservation programme as these water bodies have several species of animals, birds, and plants. Bhindawas wetland is also identified as an important wetland under International Wetland Conservation Programme of Ministry of Environment, Forests and Climate Change (MOEFCC), Govt. of India, because of its ecological importance**.** Over the past years, due to declining water level in the Bharatpur Bird Sanctuary in Rajasthan, the migratory water birds have started nesting in Bhindawas wetland during winter season (Haritash et al. 2015). Some studies have been carried out for physiochemical characterization of water and its eutrophication potential (Kumar and Dhankhar 2012; Saluja and Garg 2014), but studies dedicated to evaluation, overall quality and contamination due to heavy metals are very limited. Poor water quality has become an important issue, which ultimately affects ecosystem health and functioning and underpins biodiversity and is also responsible for major alterations in water use (Liang et al. 2016). In view of its ecological significance, it becomes imperative to monitor the water quality of Bhindawas wetland regularly. Since the water quality is an aggregated outcome of number of parameters, single numerical value called water quality index (WQI) has been suggested to represent it in simpler terms. It is evaluated by assigning weightage and rating to individual parameters against their standard prescribed values. Similarly, heavy metal evaluation index (HEI), heavy metal pollution index (HPI) and degree of contamination (C_d) can also be used to represent quality with respect to toxic heavy metals. Since the water of Bhindawas supports aquatic life and irrigation (in downstream stretch), and it receives the wastewater from adjoining villages, agricultural fields, and overflow from drain no. 8; addition of excess nutrients, organic impurities, and pollutants like heavy metals are significantly affecting the quality of water subsequently causing an effect of aquatic flora and fauna, agricultural crops and soil health. The present study was devised to carry a comprehensive assessment of water quality and metal toxicity in Bhindawas wetland.

Materials and methods

Study area

The present study was undertaken in Bhindawas wetland, which is located in Jhajjhar district of Haryana state. The wetland is located in a semiarid region with average temperature of 14 °C in winter and 47 °C during summer with annual average rainfall of 800 mm. The wetland was a protected forest and was declared as a Bird Sanctuary in the year 1986 by Ministry of Environment and Forests. Bhindawas wetland is located at 28°32′ North latitude and 76°32′ East longitude in Jhajjhar district of Haryana state. The wetland spreads over an area of 1074 Acres and gets water from monsoonal rains and excess spilling of JLN Canal. Almost 50% of the area of wetlands provide permanent water surface to plants like *Eichhornia, Valliseneria, Salvinia* etc. The wetland has sufficient number of trees like *Acacia, Eucalyptus, Albizia* etc. which provide surface for nesting within the wetland. It is regarded as Kaeoladeo National Park of Haryana (Gupta et al. 2011). This saucer shaper marshy area receives water from Jawahar Lal Nehru Canal during power failure and release extra water in Drain no.8. It is the largest wetland in all over Haryana with a periphery of 12 km and average water depth 6 feet. Due to water crisis in Kaeoladeo National Park (Bharatpur) last few years, Bhindawas wetland has gained increasing interest by migratory birds. It is being used by the migratory and resident birds of about 250 species. This sanctuary is situated in the centre of a number of villages namely Kanwah, Bilochpura, Nawada, Redhuwas, Shahjahanpur and Chadwana.

Collection of samples and analysis

For collection of samples of surface water, the wetland was divided into grids (20 in number) and one representative sample was collected from each grid (Fig. 1). The sampling was performed during the day time (11 a.m. to 2 p.m.) in during monsoon season (September, 2015). One sample each was collected JLN Canal (inlet of water) and drain no. 8 (outlet) to arrive at difference of different parameters within the wetland. Some of the parameters like pH, EC,

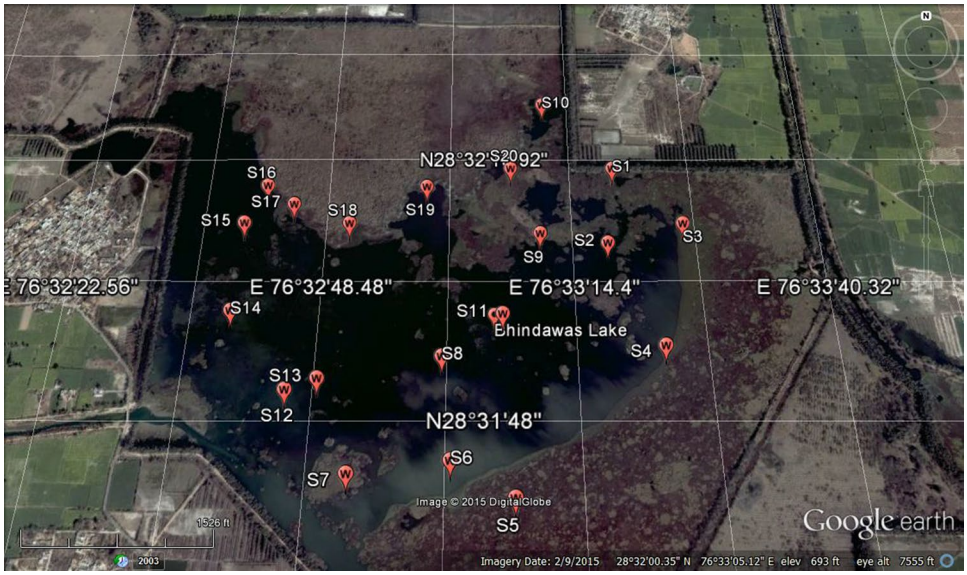


Fig. 1 Location of sampling sites in Bhindawas wetland

TDS, temperature and DO were analysed onsite using Orion (USA) make Star 320A + model. The samples were collected in pre-rinsed polypropylene bottles (1L), stored in an ice box and were transported to the laboratory immediately after collection. Total suspended solids (TSSs) were measured using gravimetric analysis; chlorides (Cl^{-}), total alkalinity (TA), and total hardness (TH) were determined by volumetric analysis; sulphate, nitrate, phosphate and dissolved ammonia were measured spectrophotometrically (Labtronics make LT-290 model); Na^{+}, K^{+}, Mg^{2+} and Ca^{2+} were measured flame photometrically (Systronics make 128μ model); and heavy metals (Cu, Fe, Ni, Pb, Zn) were measured using atomic absorption spectrophotometer (Analytik Jena Nova 350). All the experiments were conducted in triplicates using AR Grade chemicals and type-1 ultrapure water; and as per the standard methods prescribed by APHA (2005). The calculation of WQI for aquatic life and irrigation; and HPI and HEI for metal contamination was also undertaken. Statistical analysis of data was performed on SPSS version 20 to establish spatial similarities using cluster analysis; and major parameters affecting water quality by principal component analysis (PCA).

Water quality index

The water quality index (WQI) is a useful technique of rating that determines the collective influence of individual water quality parameter on the overall quality of water (Al-Mohammed and Mutasher 2013; Singh et al. 2015). WQI of Bhindawas Bird Sanctuary was calculated using the weighted arithmetic index method Brown et al (1972). For calculating the WQI in the present study, total 21 parameters (namely temperature, pH, total dissolved solids, total suspended solids, dissolved oxygen, chemical oxygen demand, chlorides, nitrate nitrogen, ammonium nitrogen, phosphate, sulphate, Ca^{+2}, Mg^{+2}, Na^{+}, K^{+}, Cu^{+2}, Fe^{+2}, Ni^{+2}, Pb^{+2}, Zn^{+2}) were considered (Table 1). The quality rating (Qi) for each parameter (Qi) was calculated by using this expression: $\text{Quality rating(Qi)} = \left[(\text{Cn} - \text{Ci})/(\text{Cs} - \text{Ci})\right] * 100$ where Cn: actual value of nth parameter; Ci: ideal value of this parameter; Ci = 0 (except Ci-7 for pH; and Ci-14.6 mg/l for DO); and Cs*:* recommended standard of corresponding parameter. Relative weight (Wi) was calculated by a value inversely proportional to the recommended standard (Si) of the corresponding parameter.

$$\text{Wi} = 1/\text{Si}$$

Overall WQI was calculated by using equation: $\text{WQI} = \left(\sum i = 1 - n\text{QiWi}\right)/\left(\sum i = 1 - n\text{Wi}\right)$ where Qi is the sub quality index of *i*th parameter; W is the unit weight of each parameter; and *n* is the number of parameters.

The classification of water quality based on WQI was done as per the criteria given below (Chatterji and Raziuddin 2002).

WQI value	Rating of water quality	Grading
0–25	Excellent	A
26–50	Good	B
51–75	Poor	C
76–100	Very poor	D
Above 100	Unsuitable for drinking	E

Calculation of metal contamination

Indexing concept helps in determining the status of water quality with respect to selected parameters. To understand

Table 1 Average concentration of physico-chemical characteristics of water samples collected from Bhindawas wetland

Parameter	Mean ± SD	Range (Min.–Max.)	Prescribed standard (Si)	
			Aquatic life[a]	Irrigation[b]
pH	8.0 ± 0.62	7.1–9.6	8	8.5
Temp	33.0 ± 0.47	32.2–33.9	20	–
EC	472.5 ± 26.15	403–508	–	3000
TDS	234.3 ± 9.62	203–250	500	2000
TSS	464.5 ± 111.60	290–684	+ 25	–
TA	100 ± 24	0–117	20	–
TH	166.40 ± 24.15	108–224	–	–
BOD	5.16 ± 1.55	3.14–9.46	–	–
DO	8.1 ± 2.08	5.3–13.3	5.5	–
COD	23.2 ± 0.45	23–24	7	–
DOC	17.48 ± 5.55	8.68–38.76	–	
Cl^-	58.2 ± 7.62	52–88	120	1063
SO_4^{2-}	85.3 ± 50.99	27–240	–	960
Ca^{2+}	18.8 ± 1.06	15–20	–	400
Mg^{2+}	17.3 ± 6.03	7–28	–	60
Na^+	31.3 ± 1.30	30–34	–	919
K^+	2.4 ± 0.38	1.1–3.0	–	2
NO_3^-N	14.2 ± 11.43	0–44.1	2.93	10
NH_3^+N	0.5 ± 0.26	0.2–1.4	1.37	5
TKN	4.20 ± 4.56	1.4–22.4		
PO_4^{3-}	4.9 ± 2.98	0.4.2–10.6	–	2
Cu^{2+}	0.08 ± 0.04	0–0.11	0.004	0.2
Fe^{2+}	0.7 ± 0.37	0.42–1.90	0.3	5
Ni^{2+}	0.2 ± 0.04	0.17–0.32	0.025	0.2
Pb^{2+}	1 ± 0.71	0.21–3.5	0.007	5
Zn^{2+}	0.01 ± 0.04	0–0.16	0.050	2

*Except pH, temperature (°C), and EC (μsiemens/cm) all parameters have units in mg/l

[a]CCME-Canadian Council of Ministers of the Environment; [b]FAO-Food and Agriculture Organization

the contamination level of the lake with respect to heavy metal, two different quality indices were used to calculate the metal pollution of Bhindawas wetland.

Heavy metal pollution index (HPI)

HPI index helps in determining the water quality status of surface water with respect to various selected parameters. For calculating HPI, each chosen parameter is assigned a rating or weightage (Wi) which is an arbitrary value, lies between 0 and 1 and defined as inversely proportional to standard permissible values (Si) for each selected parameter. HPI is calculated by using following equation (Mohan et al. 1996):

$$\mathrm{HPI} = \left(\sum i = 1 - \mathrm{n}\ \mathrm{QiWi}\right) \Big/ \left(\sum i = 1 - \mathrm{nWi}\right)$$

where Wi and Qi are the unit weight of *i*th parameter and sub index, respectively; and *n* is the number of parameters considered. The value of sub index (Qi) is calculated by using equation:

$$\mathrm{Qi} = \sum i = 1 - n\left[(\mathrm{Ci} - \mathrm{Ii})/(\mathrm{Si} - \mathrm{Ii})\right] * 100$$

where Ci, Si, and Ii are the observed value of heavy metal, standard value of *i*th parameter, and ideal values, respectively. And the (−) sign denotes the numerical difference between two values, ignoring the algebraic sign.

Heavy metal evaluation index (HEI)

In metal index, the relative contamination of different metals was calculated and a single representative number is identified which is calculated by dividing sum of all the concentrations of heavy metals by sum of maximum allowable concentrations of each heavy metal studied. If the concentration of a metal is higher as compared to its respective MAC value, it represents the worse quality of the water. MI value > 1 is a threshold of warning (Bakan et al. 2010). The MI is calculated by using the following formula (Edet and Offiong 2002):

$$\mathrm{HEI} = \sum i = 1 - n(\mathrm{Ci}/\mathrm{MACi})$$

where Ci is the observed concentration of each element, and MAC is the maximum allowable concentration.

Degree of contamination (C_d)

Combined effect of various water quality parameters is summarized in a single index called contamination index (Backman et al. 1997) and is expressed as:

$$\mathrm{Cd} = \sum i = 1 - n\ \mathrm{Cfi}$$

where$\mathrm{Cfi} = (\mathrm{Coi}/\mathrm{Cni}) - 1$. where Cfi, Coi, and Cni are contamination factor, observed value, and normative value, respectively. Here, normative value is taken same as MAC value.

Results and discussion

The physicochemical characteristics of collected samples of water revealed that almost all the cations, anions, and other organic pollutants were present to detectable levels (Table 1). Another important observation was the significant differences in levels of dissolved oxygen, phosphate, sulphate, nitrate, dissolved oxygen, and organic carbon in inlet and outlet, as well as spatially within the wetland.

General observations

The average temperature of water within the wetland was 33 °C with negligible/insignificant spatial variations, but it was significantly higher than the temperature at inlet and outlet. The reason for the relatively low temperature of inlet and outlet may be ascribed to more depth and turbulence, respectively. This could be affirmed by relatively lower value of dissolved oxygen (0.7 mg/l) in inlet (owing to more depth); and higher value (6.6 mg/l) at outlet. Rest of the parameters like TDS, pH, TSS and chloride had minimum spatial variations with mean value of 243 mg/l, 8.0, 465 mg/l and 58 mg/l respectively. The levels of dissolved oxygen were observed to be higher at location S11 (12.3 mg/l) and S15 (13.3 mg/l) which may be ascribed to lower depth at the locations, which favoured rich growth submerged vegetation particularly *Hydrilla* and *Vallisneria* at these locations. Higher rate of subsurface photosynthetic activity at such locations may raise levels of dissolved oxygen. Similar observations have been reported in other studies too (Haritash et al. 2015). Average concentration of organic impurities in terms of BOD_5 and COD was 4 mg/l and 23 mg/l, respectively. Major source of organic impurities in the wetland in autochthonous originating from death and decomposition of plant and animal biomass. Relatively high values of BOD at location S19 and S20 may be attributed to accumulation of dead plant biomass of *Eichhornia, Salvinia* etc. along the northern shallow region of wetland. The level of hardness and alkalinity, calcium, magnesium, sodium, and potassium were almost uniform throughout the wetland. Lower level of calcium at location S11 may be ascribed to its precipitation as $CaCO_3$ at pH $>$ 9.0. It may also result in precipitation of phosphate subsequent upon its binding with calcium. Lower concentration of phosphate at S11 location (0.42 mg/l) is a confirmation of co-precipitation of Ca^{2+} and PO_4^{3-}. Similar relationship of phosphate was observed with pH at other locations too. Very high concentration at locations S10 (10.6 mg/l) and S18 (10.2 mg/l) is contributed by more bird droppings at these locations. These locations are near to the shore with sufficient number of trees serving as nesting grounds for the aquatic birds. Disposal of PO_4^{3-} rich droppings at these locations resulted in higher concentration, and relatively low pH at these locations keeps the phosphate in water soluble state. Significant difference in concentration of phosphate in inlet (1.1 mg/l) and outlet (7.8 mg/l) indicates that phosphate has accumulated in the water body over the period of time and most of it circulates within the wetland ecosystem. The other non-quantified source of import of phosphate to the wetland is pumping of water from adjoining water-logged fields during monsoon and post monsoon. Similarly, the concentration of nitrate in inlet (9 mg/l) and outlet (19.7 mg/l) had significant difference, which may also be ascribed to import from pumped of water from agricultural fields. Spatial variations in level of nitrate were observed within the wetland with values varying from 3.8 to 44.1 mg/l. Higher values at certain locations may be an outcome of higher oxidising potential and improved microbial activity. Dissolved ammonia, on the other hand, was very low in concentration with an average value of 0.5 mg/l. Its concentration was high in inlet (1.1 mg/l) and less in the wetland owing to relatively stronger oxidising conditions in the wetland. Similarly, Total Kjeldahl's nitrogen (TKN) was moderate in concentration with an average value of 3.4 mg/l.

Apart from the general parameters, heavy metals (Fe, Zn, Ni, Cu, Pb) were also reported in water of Bhindawas wetland. The average value of iron, zinc and nickel was found to be 0.7 mg/l, 0.01 mg/l and 0.2 mg/l, respectively, which may be a cause of concern for the aquatic life present in the wetland since these metals are bio accumulative in nature. Similarly, high mean concentration of copper (0.03 mg/l) and lead (0.96 mg/l) that the metals are gradually accumulating in the wetland compared to the level in inlet confirms that these metals are gradually accumulating in the wetland (Table 2).

Exceedance of levels and its effect depends on degree and frequency of exceedance, chemical nature of the pollutant, its toxicity, and its interaction with other pollutants. Comparison of individual parameters against the prescribed standards will not provide sufficient information to arrive at aggregated quality. Water quality index is an aggregated numerical value which represents the quality of water based on values of given set of parameters. In the present study, evaluation of WQI has been done to determine the water quality for survival of aquatic life and use in irrigation. Based on the values of WQI for agriculture use, the water was classified as very poor or unsuitable for use upon comparison with standards prescribed by FAO (1994). The major parameters, which exceeded the prescribed standards, were nitrate, phosphate, and potassium at most of the location, which clearly indicates the input from NPK fertilisers being employed in fields. The water being pumped into the wetland from adjoining water-logged fields during monsoon is, therefore, the dominant source and reason for compromised quality of water. Based on WQI for survival of aquatic life, the water was found to be unsuitable upon comparison with standards prescribed by CCME (2007).

Heavy metal pollution index (HPI)

HPI is based on the weight (Wi) of the metals analysed in a study. The weight or rating varies between 0 and 1 and it indicates the relative significance of an individual metal and it is evaluated by inverse proportion of the standard/normative value (Si) of that metal. The normative value used for calculated HPI values for irrigation and aquatic use

Table 2 Classification of water in Bhindawas wetland based on HPI, HEI, degree of contamination and WQI

Index	Aquatic life				Irrigation			
	Modified category	Degree of pollution	No. of samples (%)	Mean (Range) (Min.–Max.)	Modified category	Degree of pollution	No. of samples (%)	Mean (Range) (Min.–Max.)
HPI	< 50	Low	13 (65)	48 (25–161)	< 1	Low	19 (95)	0.7 (0.5–1.3)
	50–100	Medium	06 (30)		1–2	Medium	1 (5)	
	> 100	High	01 (05)		> 2	High	–	
HEI	< 160	Low	13 (65)	160 (68–519)	< 2	Low	17 (85)	1.7 (1.1–2.2)
	160–320	Medium	06 (30)		2–4	Medium	03 (15)	
	> 320	High	01 (05)		> 4	High	–	
C_d	< 150	Low	13 (65)	152 (63–514)	< − 3	Low	17 (85)	− 3.3 (− 3.9 to − 2.8)
	150–320	Medium	06 (30)		− 3–0	Medium		
	> 320	High	01 (05)		> 0.0	High	03 (15)	
WQI	0–25	Excellent	Nil	194 (101–384)	0–25	Excellent	Nil	143 (53–257)
	25–50	Good	Nil		26–50	Good	Nil	
	51–75	Poor	Nil		51–75	Poor	2 (10)	
	76–100	Very Poor	Nil		76–100	V. Poor	4 (20)	
	> 100	Unsuitable	20 (100)		> 100	Unsuitable	14 (70)	

have been extracted from FAO (1994) and CCME (2007), respectively. In the present study, the degree of pollution was observed to be low for 65%; medium for 35% and high for only 5% of samples for aquatic life. The observed values of HPI for aquatic life varied from 25 to 160. Relatively higher values were observed at locations S12 to S18 adjoining Chadwana village in northeast region of the wetland. Higher levels of heavy metals in that particular region may be ascribed to pumping of domestic sewage from the adjoining village. This indicates that the water quality is questionable for aquatic life. It becomes a cause of concern since Bhindawas wetland is a temporary nesting ground for the migratory birds during the winter season. On the other hand, degree of pollution was low for use in irrigation. The values varied between 0.4 and 0.9 indicating that the heavy metals were within the safe limits for use in agriculture. Since the heavy metal uptake by plants is selective and restrictive, relatively smaller fractions are assimilated in the plant tissue. Therefore, the standard limits for use in irrigation are relaxed/higher which results in low values of HPI for agricultural use. The classification of degree of pollution for aquatic and agricultural use is given in Table 3.

Heavy metal evaluation index (HEI)

HEI is an indication of aggregated quality of water with respect to the analysed heavy metals. It is used for a clear understanding and validation of other pollution indices by synchronising the classification criteria of the other indices. Based on the values of HEI, 65% of the samples add low degree of pollution; 30% had medium; and 5% of the sample had the high degree of the pollution as observed in case of HPI. The only sample, which had maximum HEI value, was collected from S12 location adjoining Chaddwana village. Unlike HPI 85% of the samples add low degree of pollution and 15% add medium degree of pollution for use in irrigation. The locations S12 and S15 were observed to have relatively high metal concentration. The classification of water based on HEI is given in Table 3.

Degree of contamination (C_d)

The degree of contamination/contamination index (C_d) is another representation of the quality of water with respect to the heavy metal concentration. Based on the values of C_d, 65% samples had low; 30% medium; and 5% samples having high degree of pollution confirming to the classification as observed for HPI and HEI of water for aquatic life. Similarly, for use in irrigation, 85% samples had low and 15% had medium degree of pollution, which was in conformance with the classification observed for the HEI. The details of C_d are given in Table 3. Based on various indices (HPI, HEI, C_d), for degree of pollution most of the locations were having good quality of water for irrigation and aquatic life with the only exception of region adjoining rural habitation in north east region of the wetland. For aquatic life, all three indices had similar classification for quality and validated each other. For agricultural quality, C_d and HEI had similar values and were 85% in conformance with HPI.

Table 3 Principal component analysis for selected parameters of water in Bhindawas wetland

Variables	Components							
	1	2	3	4	5	6	7	8
pH	− 0.86	0.35	− 0.11	0.29	− 0.02	− 0.06	0.02	− 0.05
Temp	− 0.17	**0.58**	0.03	0.32	0.22	0.13	− 0.34	− 0.28
EC	**0.80**	0.24	0.20	0.21	0.23	− 0.28	0.06	− 0.04
TSS	**0.50**	**0.65**	− 0.32	− 0.17	− 0.15	0.07	0.12	0.00
DO	− 0.82	0.17	− 0.13	0.35	0.17	0.17	− 0.07	− 0.06
BOD	− 0.73	0.34	0.14	0.18	0.15	0.16	0.27	− 0.16
Cl^-	− 0.17	**0.50**	0.05	− 0.23	− 0.28	− 0.28	0.08	0.19
SO_4^{2-}	0.12	− 0.01	− 0.09	0.39	0.36	0.01	− 0.62	0.46
Ca^{2+}	**0.74**	0.20	0.33	0.06	0.45	0.03	0.05	0.15
Mg^{2+}	0.31	− 0.36	− 0.37	**0.67**	0.04	0.10	0.15	− 0.11
Na^+	0.24	**0.75**	0.25	0.38	− 0.06	0.01	0.11	− 0.20
K^+	**0.80**	0.09	0.45	0.03	0.26	0.03	0.21	− 0.02
NO_3^-	− 0.20	− 0.68	0.23	0.06	− 0.13	− 0.27	− 0.21	− 0.32
NH_3	0.38	0.17	− 0.01	− 0.27	− 0.39	**0.55**	− 0.37	− 0.05
TKN	0.28	− 0.14	− 0.57	0.12	0.08	0.42	0.36	− 0.09
PO_4^{3-}	0.18	− 0.56	0.40	− 0.14	0.23	0.42	− 0.02	− 0.30
Cu	0.43	− 0.29	− 0.69	0.02	0.11	0.05	− 0.01	0.13
Fe	0.08	− 0.56	0.29	**0.56**	− 0.39	− 0.15	0.10	0.05
Ni	− 0.58	− 0.23	− 0.03	− 0.15	0.46	− 0.05	0.31	0.35
Pb	− 0.58	− 0.08	0.46	− 0.15	0.16	0.43	0.06	0.23
Zn	0.12	0.03	0.32	0.41	− 0.56	0.31	0.15	0.46
Eigen values	5.40	3.37	2.14	1.85	1.60	1.28	1.12	1.04
CV%	25.73	41.77	51.93	**60.76**	68.36	74.45	79.79	84.74

Vaues in bold represent $p < 0.05$

Statistical analysis

The statistical analysis of the obtained data was performed to determine significantly correlated parameters which affect the different biochemical processes in a wetland. The correlation coefficients of selected parameters were determined using the data of all 20 locations. Significant positive correlation was observed for DO and BOD, Ca and K, Mg, Cu, and Fe, PO_4^{3-} with NO_3^-, and Pb, NO_3^- and Fe. PCA was carried out to compare the compositional pattern between the physicochemical parameters along with heavy metal concentration of water samples from various sites of the wetland and to identify the factors influencing each one. PCA for the total data set is shown in Table 3, which explained eight components with eigen values > 1 having about 84.74% of total variance in the water quality. The first component accounting for 25.73% of the total variance and was correlated (loading ≥ 0.50) with EC, TSS, Ca^{2+} and K^+. Cluster analysis is applied to the water quality parameters (both physiochemical and heavy metal concentration) of various sites of the lake. A dendrogram was obtained where all the 20 sampling sites of the lake were grouped into three statistically significant clusters. Cluster 1 comprises of S1 to S9 and S12 to S20; cluster 2 comprises of only S10; and lastly cluster 3 comprises of S11, respectively (Fig. 2).

Conclusion

The physicochemical characterisation revealed that the water quality of Bhindawas was compromised w.r.t. the levels of heavy metals. The general parameters (cations and anions) had minimum spatial variations and were within the safe limits level for aquatic life. The growth of submerged vegetation facilitated improved water quality, adding photosynthetically produced oxygen to water and filtering the suspended impurities. But, the addition of wastewater from adjoining villages and runoff from nearby fields has resulted in increased concentration of phosphates and heavy metals. Since these chemical species cause eutrophication and impart toxicity, respectively, the health of ecosystem and aquatic life is affected adversely. The results of this study are even more important when migratory birds use this wetland as a temporary refuge during winters. The present quality of water is unsuitable for aquatic life based on various indices of heavy metal pollution (HPI, HEI, C_d). The wetland

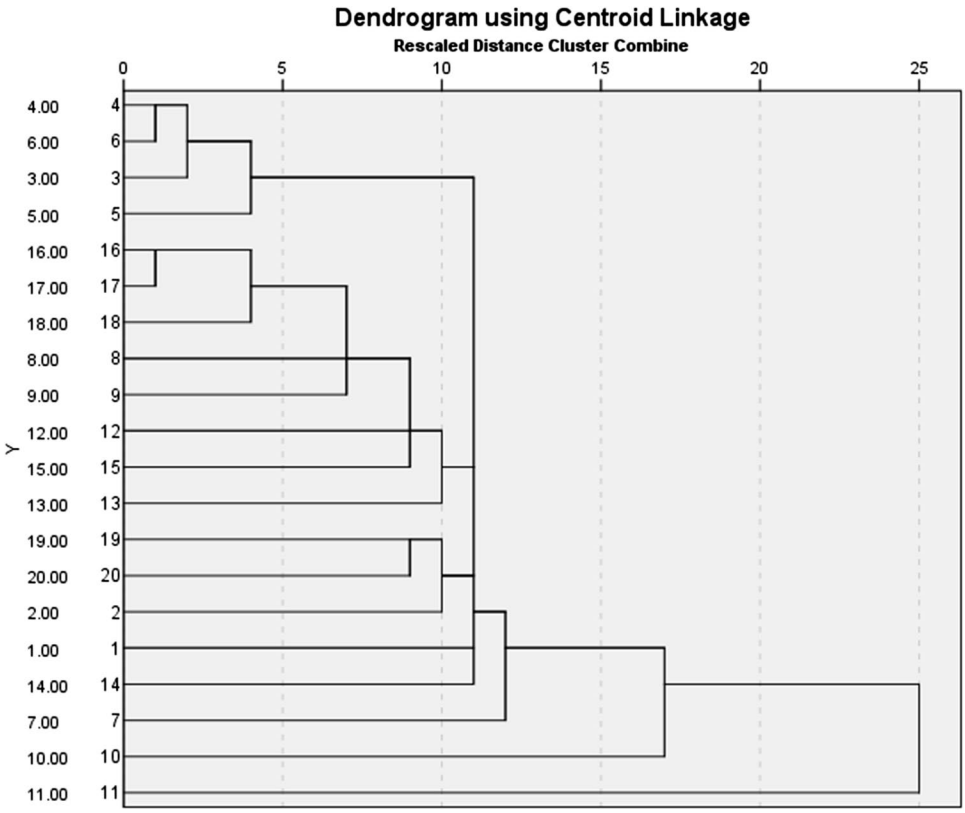


Fig. 2 Dendrogram for cluster analysis of sampling locations in Bhindawas wetland

requires regular monitoring of water quality; and treatment of wastewater before being pumped into it during monsoon.

Acknowledgements The authors acknowledge help of the Chief Wildlife Warden, Haryana for the permission and help during collection of samples.

Funding The authors received no specific funding for this work.

Compliance with ethical standards

Compliance with ethics requirements manuscript includes no issue related to ethical compliance.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Al-Mohammed FM, Mutasher AA (2013) Application of water quality index for evaluation of groundwater quality for drinking purpose in Dibdiba Aquifer, Kerbala City, Iraq. J Babylon Univ Eng Sci 21(5):1647–1660

APHA: Standard Methods for the Examination of Water and Wastewater (2005) 21st edn. APHA, AWWA, WPCF Washington DC, USA

Backman B, Bodis D, Lahermo P, Rapant S, Tarvainen T (1997) Application of a groundwater contamination index in Finland and Slovakia. Environ Geol 36:55–64

Bakan G, Ozkoc HB, Tulek S, Cuce H (2010) Integrated environmental quality assessment of Kızılırmak River and its coastal environment. Turk J Fish Aquat Sci 10:453–462

Brown RM, McCleiland NJ, Deininger RA, O'Connor MF (1972) A water quality index—crossing the psychological barrier. In: Jenkis SH (ed) Proceedings in international conference on water pollution research, Jerusalem, vol 6, pp 787–797

CCME (Canadian Council of Ministers of the Environment) (2007) For the protection of aquatic life 2007. In: Canadian environmental quality guidelines, 1999, Canadian Council of Ministers of the Environment, 1999, Winnipeg

Edet AE, Offiong OE (2002) Evaluation of water quality pollution indices for heavy metal contamination monitoring. A case study from Akpabuyo-Odukpani area, Lower Cross River Basin (southeastern Nigeria). GeoJournal 57:295–304

Fisher J, Acreman MC (2004) Wetland nutrient removal: a review of the evidence. Hydrol Earth Syst Sci 8(4):673–685

Giri S, Singh AK (2014) Assessment of surface water quality using heavy metal pollution index in Subaranarekha River, India. Water Qual Expo Health 5:173–182

Gupta RC, Parasher M, Kaushik TK (2011) An enquiry into the avian biodiversity of Bhindawas Bird Sanctuary in Jhajjar District in Haryana State in India. J Exp Zool 14(2):457–465

Haritash AK, Sharma A, Bahel K (2015) The potential of Canna lily for wastewater treatment under Indian conditions. Int J Phytoremediation 17(103):999–1004

Haritash AK, Dutta S, Sharma A (2017) Phosphate uptake and translocation in a tropical Canna based constructed wetland. Ecol Process 6(12):1–7

Kaur S, Singh I (2012) Accelerated phosphate and nitrate level: factors to blame for Eutrophication in Yamuna River, Delhi, India. Int J Plant Anim Environ Sci 2:183–187

Kumar S, Dhankhar R (2012) Trophic State Index and assessment of water quality for domestic and agriculture purpose of Bhindawas wetland, Jhajjhar, Haryana (India). Ann Biol 28:144–151

Liang J, Liu J, Yuan X, Zeng G, Yuan Y, Wu H, Li F (2016) A method for heavy metal risk exposure risk assessment to migratory herbivorous birds and identification of priority pollutants/areas in wetlands. Environ Sci Pollut Res 23:11806–11813

Mohan SV, Nithila P, Reddy SJ (1996) Estimation of heavy metal in drinking water and development of heavy metal pollution index. J Environ Sci Health A31:283–289

Mohanty A, Ray S, Yadav AK, Chaudhury GR (2014) Optimization study: biological removal of inorganic nitrogen along with chemical oxygen demand from wastewater using response surface methodology. J Clean Soil Air Water 42(11):1583–1592

Prasanna MV, Praveena SM, Chidambaram S, Nagarajan R, Elayaraja A (2012) Evaluation of water quality pollution indices for heavy metal contamination monitoring: a case study from Curtin Lake, Miri City, East Malaysia. Environ Earth Sci 67:1987–2001

Rogers FEJ (1983) Wetlands as accreting systems: waste water treatment. J Limnol Soc South Afr 9(2):110–116

Saluja R, Garg JK (2014) Surface water quality assessment of Bhindawas Lake (Haryana, India) using multivariate statistical techniques. J Glob Ecol Environ 2:2454–2644

Senthil SL, Ajith Kumar TT, MarudhuPandi T, Dhaneesh KV, Murugan JB, BalaSubramanian T (2012) Metal contagion in ecologically important estuary located in Bay of Bengal. Water Qual Expo Health 4(3):137–142

Sharma M, Kansal A, Jain S, Sharma P (2015) Application of multivariate statistical techniques in determining the spatial temporal water quality variation of Ganga and Yamuna rivers present in Uttarakhand State, India. Water Qual Expo Health 7:567–581

Shelef O, Golan-Goldhirsh A, Gendler T, Rachmilevitch S (2011) Physiological parameters of plants as indicators of water quality in a constructed wetland. Environ Sci Pollut Res 18:1234–1242

Tanner CC, Clayten JS, Upsdell MP (1995) Effect of loading rate and planting on treatment of dairy farm wastewaters in constructed wetlands—I. Removal of oxygen demand, suspended solids and faecal coliforms. Water Res 29(1):17–26

Zhang W, Liu X, Cheng H, Zeng EY, Hu Y (2012) Heavy metal pollution in sediments of a typical mariculture zone in South China. Mar Pollut Bull 64(4):712–720

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

© The Author(s) 2021. This work is published under https://creativecommons.org/licenses/by/4.0/(the “License”). Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.

arXiv:2101.01899v1 [cs.HC] 6 Jan 2021

Exploring Semi-Supervised Learning for Predicting Listener Backchannels

Vidit Jain*
vidit17370@iiitd.ac.in
IIIT-Delhi
New Delhi, India

Maitree Leekha*
maitreeleekha_bt2k16@dtu.ac.in
Delhi Technological University
New Delhi, India

Rajiv Ratn Shah†‡
rajivratn@iiitd.ac.in
IIIT-Delhi
New Delhi, India

Jainendra Shukla†
jainendra@iiitd.ac.in
IIIT-Delhi
New Delhi, India

ABSTRACT

Developing human-like conversational agents is a prime area in HCI research and subsumes many tasks. Predicting listener backchannels is one such actively-researched task. While many studies have used different approaches for backchannel prediction, they all have depended on manual annotations for a large dataset. This is a bottleneck impacting the scalability of development. To this end, we propose using semi-supervised techniques to automate the process of *identifying* backchannels, thereby easing the annotation process. To analyze our identification module's feasibility, we compared the backchannel *prediction* models trained on (a) manually-annotated and (b) semi-supervised labels. Quantitative analysis revealed that the proposed semi-supervised approach could attain 95% of the former's performance. Our user-study findings revealed that almost 60% of the participants found the backchannel responses predicted by the proposed model more natural. Finally, we also analyzed the impact of personality on the type of backchannel signals and validated our findings in the user-study.

CCS CONCEPTS

• **Human-centered computing** → *User studies*; **Empirical studies in HCI**; • **Computing methodologies** → **Semi-supervised learning settings**.

KEYWORDS

Conversational Agents, Backchanneling, Multimodal analysis.

*The authors contributed equally, and wish they be regarded as joint first authors.

†Jainendra Shukla and Rajiv Ratn Shah are partly supported by the Infosys Center for AI and Center for Design and New Media at IIIT Delhi.

‡Rajiv Ratn Shah is also partly supported by the ECRA Grant (ECR/2018/002776) by SERB, Government of India.

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 ACM 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.
CHI '21, May 8–13, 2021, Yokohama, Japan
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445449

ACM Reference Format:
Vidit Jain, Maitree Leekha, Rajiv Ratn Shah, and Jainendra Shukla. 2021. Exploring Semi-Supervised Learning for Predicting Listener Backchannels. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3411764.3445449

1 INTRODUCTION & RELATION TO PRIOR WORK

Human conversations, even the most casual ones, have a lot of complexity associated with them. Two people actively engaged in a conversation frequently respond to each other, not only with respect to the content of the conversation, but also to behavioral aspects, such as facial expressions and prosody. Developing Embodied Conversational Agents (ECAs) [9] and spoken dialogue systems capable of incorporating these complex elements to converse naturally is a challenging task and has been a constant focus of the Artificial Intelligence and Human Computer Interaction research communities. Of these complex conversational constructs, dyadic components like listener backchannels are among the most crucial for modeling virtual humans and are also the main focus of this study[1].

In a peer-to-peer conversation, a backchannel occurs when one of the participants is speaking, and the other (the listener) interjects a short response to the former [41]. These responses do not interrupt the flow of the conversation; rather, they convey the listener's state-of-mind about the speaker's dialog. They also reflect cooperation and understanding between the two parties [17]. Backchannels can be verbal, non-verbal (visual) or both. Vocalisations like 'hmm' or 'uh-huh', gestures such as head nods or head shakes, and a combination of verbal and non-verbal responses are common examples of backchannels.

In the past few years, the research community has shown a keen interest in modeling the listener's backchanneling behavior. A large number of such studies on backchannel prediction have focused on the use of rule-based classifiers. Ward [38], Truong *et al.* [36], Ward and Tsukahara [39] utilized different acoustic features of the speaker such as pitch, pausal information, etc. to predict backchannel opportunities. A recent study by Park *et al.* [29] on backchannel

[1]A teaser video demonstrating our work can be found in the video figure section of the submission. It illustrates a virtual listener that emits backchannels to the speaker's context using the models proposed with this work.

prediction for children also used similar prosodic features and hand-crafted rules. Expanding the feature set, Moubayed *et al.* [1] used both visual and prosodic speaker features.

Transitioning towards data-driven automatic prediction from hand-crafted rules, Solorio *et al.* [34] used prosodic features with locally weighted linear regression to predict backchannel opportunities. Morency *et al.* [27] used multimodal (visual and acoustic) speaker features and a hidden markov model to predict backchannels. More recently, many researchers also used deep learning techniques for predicting backchannels. Ruede *et al.* [32] used Long short-term memory (LSTM) based model with acoustic features. Hara *et al.* [16] also used LSTMs to predict turn-taking, filler words, and backchannels in a multitask learning paradigm. In a recent study, Goswami *et al.* [15] used state of the art machine learning and deep learning-based time-series classification techniques to model backchannel opportunities in children with multimodal features. Inspired by prior work, we too explore the use of machine learning and deep learning based time series models and include several similar multimodal features in our analysis.

All these studies, although well-founded, have the following limitations:

(1) Most of them have focused only on the backchannel opportunity prediction task, with a very few involving the next step of the problem *i.e.,* predicting the type of backchanneling signals.
(2) All studies in the literature have relied on data with annotated backchannel instances during their modeling phase, where they develop a backchannel predictor using the speaker context. For instance, the datasets used by prior works, including the Switchboard Dialog Act Corpus (SwDA) [23], Iraqi Arabic [40], P2PStory [33], data collected by Morency *et al.* [27], and many others, had to be manually annotated for listener backchannels by multiple coders. This annotation process can be extremely time-consuming, depending on the amount of data present. Furthermore, this approach also does not scale well when trying to expand the scope of development by collecting more data. For instance, in many studies, including the ones cited above, the datasets used cover a relatively small population size. The size of the dataset used is primarily constrained by the amount of time it will take to annotate it, which further impacts the generalizability of the study. Similar challenges are faced when considering conversations from low-resource languages, which may be harder to annotate by virtue of their limited resources. This is one of the major issues we address in the present study *i.e.,* can we automate the process of identifying when and how the listener backchannels from the data, thereby easing the annotation process?

Our novel contributions include:

(1) Use of self-training based semi-supervision for labeling the instances in a dataset for the presence or absence of listener backchannels, and therefore, exploring computational techniques to guide the development of conversational agents. This *identification* model partly replaces the human annotator, and therefore uses the listener's multimodal features to detect his/her backchannel responses. In addition to identifying backchannel opportunities, this step also identifies the type of signals (verbal, visual or both) associated with the backchannels. To analyze the feasibility of automating the labeling process via semi-supervision, we compare the backchannel *prediction* models trained on the labels assigned by the semi-supervised identification process, with the models trained on the ground truth labels (from the annotators).
(2) Inspired by Bevacqua *et al.*'s work [4] on personality contingent listener backchannels, we statistically analyzed how people with varying personalities emit different types of backchannel responses. In particular, we study the impact of the extraversion trait of a subject on their preference of modality [5] for their backchannel response.
(3) Finally, unlike most prior works, in addition to predicting the backchannel opportunities, we also predict the signal for the listener agent. The signal prediction task itself is way more challenging than the opportunity prediction task as the signals may vary significantly from person to person. We approach this task by first predicting the type of signal to emit (visual, verbal, or both). Then, based on our findings in (2) and depending on what personality we want our virtual listener to embody, we select the exact signal combination for it to express.

With this work, unlike most of the past studies that have focused on English datasets, we are also amongst the first to use peer-to-peer conversations in Hindi, which is a low-resource language, for listener backchannels[2]. In particular, we use the conversational dataset collected by Khan *et al.* [24], and annotate it for analyzing backchannels. Our quantitative and subjective evaluations reveal that:

(i) By leveraging semi-supervision for identification of listener backchannels, we were able to detect the presence of backchannels $\sim$ 90% of the times, and the type of signals associated $\sim$ 85% of the times, with only a small subset (25%) of labeled data.
(ii) Comparing the prediction models trained on the labels generated by the identification models, with those trained on manually-annotated labels: the former setting is able to reach $\sim$ 93% of the latter's performance in case of opportunity prediction, and $\sim$ 96% for signal category prediction. Note that the cost performance is significantly less for the former as it needs only a small amount of labeled data, thereby substantially reducing the efforts required in annotating the data.
(iii) Subjective analysis in the form of a user study with twenty seven participants supported our quantitative observations. Approximately 75% of the subjects found the backchannels produced by our proposed model more or equally natural to the responses by the model trained on annotated labels. The participants also confirmed our observations of personality impacting the preference of modality for backchannel responses.

[2]However, this does not entail that the techniques we propose cannot be used for datasets from other languages

The rest of this paper has been organized as follows: Section 2 discusses the dataset used in this work, the annotation process, and the initial data analysis. In Section 3, we formally describe the problem statements for the identification and prediction tasks and discuss how we model them. Additionally, this section also briefly discusses the features used for modeling these tasks. Section 4 begins by detailing the complete experimental setup and follows with a quantitative evaluation of all the tasks. The section also discusses the observations of a user study, performed to analyze the efficacy of our models in real-time. Finally, we conclude the paper with Section 6, which discusses the future scope of this work.

2 DATASET

In this work, we use *Vyaktitv*, a peer-to-peer Hindi conversations dataset, curated by Khan *et al.* [24]. The dataset provides audio and video recordings of participants involved in a dyadic conversation. There are a total of 25 conversations (50 individual recordings) with each one lasting 16 minutes and 6 seconds on an average. A total of 38 subjects (24 Male, 14 Female) were a part of the dataset. It also provides the Big Five personality traits [10] for all the subjects. Note that, to the best of our knowledge, no work has used Hindi conversations so far for analysing listener backchannels.

2.1 Annotating Listener Backchannels

The audio and visual feeds for the individual speakers were annotated for verbal and visual activity based backchannels. Specifically, three annotators used the ELAN annotation software [8] with a custom tier template to mark the onset and offset time for different backchannel signals including- *nod*, *head-shake*, *mouth*, *eyebrow*, and *short-utterance* (Table 1). The overall agreement amongst the annotators with respect to the presence of backchannels (*i.e.,* if any backchannel activity was present in a particular time range) was near perfect with a Fleiss' κ of 0.86. For the individual signals (κ): substantial agreement was observed for *nod* (0.71), *head-shake* (0.64), and *mouth* (0.63), and moderate agreement in case of *eyebrow* (0.45) and *short-utterance* (0.49). Similar values were also reported by other dataset annotators, like [1, 33]. A primary reason for relatively low values of annotator agreement in such datasets is the subjective nature of backchannels.

A consensus strategy was adopted to combine the three sets of annotations from different coders. In particular, we assigned a positive label to an instance, provided at least two annotators agreed upon the existence of backchannel activity during that time frame. The feedback signals associated with these backchannel instances (like a nod, smile, etc.) were also decided in a similar fashion using consensus voting, where a particular feedback signal is included in the final annotation if at least two annotators observed that signal's presence. Figure 1 depicts this strategy using an example where two annotators agree upon the existence of backchannel activity.

Note that the onset and offset times annotated by the three coders were bound to have slight variation even for the same instance. Therefore, to resolve this issue while combining the annotations, any two instances from different annotation sets (labeled by different coders) whose onset and offset time were less than 1 second apart, were considered to point to the same instance. The final value for the onset and offset time was taken as the average of the timestamps by different annotators.

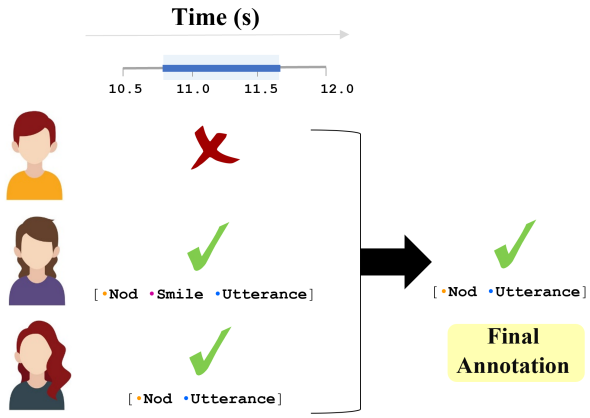


Figure 1: Sample depicting the consensus strategy adopted for combining the annotations from different coders.

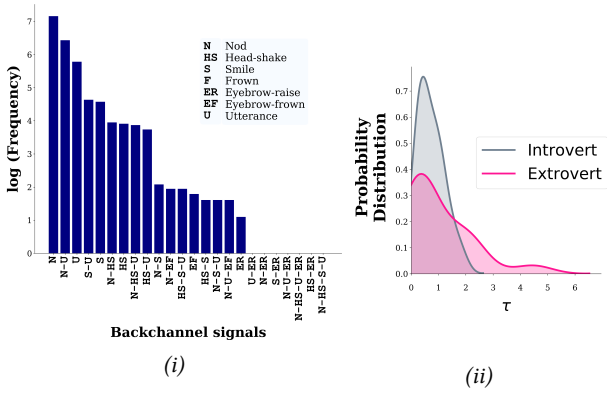
BC Signal	Labels (*Default*)	N	Mean Freq.	Mean Dur. (s)
Nod	*none*, nod	2037	42.4	1.7
Head-shake	*none*, head-shake	207	4.3	1.7
Mouth	*neutral*, smile/laugh, frown	227	4.7	1.8
Eyebrow	*neutral*, raise, frown	27	0.6	2.1
Utterance	*none*, short-utterance (eg: *"ohh", "okay"*)	1161	24.2	1.4

Table 1: Backchannel Signals and their Descriptive Statistics. (after taking consensus) N represents the total number of a particular feedback signal observed across the complete dataset. Mean Freq. is the average number of times a participant emits a particular signal during a conversation. Mean Dur. is the average duration of the signals in seconds (s).

2.2 Data Analysis

A total of 2781 backchannel instances were observed across all the participants in the dataset. In Table 1, we present a descriptive analysis for these instances in terms of the type of feedback signals. In particular, we record the total number (**N**), average frequency per participant (**Mean Freq.**), and the average duration (**Mean Dur.**) for each signal type. Note that a backchannel instance could have multiple feedback signals (eg., nod, and smile), and therefore, has been included in each of the possible signal categories. Observations from the analysis reflect how all the participants frequently used nods and short utterances. On the other hand, backchanneling via head-shakes, mouth, and eyebrow movements was not common. In particular, instances of backchanneling through eyebrow movements were scarce in the dataset (only 27), and therefore, we refrained from considering them in our predictive analysis. Furthermore, the average duration for all these backchanneling signals was around 1.8 seconds.

With Figure 2*(i)*, we aim to dig deeper by assessing the signals that co-occur frequently. In this study, we refer to a backchannel instance as *multimodal* if it has multiple associated signals; else, it is a *unimodal* instance. Note how unimodal nods and utterances



(i) *(ii)*

Figure 2: *(i)* Frequency of different combinations of backchannel signals emitted together by the subjects (across all subjects). *(ii)* Probability density distribution plot depicting the relation between extroversion and the ratio of multimodal to unimodal backchannel instances (τ) emitted by the subjects.

were amongst the top frequent signals emitted. Furthermore, most of the frequent multimodal instances included at least either a nod or an utterance. Multimodal signals where more than three types of signals co-occurred were infrequent.

Finally, we utilize the Big Five traits provided in the dataset to analyze if personality traits influenced the type of signals (multimodal or unimodal) emitted by the subjects. For our analysis, we defined the variable τ as the ratio of the number of multimodal to unimodal backchannel instances emitted by a subject. Then, we used a two-sample Kolmogorov-Smirnov (K-S) test, with the null hypothesis that the distribution of τ for the two categories of participants (based on a particular personality trait) was similar. The test presented an interesting insight in terms of the subjects' preference for multimodal and unimodal feedbacks. Of the five traits, the tests suggested a significant difference in the distribution of τ when considering the extraversion trait. The probability density distribution in Figure 2 *(ii)* shows how subjects with low extraversion scores (introverts) tended to have a lower value for τ, indicating their preference for unimodal signals, while extroverts used multimodal feedback more often. Quantitatively, the average probabilities of an extrovert emitting multimodal or unimodal backchannels were 0.51 and 0.49, respectively. The same values for an introvert were 0.35 and 0.65, respectively. This observation has been utilized later in this study while deploying the backchannel prediction models to a virtual-agent, thereby lending it a 'personality' based on which it can *choose* to emit multimodal or unimodal feedbacks in a probabilistic fashion (Section 4.3).

2.3 Negative Samples

Negative samples are the instances from the conversations where the listener did not emit any backchannel. Specifically, we evaluated the following two conditions while extracting such instances:

(1) the listener is not speaking, and
(2) the listener is not backchanneling in that time frame.

We used the Audacity tool[3] to extract the voice-activity of the listener. Combining the voice-activity and the annotations, we extracted regions from the conversations that met the above two conditions. We sampled disjoint instances from these regions, where the length of each instance (in seconds) was taken as a random floating-point number in the range [1.06, 5.43]. This range also determined the lengths of the backchannel instances, and hence the choice. Using this approach, we sampled a total of 2670 negative instances.

3 METHODOLOGY

This section elaborates on the two-step methodology followed in the present study. We begin with the backchannel opportunity and signal **identification** module, which utilizes the *listener's features* to automatically classify the instances into different categories based on the presence or absence of backchannel activity. This step employs a semi-supervised training paradigm and aims to simplify the manual annotation task. Next, we discuss the **prediction** module, which uses the *speaker's contextual* features to predict these backchannels. The training step of this latter module makes use of the labels generated by the former semi-supervised step. Figure 3 summarises this workflow.

3.1 Semi-Supervision for Backchannel Opportunity and Signal Identification

Task Formulation. We begin this module by first formally defining the task of identifying backchannel opportunities and signals. Consider a time frame T_{ij}, which starts at the ith second and ends at the jth, and let L_{ij} represent the listener's visual and acoustic features in that time frame. Then, our aim for the backchannel opportunity identification task is to learn a function $\mathcal{F}_{bc}$ mapping a time series in the listener's feature space to the corresponding backchannel opportunity label $\mathcal{BO}_{ij}$ (binary label signifying the presence or absence of backchannels in the time T_{ij}), *i.e.*,

$$\mathcal{F}_{bc}(L_{ij}) \mapsto \mathcal{BO}_{ij} \tag{1}$$

We model (identify and predict) the backchannel signals differently from the literature. Instead of identifying the different signals (like nod, head-shake, etc.) individually, we categorize them into two types- *visual* (nod, mouth, head-shake) and *verbal* backchannels (utterances). This is done because the task of predicting the exact backchannel signal that a listener must emit based on the speaker's context (*i.e.,* the signal prediction task) is challenging, primarily because of the subjective nature of these signals. For instance, subject A may emit a smile to a particular speaker context, whereas another subject B may emit a nod in response to the same. Grouping signals together into visual and verbal categories simplifies the tasks at hand.

A backchannel instance could be associated with visual, verbal, or both kinds of signals. Therefore, the signal identification task aims at finding the type of signal the listener emits whenever s/he backchannels. The goal is to learn a mapping function $\mathcal{F}_{sig}$ from the listener's feature space to one of the three signal categories (verbal, visual, both) ($\mathcal{BS}_{ij}$), *i.e.,*

[3]https://www.audacityteam.org/

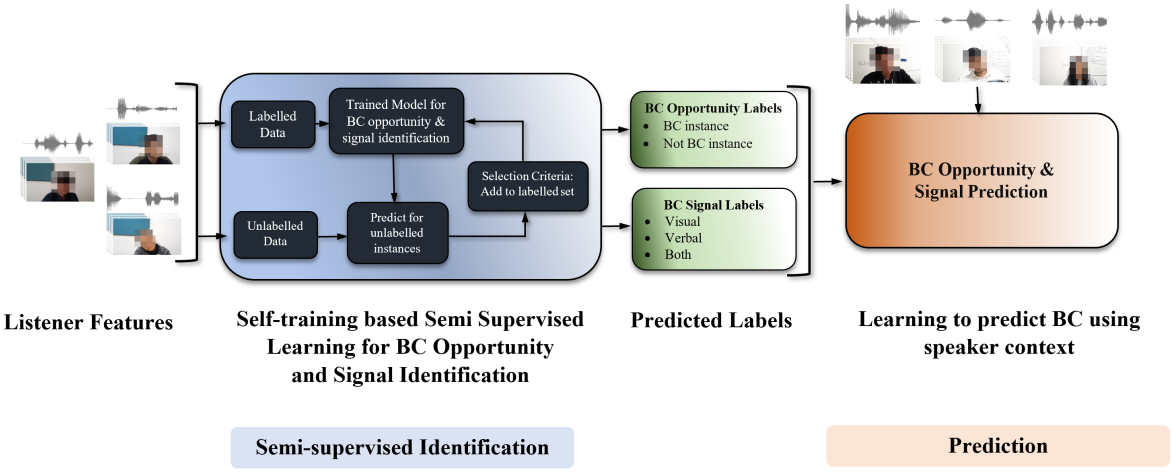


Figure 3: Methodology: *(i)* Semi-supervised learning for identifying backchannels and type of signals emitted using a subset of labeled data. *(ii)* Learning to predict these instances and signals using the speaker's context.

$$\mathcal{F}_{sig}(L_{ij}) \mapsto \mathcal{BS}_{ij} \tag{2}$$

It is important to note that both the identification tasks make use of only the listener's features. This is a crucial distinction from the prediction module. The identification module identifies backchannel opportunities and signals much like a human annotator, by paying attention to the listener. On the other hand, as expected, the prediction module would use only the speaker's contextual features to predict these backchannels.

Modelling: Semi-Supervised Learning. Annotating large datasets is a tedious task, and researchers have long been exploring ways to ease out and automate this manual process [2, 6, 19, 31]. Even in the context of the present study, the annotation process took around 90 hours[4], where the three annotators viewed and labeled all the conversations. This indeed is a bottleneck! To the best of our knowledge, this challenge in developing ECAs, specifically for modeling the backchannel behavior of an active listener, is an open research gap that has not been investigated by prior literature. The identification module of our workflow is a step towards tackling this challenge.

Several AI techniques based on Semi-Supervision [14, 21] and Weak-Supervision [11, 26] have been utilized to decrease the annotation costs in different application domains. Here we explore self-training based semi-supervised learning paradigm [37] for identifying listener backchannel instances and the signals associated with them, using only a small subset of the manually annotated data. The following steps summarise the general approach to self-training based semi-supervision adopted here (Figure 3):

(1) We start with the labeled portion of the dataset ($\mathcal{L}$) to train an initial classifier (C) that learns to identify the backchannel instances and the signals associated based on the listener's features.
(2) C is then used to predict the labels for the unlabeled data ($\mathcal{U}$).
(3) Of these predictions, the instances which meet a specific *selection criterion* are removed from $\mathcal{U}$ and added, along with their predicted pseudo labels, to the training set. This updated training set comprising the initially labeled ($\mathcal{L}$) and the newly added pseudo labeled instances are used to train the classifier C again.
(4) The cycle continues until no new instance matches the selection criteria.

In the present study, we have the labels for the complete dataset ($\mathcal{D}$). Therefore, in our context, $\mathcal{L}$ and $\mathcal{U}$ are disjoint subsets of $\mathcal{D}$[5]. Furthermore, as selection criteria for Step-3, we use a high threshold value of 0.90 on the predicted class probability. Note that after Step-4, we use the trained classifier C to predict the pseudo labels for all the remaining instances from $\mathcal{U}$. Our experiments revealed that such instances were very few in number.

3.2 Backchannel Opportunity and Signal Prediction

The formal definition for the backchannel opportunity prediction task is similar to the one proposed in prior literature [15]. Consider a time window T_{ij}, and let S_{ij} be the speaker's visual and acoustic features for that period. The backchannel opportunity prediction task entails predicting whether the listener would backchannel after T_{ij} (*i.e.,* the label $\mathcal{BO}_{ij+}$) using only the speaker's features (context). Similarly, the signal prediction task aims to predict the type of feedback signal (visual, verbal, or both) (*i.e.,* $\mathcal{BS}_{ij+}$) that will be emitted by the listener using the speaker's context. The following function mappings represent these tasks:

[4] 30 hours each annotator, taking the average time to annotate one side of a conversation as 35 minutes. The total amount of recorded content being ~ 13.5 hours long.

[5] This notion becomes more apparent as we discuss the experimental setup in Section 3.4

$$\mathcal{G}_{bc}(S_{ij}) \mapsto \mathcal{BO}_{ij+} \tag{3}$$

$$\mathcal{G}_{sig}(S_{ij}) \mapsto \mathcal{BS}_{ij+} \tag{4}$$

Inspired by literature [15, 28] we extract a 3 second context window before each instance in our dataset and use the speaker's visual and acoustic features for that time frame to predict listener backchannels.

The prediction tasks are performed in a supervised fashion, with the labels derived from the semi-supervision based identification module. Since the identification module used the listener's channel (features), the prediction module is essentially learning based on cross-channel semi-supervised labels.

3.3 Feature Extraction

This section elucidates all the features extracted from the dataset for modeling the identification and prediction tasks. Table 2 summarises these visual and prosodic features. In particular, inspired by prior work [15, 22] we use OpenFace [3] to extract- 18 facial action units (FAUs), velocity and acceleration of gaze, translational and rotational head velocities and accelerations, blink rate, pupil location, and smile ratio. Additionally, we also find the gaze state as a categorical variable, taking up three values- `left`, `right`, or `blinking`[6]. These features from the listener's channel are used for backchannel opportunity, and signal **identification** tasks and the **prediction** tasks utilize the corresponding features from the speaker's channel.

As prosodic features, in addition to the voice activity (discussed in Section 2.3), we also extract the fundamental frequency (`F0`), the energy, and the first 13 Mel-Frequency Cepstral Coefficients (MFCC) using the pyAudioAnalysis library [13]. All of them are used for the identification as well as the prediction tasks.

3.4 Experimental Setup

We now discuss the experimental settings for our work. Since each task can be analyzed and evaluated based on several dimensions, such as the type of features used (multimodal vs. unimodal), and the amount of initial 'labeled' data in semi-supervision ($\mathcal{L}$), we could theoretically have had a combinatorial number of settings corresponding to all pairs of values for these variables. However, we follow a more structured pattern to limit the number of settings and draw focus on the most crucial elements.

3.4.1 Backchannel Opportunity and Signal Identification. Separate models were trained for backchannel opportunity and signal identification tasks. We tried several different machine learning algorithms for training the classifier C in a semi-supervised paradigm described in Section 3.1. Specifically, we experimented with Random Forests (**RF**), Support Vector Machine Classifier (**SVC**), K-Nearest Neighbour Classifier (**KNN**), AdaBoost (**ADA**), and ResNet (**ResNet**). ResNet is the state-of-the-art deep learning model for time series classification [12]. Additionally, we also experimented with the widely used Label Spreading algorithm (**LSpread**, implemented as a part of the scikit-learn python library [30]) for semi-supervised learning as a baseline. All but for the ResNet model were trained using the mean and standard deviation aggregates of the time series based listener features. ResNet was trained using the detailed time series features for the complete time window. Finally, for the proportion (x) of the dataset $\mathcal{D}$ taken as the initial 'labeled' data ($\mathcal{L}$) for training C, we experimented with all values in the range (5%, 100%) with a step size of 5%. Although semi-supervision is applicable only when the amount of unlabelled data exceeds the labeled, experimenting with the whole range helps in the analyzing the models' sensitivity, *i.e.*, how the performance changes when we approach the fully supervised setting (by increasing x).

Manifold evaluation for the identification tasks can be easily understood from Figure 4. First, we randomly create 5 folds from the data ($\mathcal{D}$). Four of these folds are used for training C via semi-supervision; *i.e.,* a random x% sample of the data from these four folds serve as the initial 'labeled' set ($\mathcal{L}$), while the rest is termed as 'unlabelled' ($\mathcal{U}$). Once the model has been trained, it is evaluated on the 5th fold, *i.e.*, the pseudo labels produced by the identification models are compared against the ground truth labels. To ensure that our models do not over-fit to a particular random split, we run 10 simulations of training and evaluation for each pair of values of C and x, and report the average results across all the simulations[7]. As evaluation metrics, we use the weighted average precision, recall, F1-score, and overall accuracy, for both the opportunity and signal identification tasks.

Features	Description
Visual Features	
`FAUs`	Regressive values of 18 Facial Action Units: `AU01_r, AU02_r, AU04_r, AU05_r, AU06_r, AU07_r, AU09_r, AU10_r, AU12_r, AU14_r, AU15_r, AU17_r, AU20_r, AU23_r, AU25_r, AU26_r, AU28_r, AU45_r`
`gaze_vel, gaze_acc`	Velocity and acceleration of eye gaze
`gaze_state`	Categorical feature signifying the direction of gaze as- `left`, `right`, `blinking`
`head_vel_T, head_acc_T`	Translational velocity and acceleration of head
`head_vel_R, head_acc_R`	Rotational velocity and acceleration of head
`blink_rate`	First order differential of Eye Aspect Ration
`pupil`	Location of the pupils.
`smile_ratio`	Stretch of the smile calculated as the ratio of two characteristic dimensions of the mouth [1]
Prosodic Features	
`F0`	The fundamental frequency of the speech signal
`energy`	The sum of squares of the signal values, normalized by the respective frame length
`mfcc`	Mel-Frequency Cepstral Coefficients 1-13
`voice_activity`	Binary state characterizing whether the person is speaking or not, based on acoustic signals.

Table 2: Visual and vocal prosodic features used in the study.

[6]The following GitHub repository was used for the same: https://github.com/antoinelame/GazeTracking

[7]The results for each simulation were taken as the average of all 5 folds.

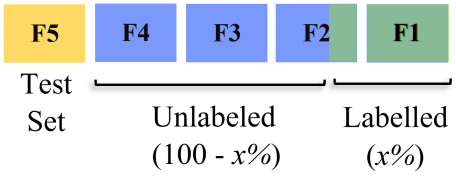


Figure 4: 5-Fold evaluation of semi-supervised models for backchannel opportunity and signal identification.

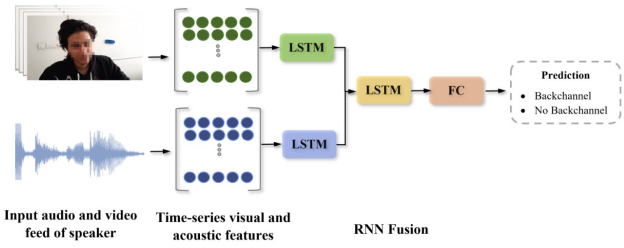


Figure 5: Multimodal RNN Fusion based architecture for backchannel opportunity prediction model.

3.4.2 Backchannel Opportunity and Signal Prediction. We begin evaluating the prediction experiments by first analyzing the impact of using different sets of features *i.e.,* unimodal vs. multimodal features. This is done by using the supervised (manually annotated) labels for both training and evaluation. Inspired by Tavabi *et al.* [35], we use a multimodal RNN fusion-based architecture for backchannel opportunity prediction, shown in Figure 5. In essence, the time-series features for video and audio modalities are first individually passed through LSTM encoders. Following this, the encoders' outputs are concatenated and again passed through a recurrent layer, enabling the model to learn the temporal dependencies between modalities. The latent representation from this fusion layer is finally passed to a softmax layer through a fully-connected layer to predict feedback opportunities. The two unimodal models simply include a softmax layer after the encoder to output predictions.

The signal prediction task was more complex, as the labels for the task also suffered from a serious class imbalance (*visual:* 1593, *verbal:* 326, *both*: 835) which was adversely impacting the predictive performance of the models[8]. Therefore, we use the mean and standard deviation aggregates of the time series based speaker context features, along with `SVM-SMOTE` [7] to handle this imbalance[9]. Note that signal prediction was made using only the feature set (unimodal/multimodal) that performed the best for the opportunity prediction task. Using the upsampled data[10], we tried several models for signal prediction, including Random Forests (**RF**), Support Vector Machine Classifier (**SVC**), AdaBoost (**ADA**), K-Nearest Neighbors (**KNN**), and a Multi-Layered Perceptron model (**MLP**)[11].

The next set of experiments analyze the feasibility and the performance to cost trade-off when adopting semi-supervision to label the dataset. For this, we compare the prediction tasks in two paradigms:

(1) A complete supervised setting, using true labels provided by the annotators during the training and evaluation phases. Note that previous evaluations of the prediction tasks already produced the results for this setting.
(2) The proposed setting using labels generated by the semi-supervised identification modules for training, while evaluating on the true labels. For this, we use only the best feature set (unimodal/multimodal) found via the previous evaluations, along with the pseudo labels produced by the best-found pair of C and x from the identification tasks.

With these two paradigms, we assessed how far semi-supervised backchannel identification fetches in comparison to human annotations for the downstream prediction tasks.

For evaluating the backchannel prediction models, several recent works have used the leave-one-subject-out approach [22]. In the present study, we use a slightly modified version of this technique to evaluate our prediction models in different paradigms. Instead of having the data from each subject comprise a test fold, we divide the subjects into 6 groups, and data from each of these groups form a test fold, *i.e.,* leave-one-group-out. This modification was primarily done because some subjects had very few backchannel instances in their conversations. Combining instances from multiple subjects to form a test set helped in better analyzing the model predictions and keeping the number of folds tractable. As metrics for the backchannel opportunity prediction task, we report the positive class' precision, recall, and F1-score, as well as the overall accuracy. For signal prediction, we report the weighted average precision, recall, and F1-score, the overall accuracy, and the confusion matrices.

4 RESULTS

4.1 Identification Tasks

In Figure 6, we have the sensitivity analysis in terms of accuracy of all the models we tried for the backchannel opportunity and signal identification tasks. Each curve in these plots represents how the performance of a particular classifier (C) for the identification task changes as we increase the amount of initial labeled data available (x) for semi-supervision. Using this information, we looked for the best possible pair of values for C and x. However, in addition to achieving high performance, we also wanted to limit x. For this, we manually found the *elbow* point for the best performing classifier in each task. The elbow value for x is one where decreasing it further would cause a drastic drop in performance, while increasing it would not change the performance significantly. Ideally, we also wanted x to be less than 50%, where we could have more unlabelled than labeled data for semi-supervision, thereby improving the annotation cost trade-off.

[8]Note that though the same data was also used for the signal identification task, the imbalance did not cause a degradation in the performance there.
[9]SMOTE takes data of the shape ($n_samples$, $n_features$) as input, and therefore, time-series data had to be aggregated. This also eliminated the utility of models like ResNet for the task.
[10]Only the training set was upsampled i.e., the test set remained as is.
[11]Please refer the supplementary material for details on the hyper-parameters used for all the different models in the identification and prediction tasks

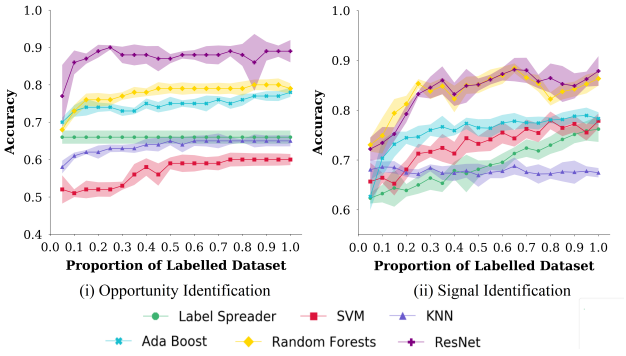


Figure 6: Sensitivity Analysis using accuracy as the metric for semi-supervised backchannel opportunity and signal identification models. The shaded portion represents standard deviation.

Model	Opportunity Identification				Signal Identification			
	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
LSpread	0.66	0.66	0.66	0.66	0.67	0.65	0.66	0.65
SVC	0.56	0.52	0.56	0.54	0.69	0.70	0.68	0.71
KNN	0.63	0.63	0.63	0.63	0.63	0.67	0.62	0.67
ADA	0.74	0.74	0.74	0.74	0.75	0.75	0.69	0.75
RF	0.76	0.76	0.76	0.76	**0.86**	**0.87**	**0.86**	**0.85**
ResNet	**0.90**	**0.90**	**0.92**	**0.90**	0.81	0.84	0.83	0.83

Table 3: Backchannel opportunity and signal identification: Detailed results for x = 25%.

Figure 6 (i) shows how ResNet outperforms all the other models in terms of the listener backchannel opportunity identification task. Observe how increasing x beyond 25% does not change the performance a lot. It also meets all the other criteria we discussed above. Therefore, we choose $x = 25\%$ as the final seed value (used to initialize the amount of labeled data) for the opportunity identification task. For the signal identification task (Figure 6 (ii)), we observe that ResNet and Random Forests both have somewhat overlapping accuracy for different values of x. However, by taking a closer look, we analyze that for values of $x < 30\%$, Random Forests lead ResNet. The elbow point for Random Forests is at 25%, while that for ResNet (with a similar accuracy value) is observed at 35%. Therefore, for the signal identification task, our final choice is Random Forests with $x = 25\%$. The detailed results obtained for other classifiers with x set to 25% for both the opportunity and signal identification tasks are shown in Table 3.

4.2 Prediction Tasks

We begin discussing the prediction models with Table 4, which records the results of our backchannel opportunity prediction task performed using the annotated labels for training, while experimenting with different subsets of input features. The unimodal-audio model, with positive class F1 of 0.74 and overall accuracy of 0.70, outperforms the unimodal-video model, with the corresponding metric values of 0.71 and 0.66, respectively. The multimodal model utilizing both the video and audio features beats the audio model by a small margin, attaining an F1 of 0.75 and 0.72 accuracy.

Feature set	Precision	Recall	F1-score	Accuracy
Video	0.61	0.86	0.71	0.66
Audio	0.64	0.90	0.74	0.70
Video + Audio	**0.66**	**0.89**	**0.75**	**0.72**

Table 4: Backchannel opportunity prediction model trained trained across different sets of features using supervised (manually-annotated) labels for training. Overall, the multimodal feature set performed the best.

Labels used	Precision	Recall	F1-score	Accuracy
Supervised	0.66	0.89	0.75	0.72
Semi-Supervised	0.62	0.82	0.70	0.66

Table 5: Backchannel opportunity prediction: comparison of model trained using supervised manually-annotated labels with the one using labels generated by the identification module.

Model	Supervised Labels				Semi-Supervised Labels			
	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
SVC	0.80	0.80	0.80	0.80	0.75	0.75	0.75	0.75
KNN	0.74	0.74	0.73	0.74	0.72	0.72	0.71	0.72
ADA	0.81	0.81	**0.81**	0.81	0.78	0.78	**0.78**	0.78
RF	0.78	0.78	0.78	0.78	0.76	0.75	0.75	0.75
MLP	0.80	0.81	0.80	**0.87**	0.75	0.74	0.74	**0.84**

Table 6: Backchannel signal prediction: comparison of models trained using supervised (manually-annotated) labels, and those using labels generated by the semi-supervised signal identification Random Forest model (with 25% initial labeled data).

True \ Predicted	C1	C2	C3
C1	0.71	0.07	0.23
C2	0.07	0.87	0.06
C3	0.27	0.05	0.67

(i)

True \ Predicted	C1	C2	C3
C1	0.70	0.09	0.21
C2	0.08	0.80	0.12
C3	0.23	0.06	0.72

(ii)

Table 7: Confusion matrices for the best backchannel signal prediction model (i) trained on manually-annotated labels, (ii) trained using labels generated by the signal identification models. Here, C1, C2, C3 refer to the 'visual', 'verbal', and 'both' class labels, respectively.

Since this set has overall the best performance for the baseline opportunity prediction task with supervised labels, we use the same to train and analyze models for subsequent tasks as well.

Continuing with opportunity prediction, we now discuss the results obtained using the semi-supervised labels generated by the opportunity identification model for training (the evaluation/test set used the same supervised labels). Table 5 records these results. Evidently, with the labels generated using just 25% of the annotated data (in the identification step), we are able to achieve nearly 93% of the supervised F1 and 92% of the supervised accuracy scores on the opportunity prediction tasks. The values of these metrics observed here are 0.70 and 0.66, respectively.

Moving on to backchannel signal prediction for determining the category of signals for the listener to emit, Table 6 shows the results obtained for all the models trained in different settings. When using the manually-annotated labels for training, we observe that the **MLP** model has the best accuracy of 0.87, however in terms of the F1-Score, the **ADA** model outperforms all the others. Given the data imbalance in the signal prediction task, we use F1-score as the deciding metric, and choose **ADA** as the best performing model, with an F1-score of 0.81. We observe a similar performance trend when using the semi-supervised signal identification model to generate the training labels. Here as well, **ADA** is the best performing model in terms of F1-score, with a value of 0.78, which is 96% of the corresponding metric for the supervised setting. In Table 7, we report the confusion matrices for the best (**ADA**) signal prediction models found for the two settings. We used the worst-case confusion matrix [25] computed using all the matrices generated across manifold evaluation of the model. The performance for the 'visual' (C1) class remained almost unchanged in the two settings. The false positives between the 'visual' and the 'both' (C3) classes reduced when using semi-supervised labels. The 'verbal' (C2) class performance dropped slightly. On the other hand, we performed better for the class C3, but that came at the cost of a slight increase in the false positives from C2. Overall, the performance of the model trained using semi-supervised labels was indeed comparable with the one trained on manually-annotated labels.

4.3 Subjective Evaluation: User Case Study

For problems like predicting listener backchannels, quantitative evaluations based on different metrics may not be sufficient to analyze the models' efficacy. For such cases, qualitative assessments become extremely important. Many prior studies have used ECAs or robots to deploy their models and assess the backchannel predictions. However, in the present study, we follow a slightly different approach, which is partly inspired by [15] and [28]. Specifically, using Apple's Memoji feature, we created two virtual avatars- *Arjun* (introvert) and *Karan*[12] (extrovert). We used the same Memoji technology to record (as short clips) the neutral state and different combinations of backchannel signals, mentioned in Table 8, for each of them. Furthermore, we prepared a short video compilation (nearly 4 minutes long) of a few speakers from the dataset used in this work. Using our prediction models[13], along with a personality contingent signal combination sampling technique, we recorded the backchanneling responses that Arjun and Karan would emit for the speakers in the compilation. Although the sampling is not the main focus of this work, we briefly describe below the steps followed for re-use by future works:

- After getting the predictions for backchannel opportunity, and the signal category from our models, we use inverse transform sampling to decide whether to emit unimodal or multimodal signals (in case of visual and verbal categories only) based on the personality being modeled (extrovert or introvert). Note that the probabilities used here are calculated from the data and are mentioned in Section 2.2.
- Now, we choose the exact signals to emit from Table 8 using inverse transform sampling, where the probabilities for each signal combination are calculated from the data[14].

	Unimodal	Multimodal
Visual	nod (0.83), head-shake (0.08), smile (0.09)	nod + smile (0.60), head-shake + smile (0.40)
Verbal	utter: top short utterances like *okay*, *hmm*, *haan* (yes), etc., were proabbilistically sampled	nod+utter (0.80), smile+utter (0.13), head-shake+utter (0.05), head-shake+utter+smile (0.01), nod+utter+smile (0.01)
Both	-	

Table 8: Signal combinations from different categories used in the present study. The corresponding normalized probabilities inferred from the data and used while sampling the signal combinations are also shown.

We recorded six response videos, three for each Arjun and Karan, using the following different prediction policies:

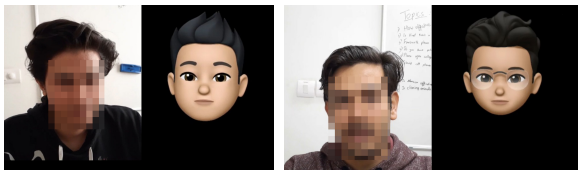
(1) Random Prediction Model (**Random** policy): A baseline model that predicted backchannel opportunities and the signal categories at **random**.
(2) Supervised Model (**MA** policy): The best opportunity (multimodal) and signal category (**ADA**) prediction models trained on the **m**anually-**a**nnotated labels.
(3) Semi-supervised Model (**SSL** policy): The opportunity (multimodal) and signal category (**ADA**) prediction models trained with the labels generated from the **s**emi-**s**upervised identification models (**ResNet** and **RF**, respectively with 25% data) used in this work.

With these short clips recording the backchannel responses as predicted by different models for Arjun and Karan, we wanted to test the following three hypotheses:

[H1]: As perceived by a human watching the response videos, the backchannels predicted by our models look more natural, and therefore, better than those predicted by the random policy model.
[H2]: From a user's perspective, there is no significant difference in the quality of backchannel responses emitted by models in the settings 2 and 3 above, *i.e.,* using semi-supervised learning with a small subset of labelled data does not impact the quality of backchannel responses generated by the final prediction models.
[H3]: Finally, to assess the extent to which backchannel responses depend on personality, and the utility of using such personality contingent signal sampling: A human can judge the extraversion personality trait for Arjun and Karan based on their response videos.

Twenty seven students were recruited as participants from a university's mailing list. They were all native Hindi speakers. As an introduction, the participants were acquainted with the concept

[12]The names used are hypothetical, and do not compromise anonymity.
[13]Backchannel opportunity prediction was done at an interval of 3 seconds. Depending on the output, the signal categories were predicted. We also ensured that the models used for different speakers in the compilation were the ones where the speaker belonged to the test set.
[14]The only step that is contingent on personality is deciding on the unimodal vs. multimodal signals. The second step of signal sampling uses common probabilities for both characters.



(*i*) (*ii*)

Figure 7: Stills from two of our user study videos- (*i*) Karan, and (*ii*) Arjun

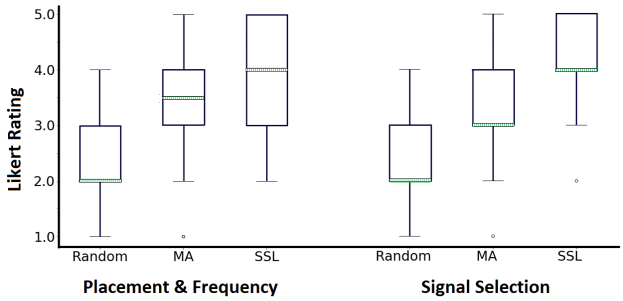


Figure 8: Box & Whisker plot for the Likert scale ratings (where 1 is *very poor* and 5 is *very good*) given by the participants to each of the three policies, based on (*i*) placement & frequency, and (*ii*) signal type selection.

of listener backchannels. They were instructed to watch all six response videos and carefully observe the virtual listeners' backchannel responses in each one. It was a blind study, *i.e.* the participants knew neither about the prediction policies nor about the listeners' personality traits in the videos.

Furthermore, to ensure that the appearance of the two avatars did not bias the participants' judgement, we switched the avatars (with everything else as is) for half the participants. In other words, for half of the participants, Arjun acted as an extrovert, and vice versa. With this, we will be able to analyze whether backchannels indeed depend on the extraversion personality trait.

After watching all the videos, the participants were presented with a questionnaire meant to test the three hypotheses mentioned above. In particular, we asked the participants to

- Rank the three policies in order based on the quality of backchannels, taking into consideration factors like the placement, frequency, and selection of backchannel signals.
- Rate each one on a Likert scale based on the above parameters.
- Identify the personality traits of the avatars based on the videos.

Post survey analysis of the participants' responses lead to some exciting findings which supported our hypothesis:

(1) **Models from both the manually-annotated and semi-supervised settings performed better than random:** 95% of the participants ranked our prediction models to be better than the random prediction policy. In addition, the Likert scale ratings suggested that our models performed *significantly* better than the random policy, both in terms of the placement & frequency, as well as the signal selection. We used the Wilcoxon signed-rank test, and the corresponding p-values are [**P**- placement & frequency, **S**- signal selection, $*$ indicates significant difference at 95% confidence interval]:
   - **MA** > **Random**: 0.0019 (**P***); 0.0013 (**S***)
   - **SSL** > **Random**: 0.00011 (**P***); 0.0001 (**S***)

   A box-plot for these ratings is shown in Figure 8. Notice how the boxes for **MA** and **SSL** do not overlap with **Random**. This also indicates that the former two certainly received better ratings than the later. This validates our first hypothesis **[H1]** that with a data-driven approach, we were able to emit more natural and human-like backchannels.

(2) **Prediction models trained on semi-supervised labels produced more natural backchannels:** Comparing the ratings provided for the **MA** and the **SSL** policies, we found that in terms of the frequency and placement (**P**), there was no *significant* difference between the two (p-value=0.1024). For signal selection (**S**), the ratings suggested that **SSL** was significantly better than **MA** (p-value=0.0122). The box-plot in Figure 8 also aligns with these results. In case of Placement & Frequency, the boxes for **MA** and **SSL** overlap, with the median line for **SSL** slightly above **MA**'s. This indicates that there is likely to be a difference between the two sets of ratings, even though it may not be significant (as found from the test). For Signal Selection, the boxes do not overlap at all, indicating a difference.
   Furthermore, as a part of the questionnaire, we also asked the participants to compare the response videos generated by these two prediction models. Of the total, 60% of the participants found the backchannel responses emitted by the **SSL** model more natural than the **MA** model (**SSL** > **MA**), and 15% observed no perceptible difference between the two (**SSL** ~ **MA**). This also aligns with our quantitative prediction results where the proposed **SSL** model was able to reach ~ 95% of the latter's performance. Thus, the two models were, both qualitatively and quantitatively, very similar.
   This confirms our second hypothesis **[H2]** with most of the participants finding the proposed model (**SSL**) similar (or more natural) to the **MA** model.

(3) **Karan and Arjun's extraversion traits were perceptible:** 80% of the participants were able to accurately identify which of the two virtual-human characters was introvert and extrovert. This confirms our third and final hypothesis **[H3]** that the type of backchannel signals emitted by an individual indeed depends on their personality.

5 DISCUSSION

Relation to prior work and some interesting findings: Our quantitative and qualitative evaluations in the previous section greatly validate how semi-supervision can be extremely useful in designing human-like ECAs, by focusing on the task of listener backchannel prediction. We found that with just 25% of manually annotated data (175 minutes), we were able to train a backchannel prediction system that performed comparably well, and even better

in terms of some parameters, as the one trained using 100% data. Most of the prior works, even the most recent ones, including [15, 16, 29, 32], have depended on a large amount of annotation. We believe that this observation can significantly benefit the HCI community. Furthermore, studies have shown that backchannel responses vary greatly with culture [18, 42], and most of the prior studies have focussed primarily on the American and European population in this regard. In this work, we worked with subjects of Indian origin, and therefore, our work holds cultural significance as well.

A particularly interesting finding from Section 4.3 was that most participants found the backchannel responses generated by the **SSL** policy more 'natural' than the **MA** policy. We hypothesise this could be attributed to some form of label noise: After annotations, we only took those positive instances where at least two raters agreed, and the rest were discarded. Our negative samples overlapped with those instances where only 1 of the raters had annotated a positive BC. The **SSL** model could be learning to predict these backchannels. In fact, in our demo videos, we found two such instances. Starting with a small amount of seed data, the SSL model could have been learning these instances as well (given there are some hints of BC), which could explain the observation that even though SSL performed comparably with **MA** quantitatively, but seemed more natural to the participants in the subjective study.

Limitations: We would also like to highlight some limitations of our work, which can form the basis for future studies. First, we only used the participants' visual and acoustic features and did not include the content of the conversation itself for predicting backchannels. The main reason for this was that the primary language used in the *Vyaktitv* dataset was Hindi. The English translations for the dialogues were not available, making the use of state-of-the-art NLP techniques non-trivial. In another aspect, our qualitative user case study involved a short 4-minute long compilation of speakers from the *Vyaktitv* dataset. A more rigorous analysis could follow by deploying the models and the avatars as a real time-system. Finally, we did not annotate eye blinks/gaze [20] as backchannel signals, primarily because they were not as apparent in the dataset. This could be a cultural difference as well.

Ethical Consideration: When developing data-driven systems leveraging data from human-subjects, it becomes imperative that we respect their privacy boundaries. We want to state that no Personally Identifiable Information (PII) was used while training or evaluating the system. We also complied with the agreements in the *Vyaktitv* dataset to ensure the safe use of data.

6 CONCLUSION

In this work, we confirmed the feasibility of using semi-supervised learning to (semi-) automate the process of identifying and labeling listener backchannel instances (both opportunities and the associated signals) from conversations. We used a Hindi peer-to-peer conversation-based multimodal dataset *Vyaktitv*, for our experiments. However, the methodology proposed in the study is general and can be adapted for other conversational datasets as well. Quantitative evaluation alongside a subjective analysis in the form of a user study strongly validated our hypothesis that- prediction models trained using semi-supervised labels performed comparably with those using manually annotated labels. Furthermore, we statistically and qualitatively confirmed that the type of backchannel signals emitted are intimately linked to an individual's personality (extraversion in particular).

Future work directions include validating the scope of semi-supervised learning for listener backchannel prediction on other datasets. Other parallel tasks, like listener disengagement prediction, can also be similarly performed. Methodologically, future studies could explore by devising heuristics and using weak supervision based techniques to identify backchannels. Our observations from analyzing the impact of personality on the type (modality) of backchannel response also open some new research questions; for instance- can we also analyze if the frequency of backchannels emitted by different individuals depend on personality [4]? Furthermore, can we use these findings to embed personality into a virtual human further? These are some exciting lines future researchers can look into by conducting more extensive analysis.

ACKNOWLEDGEMENTS

Jainendra Shukla and Rajiv Ratn Shah are partly supported by the Infosys Center for AI and Center for Design and New Media at IIIT Delhi. In addition, Rajiv Ratn Shah is also partly supported by the ECRA Grant (ECR/2018/002776) by SERB, Government of India. Finally, we would like to thank the annotators, and all the participants who took part in our case study for their time and efforts.

REFERENCES

[1] Samer Al Moubayed, M. Baklouti, Mohamed Chetouani, Thierry Dutoit, Ammar Mahdhaoui, Jean-Claude Martin, Stanislav Ondáš, Catherine Pelachaud, Jerome Urbain, and M. Yilmaz. 2009. Generating Robot/Agent backchannels during a storytelling experiment. *Proceedings - IEEE International Conference on Robotics and Automation* ICRA'09 (06 2009), 3749 – 3754. https://doi.org/10.1109/ROBOT.2009.5152572

[2] Stéphane Ayache and Georges Quénot. 2008. Video corpus annotation using active learning. In *European Conference on Information Retrieval*. Springer Berlin Heidelberg, Berlin, Heidelberg, 187–198.

[3] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* 1 (2018), 59–66.

[4] Elisabetta Bevacqua, Etienne De Sevin, Sylwia Julia Hyniewska, and Catherine Pelachaud. 2012. A listener model: introducing personality traits. *Journal on Multimodal User Interfaces* 6, 1-2 (2012), 27–38.

[5] Elisabetta Bevacqua, Maurizio Mancini, and Catherine Pelachaud. 2008. A listening agent exhibiting variable behaviour. In *International Workshop on Intelligent Virtual Agents*. Springer Berlin Heidelberg, Berlin, Heidelberg, 262–269.

[6] Patrícia Bota, Joana Silva, Duarte Folgado, and Hugo Gamboa. 2019. A semi-automatic annotation approach for human activity recognition. *Sensors* 19, 3 (2019), 501.

[7] Kevin W. Bowyer, Nitesh V. Chawla, Lawrence O. Hall, and W. Philip Kegelmeyer. 2011. SMOTE: Synthetic Minority Over-sampling Technique. *CoRR* abs/1106.1813 (2011), 321—-357. arXiv:1106.1813 http://arxiv.org/abs/1106.1813

[8] Hennie Brugman, Albert Russel, and Xd Nijmegen. 2004. Annotating Multimedia/Multi-modal Resources with ELAN. *LREC* Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04) (2004), 2065–2068.

[9] Justine Cassell, Joseph Sullivan, Elizabeth Churchill, and Scott Prevost. 2000. *Embodied conversational agents*. MIT press, Cambridge, MA, United States.

[10] John M Digman. 1990. Personality structure: Emergence of the five-factor model. *Annual review of psychology* 41, 1 (1990), 417–440.

[11] Jared A Dunnmon, Alexander J Ratner, Khaled Saab, Nishith Khandwala, Matthew Markert, Hersh Sagreiya, Roger Goldman, Christopher Lee-Messer, Matthew P Lungren, Daniel L Rubin, et al. 2020. Cross-modal data programming enables rapid medical machine learning. *Patterns* 1 (2020), 100019.

[12] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. 2019. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery* 33, 4 (2019), 917–963.

[13] Theodoros Giannakopoulos. 2015. pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis. *PloS one* 10 (2015), 1–17.

[14] Mononito Goswami, Lujie Chen, and Artur Dubrawski. 2020. Discriminating Cognitive Disequilibrium and Flow in Problem Solving: A Semi-Supervised Approach Using Involuntary Dynamic Behavioral Signals. *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (2020), 420–427.

[15] Mononito Goswami, Minkush Manuja, and Maitree Leekha. 2020. Towards Social & Engaging Peer Learning: Predicting Backchanneling and Disengagement in Children. arXiv:2007.11346 [cs.HC]

[16] K. Hara, K. Inoue, K. Takanashi, and Tatsuya Kawahara. 2018. Prediction of Turn-taking Using Multitask Learning with Prediction of Backchannels and Fillers. *INTERSPEECH* 2 (2018), 991–995.

[17] Bettina Heinz. 2003. Backchannel responses as strategic responses in bilingual speakers' conversations. *Journal of pragmatics* 35, 7 (2003), 1113–1142.

[18] Bettina Heinz. 2003. Backchannel responses as strategic responses in bilingual speakers' conversations. *Journal of Pragmatics* 35, 7 (2003), 1113 – 1142. https://doi.org/10.1016/S0378-2166(02)00190-X

[19] Réka Hollandi, Ákos Diósdi, Gábor Hollandi, Nikita Moshkov, and Péter Horváth. 2020. AnnotatorJ: an ImageJ plugin to ease hand-annotation of cellular compartments. *Molecular Biology of the Cell* 31 (2020), mbc–E20.

[20] Paul Hömke, Judith Holler, and Stephen C Levinson. 2018. Eye blinks are perceived as communicative signals in human face-to-face interaction. *PloS one* 13, 12 (2018), e0208030.

[21] Danfeng Hong, Naoto Yokoya, Gui-Song Xia, Jocelyn Chanussot, and Xiao Xiang Zhu. 2020. X-ModalNet: A semi-supervised deep cross-modal network for classification of remote sensing data. *ISPRS Journal of Photogrammetry and Remote Sensing* 167 (2020), 12–23.

[22] Rajni Jindal, Maitree Leekha, Minkush Manuja, and Mononito Goswami. 2020. What makes a Better Companion? Towards Social & Engaging Peer Learning. *Proceedings of the Twenty-forth European Conference on Artificial Intelligence. IOS Press* 325 of Frontiers in Artificial Intelligence and Applications (2020), 482–489.

[23] Daniel Jurafsky, Carol Van Ess-dykema, et al. 1997. Switchboard discourse language modeling project.

[24] S. N. Khan, M. Leekha, J. Shukla, and R. R. Shah. 2020. Vyaktitv: A Multimodal Peer-to-Peer Hindi Conversations based Dataset for Personality Assessment. *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)* 1 (2020), 103–111. https://doi.org/10.1109/BigMM50055.2020.00024

[25] Maitree Leekha, Mononito Goswami, Rajiv Ratn Shah, Yifang Yin, and Roger Zimmermann. 2019. Are You Paying Attention? Detecting Distracted Driving in Real-Time. *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)* 1 (2019), 171–180.

[26] Pierre Lison, Aliaksandr Hubin, Jeremy Barnes, and Samia Touileb. 2020. Named Entity Recognition without Labelled Data: A Weak Supervision Approach. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* 1 (2020), 1518–1533.

[27] Louis-Philippe Morency, Iwan Kok, and Jonathan Gratch. 2010. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems* 20 (01 2010), 70–84. https://doi.org/10.1007/s10458-009-9092-y

[28] Markus Mueller, David Leuschner, Lars Briem, Maria Schmidt, Kevin Kilgour, Sebastian Stueker, and Alex Waibel. 2015. Using Neural Networks for Data-Driven Backchannel Prediction: A Survey on Input Features and Training Techniques. In *Human-Computer Interaction: Interaction Technologies*, Masaaki Kurosu (Ed.). Springer International Publishing, Cham, 329–340.

[29] Hae Won Park, Mirko Gelsomini, Jin Lee, and Cynthia Breazeal. 2017. Telling Stories to Robots: The Effect of Backchanneling on a Child's Storytelling. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. Association for Computing Machinery, New York, NY, USA, 100–108. https://doi.org/10.1145/2909824.3020245

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[31] Michael Rubinstein, Ce Liu, and William T Freeman. 2012. Annotation propagation in large image databases via dense image correspondence. In *European Conference on Computer Vision*. Springer Berlin Heidelberg, Berlin, Heidelberg, 85–99.

[32] Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. 2019. *Yeah, Right, Uh-Huh: A Deep Learning Backchannel Predictor*. Springer International Publishing, Cham, 247–258. https://doi.org/10.1007/978-3-319-92108-2_25

[33] Nikhita Singh, Jin Joo Lee, Ishaan Grover, and Cynthia Breazeal. 2018. P2PSTORY: Dataset of Children as Storytellers and Listeners in Peer-to-Peer Interactions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, United States, 1–11.

[34] T. Solorio, O. Fuentes, Nigel G. Ward, and Yaffa Al Bayyari. 2006. Prosodic feature generation for back-channel prediction. *INTERSPEECH* 5 (2006), 2398–2401.

[35] Leili Tavabi, Kalin Stefanov, Setareh Nasihati Gilani, David Traum, and Mohammad Soleymani. 2019. Multimodal Learning for Identifying Opportunities for Empathetic Responses. In *2019 International Conference on Multimodal Interaction*. Association for Computing Machinery, New York, NY, USA, 95–104. https://doi.org/10.1145/3340555.3353750

[36] Khiet P Truong, Ronald Poppe, and Dirk Heylen. 2010. A rule-based backchannel prediction model using pitch and pause information. *Eleventh Annual Conference of the International Speech Communication Association* 4 (2010), 3058–3061.

[37] Jesper E Van Engelen and Holger H Hoos. 2020. A survey on semi-supervised learning. *Machine Learning* 109, 2 (2020), 373–440.

[38] N. Ward. 1996. Using prosodic clues to decide when to produce back-channel utterances. *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96* 3 (1996), 1728–1731.

[39] Nigel Ward and Wataru Tsukahara. 2000. Prosodic features which cue back-channel responses in English and Japanese. *Journal of pragmatics* 32, 8 (2000), 1177–1207.

[40] Nigel G Ward and Yaffa Al Bayyari. 2006. A case study in the identification of prosodic cues to turn-taking: Back-channeling in Arabic. *Ninth International Conference on Spoken Language Processing* 5 (2006), 2018–2021.

[41] Sheida White. 1989. Backchannels across cultures: A study of Americans and Japanese. *Language in society* 18, 1 (1989), 59–76.

[42] Sheida White. 1989. Backchannels across Cultures: A Study of Americans and Japanese. *Language in Society* 18, 1 (1989), 59–76. http://www.jstor.org/stable/4168001

Fabrication of GaN Nano-Towers based Self-Powered UV Photodetector

Lalit Goswami[1,2], Neha Aggarwal[1], Pargam Vashishtha[1,†], Shubhendra K. Jain[1,3,†], Shruti Nirantar[3], Jahangeer Ahmed[4], M. A. Majeed Khan[5], Rajeshwari Pandey[2] and Govind Gupta[1,†,*]

[1]*CSIR-National Physical Laboratory, Dr K.S. Krishnan Road, New Delhi- 110012, India.*
[2]*Department of Electronics & Communication Engineering, Delhi Technological University, New Delhi-110042.*
[3]*Functional Materials and Microsystems Research Group and the Micro Nano Research Facility, RMIT University, Melbourne, VIC 3000, Australia.*
[4] *Department of Chemistry, College of Science, King Saud University, Riyadh 11451, Saudi Arabia*
[5]*King Abdullah Institute for Nanotechnology, King Saud University, Riyadh, 11451, Saudi Arabia*
[†]*Academy of Scientific & Innovative Research, CSIR-HRDC Campus, Ghaziabad, Uttar Pradesh - 201002, India.*
[*]Corresponding Author: govind.npl@nic.in; Telephone: +91-1145609503

Abstract

Fabrication of unique tapper ended GaN-Nanotowers structure based highly efficient ultraviolet photodetector is demonstrated. Hexagonal stacked GaN nanocolumnar structure (nanotowers) grown on AlN buffer layer exhibits higher photocurrent generation which significantly enhances its responsiveness towards ultraviolet illumination and leads to outstanding performance of the device. The fabricated detector display low dark current (~12nA), high I_{Light} / I_{Dark} ratio (>10^4), fast time-correlated transient response (~433µs) upon UV (325nm) illumination. A very high photo responsivity of 2.47 A/W in self-powered (zero applied bias) mode of operation is reported. While in photoconductive mode, the responsivity is observed to be 35.4 A/W @ -3V alongwith very high external quantum efficiency (~10^4 %), lower noise equivalent power (~10^{-13} $WHz^{-1/2}$) and excellent UV-Vis selectivity. Nanotower structure with lower strain and dislocations as well as reduced trap states cumulatively contributed to augmented performance from the device. The utilization of these GaN-Nanotower structures can potentially be useful towards the fabrication of energy-efficient ultraviolet photodetector.

Keywords: Gallium Nitride, Nanotowers, Ultraviolet, Photodetector

Research Square

Preprints are preliminary reports that have not undergone peer review. They should not be considered conclusive, used to inform clinical practice, or referenced by the media as validated information.

Introduction

Gallium nitride (GaN) ultraviolet (UV) photodetectors (PDs) attracted a lot of attention due to their versatility and ability to serve in an extreme environment. Thereby, GaN-material inherits the detectors with high carrier mobility, high saturation velocity, high-temperature sustainability, high breakdown voltages, inertness and hardness towards chemical and radiation respectively [1]. With various advantageous properties, GaN UV-PD can establish its position in numerous applications such as monitoring of combustion, flame detection, ozone detection, space communication, biological sensors: phototherapy, sterilization control, solar irradiance measurement, industrial monitoring, missile plume detection and ultraviolet astronomy etc.[2] So far various UV PD structures such as p-π-n PDs,[3] p-i-n photodiodes,[4] p-n junction photodiodes,[5] Schottky-barrier photodiodes,[6, 7] and metal-semiconductor-metal (MSM) PDs,[8, 9] have been fabricated. The ease of fabrication and rapid time-correlated response ascertained the advantage of Schottky barrier type and MSM structured UV-PDs as compared to the p-n junction. Recent reports also suggest that extremely low dark currents is achieved with MSM based UV PDs.[10] Since, the reduction in dark current is affected by low defect density, thus device performance can directly be correlated with defects generated during the growth of device structure. Simultaneously, these defects induce stress-strain which are accredited to lattice mismatch, and thermal expansion coefficient difference between GaN and the substrate.[11]

Further, in the process of PDs improvisation, nanostructures (NS) have played a significant role due to very low stress-strain induced electrical as well as optical defects that reciprocate its significant impact on detector performance.[12, 13] Moreover, NSs were also recognized as a structure with high surface to volume ratio i.e., higher aspect ratio. The increased surface area in NS provides more absorption sites to incident photons for generating high density of carriers, and lower resistance path which can cumulatively increase the detection capability of the PDs. Therefore, growth of high surface-to-volume ratio and stress relaxed 3D GaN-NSs play a crucial role in the development of highly efficient optoelectronic devices. Recently, it has been reported that GaN-nanocolumnar structure can offers wavelength tunability, low dislocation density and lower lattice related strain and very good response.[14-16] However, the fabricated devices are non-operational in photovoltaic mode i.e. zero applied bias and display very high response time.

In this letter, we report the utilization of AlN buffer layer (800ºC) for growth of lower strain and stress relaxed, higher aspect ratio hexagonal stacked GaN- nanocolumnar structure i.e. nanotowers (GaN-NTs) on Si (111) substrate via PAMBE. The impact of buffer on the quality of film is discussed along with the fabrication of UV photodetector. The performance of the fabricated UV-PD is analysed based on the performance parameters such as responsivity, detectivity, noise equivalent power and external quantum efficiency under UV illumination. Besides, the spectral selectivity of the fabricated UV-PD for the wavelength ranging from 200nm to 800nm under variable applied bias has also been assessed.

Experimental

A unique geometry (as shown in figure 1) i.e. GaN-NT arrays have been grown epitaxially via. plasma assisted molecular beam epitaxy (Compact 21 Riber) system. After standard wet chemical etching and thermal cleaning procedure [17] of Si (111) substrate (figure 1(a)), the initial growth of AlN buffer layer has been performed at 800°C for 30 min with Al beam equivalent pressure (BEP) ~6.8×10^{-7} torr and growth rate of 3.2 A^{o}/sec (shown in figure 1(b)). Afterwards, a wetting layer of Gallium (Ga) has been deposited for 10 sec at a lower substrate temperature of 666°C that precedes the nucleation process of GaN growth. Through heat energy gained by providing substrate temperature, Ga molecules and N_2 ion active species acquire enough energy at the substrate surface to forms GaN. The growth of GaN-NTs has been performed at 666 °C substrate temperature for 210 min with Ga BEP of 2.8×10^{-6} torr. To fulfil the essential requirement for nanocolumnar GaN growth, nitrogen-rich condition (III/V ratio <1) was maintained by adjusting N_2 flow up to 3 sccm with 400 W RF-plasma power with growth rate 3.2 A^{o}/sec. Thereafter, the growth conditions of closely packed GaN-NCs (shown in figure 1(c)) has been altered to the tapered ended nanotower-like structure by gradually reducing the Ga-flux from 2.8×10^{-6} to 1.5×10^{-6} torr for next 600 sec (shown in figure 1(d-e))

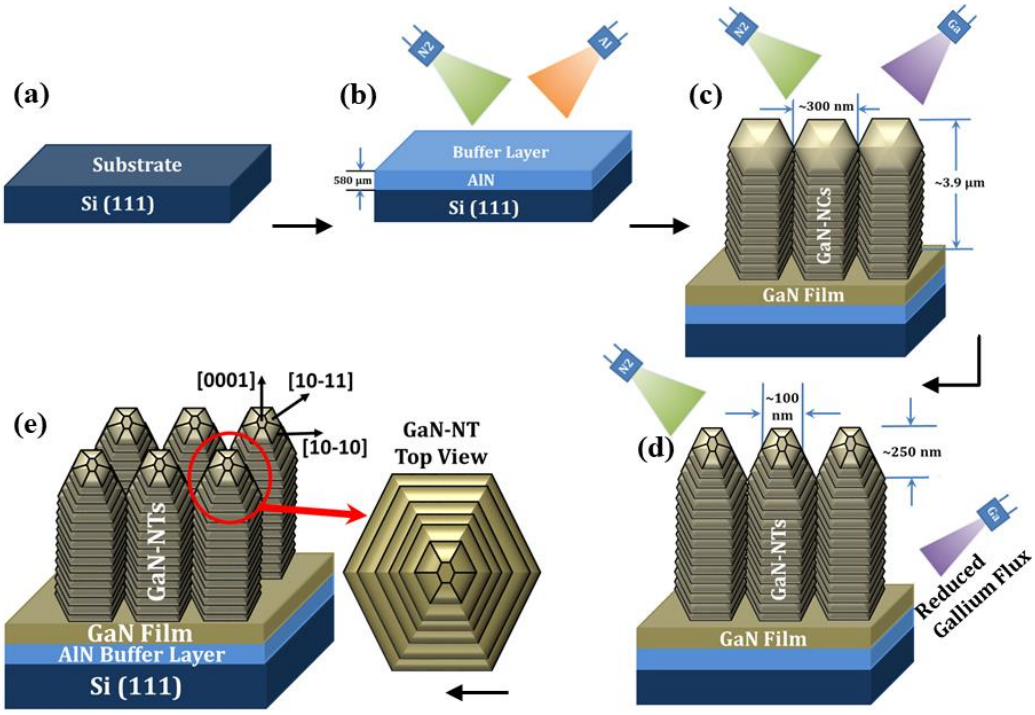


Figure 1 (a-e) A schematic growth diagram of grown GaN NTs on Si (111) substrate with sandwitched AlN buffer layer in between.

The real-time monitoring of GaN-NTs/Si (111) growth has been performed by reflection high electron diffraction (RHEED) using in-situ STAIB electron gun (12 keV). The morphological studies of grown GaN-NTs have been carried out by using Field-emission scanning electron microscopy (FEI Verios 460L FEGSEM) and Atomic force microscopy (AFM) (Multimode-V Veeco). The structural quality of the grown structure was examined by High-Resolution X-ray Diffraction (HRXRD) (PANalytical X'Pert PRO MRD System) using CuKα1 radiation (λ= 1.5406 A˚). For analysing the optical properties, room temperature (RT) photoluminescence (PL) measurement was carried out by using FLS980 D2D2 (Edinburg) system with 325nm excitation source (Hd-Cd Kimmon laser) and Raman spectroscopy was carried out in backscattering configuration, using a 514 nm argon-ion laser. The grown GaN-NTs were utilized to fabricate UV-PD in MSM structure by depositing metal electrodes of Au (~200nm)/Cr (~10nm) using thermal evaporation system. Photoresponse measurements were carried out using a probe station setup (S10 Triax probe station with Keithley 2401 acquisition unit) on the fabricated device with 1.18×10^{-3} cm^2 active area. The spectrometer is equipped with a focused laser source (λ= 325 nm with the utilized power density of 2.5 mW/cm^2 to 184 mW/cm^2).

Results and Discussion

The growth of GaN-Nanotower structure is carried out on reconstructed Si (111) 7x7 and in-situ monitored by RHEED technique (figure S1, supplementary information).The crystallinity of grown structure was evaluated by HRXRD 2θ-ω scan and its peak positions (figure S2; supplementary information) were utilized to calculate the strain via. lattice constant estimation using Bragg's law.[18] The calculated value of strain in the grown structure is realized to be 0.25% which confirms strain relaxation in the grown nanotowers. Further, to analyse the dislocation density, HRXRD omega scans along (0002) and (10-12) plane of diffraction have been performed as shown in figure 2(a). The FWHM values of omega scan along (0002) and (10-12) were found to be 0.39° and 0.84° attributing to screw dislocation density ~1.91×10^{9} cm^{-2} and edge dislocation density ~ 2.33×10^{10} cm^{-2}, respectively. These dislocation densities are found to be lower in comparison to the GaN NTs structure grown on Si (111) substrate with thicker AlN buffer layer (665 nm) at lower growth temperature (762^{o}C).[16]

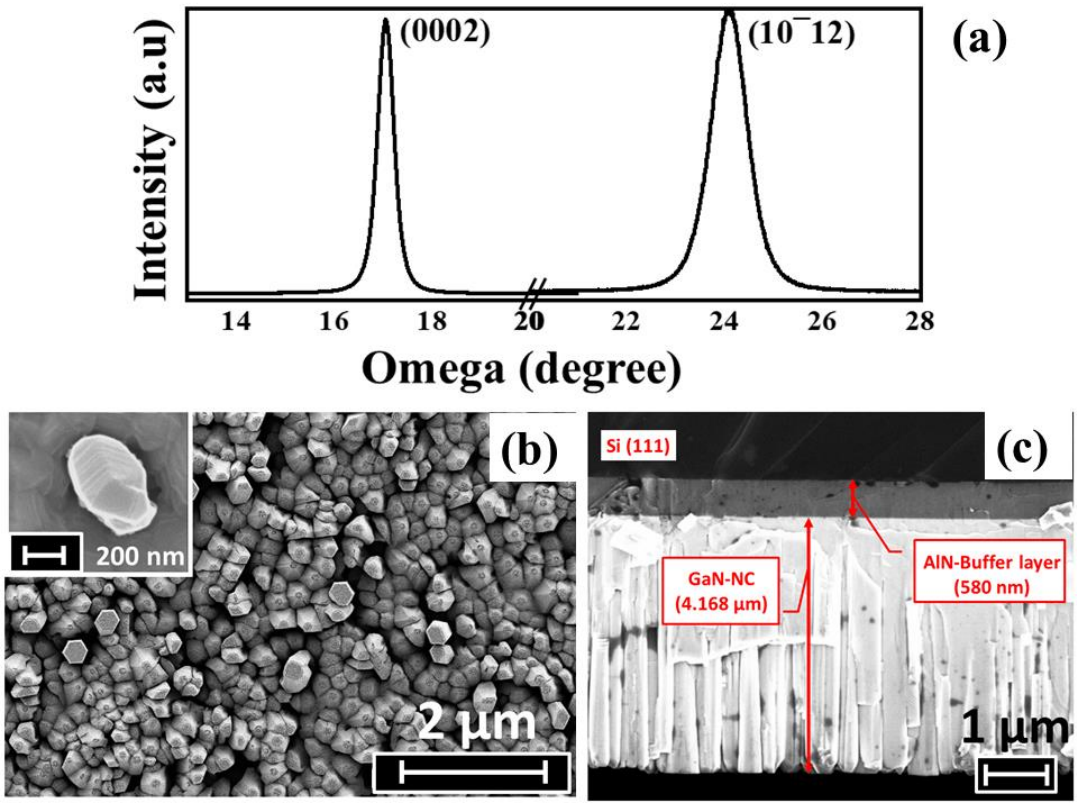


Figure 2 (a) HRXRD omega scan along (0002) & (10-12) plane of diffraction, (b) Planar view FESEM image (inset: shows a single NT), (c) Cross-sectional FESEM image of GaN NTs.

The surface morphology of the GaN-nanotowers has been probed by FESEM (as shown in figure 2(b)) where the planer view of high density closely packed GaN-NTs with hexagonal shapes nanocolumns was shown with tapered ended geometry. Inset of figure 2(b) demonstrates a high magnification image of layer by layer growth of single GaN-NT. Figure 2(c) displayed the cross -sectional view of the novel hexagonal GaN-NTs having length of ~4.168 µm grown along the direction perpendicular to the substrate with AlN-buffer layer (~580nm) sandwiched between them. The notable layered structure is having stacking of uniform hexagonal GaN layers up to the length of ~4µm. Afterward, lowering of Ga flux resulted in a gradual decrease in the surface areas of stacking of GaN layers which help in formation of unique nanotowers like structure ended with a tip of width: ~100nm. The distribution density of grown NTs is calculated to be 3.5×10^{10} cm^{-2}. It was also observed that the high quality AlN buffer layer grown at higher substrate temperature (800^{o}C) resulted in better structural and morphological properties in comparison to the structure grown at AlN buffer grown at lower substrate temperature[19, 20] Besides this, the increment in surface area is evaluated in concurrence with FESEM & AFM measurements which confirm the enhancement from 9 μm^2 to 15.6 μm^2 from planar to tapper ended GaN-NTs morphology.

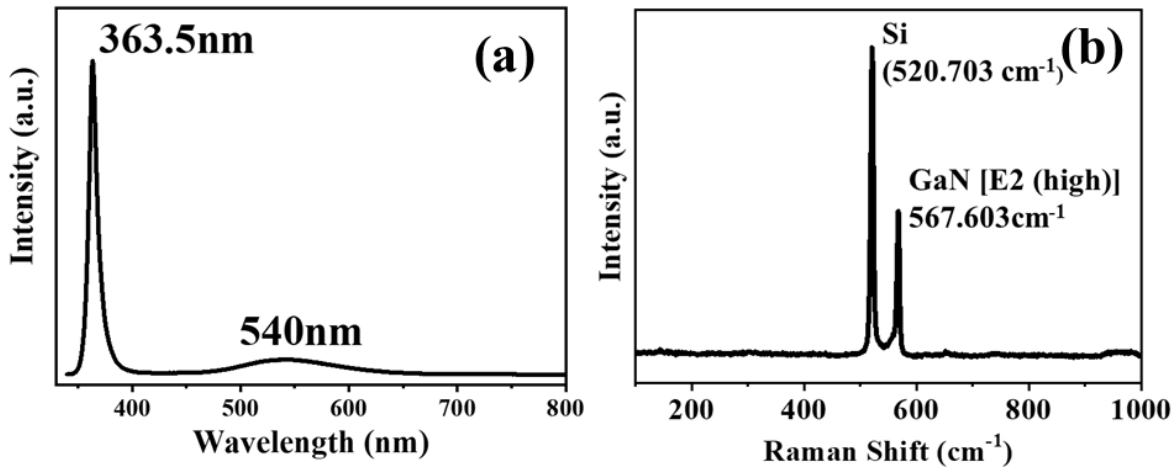


Figure 3 (a) RT-PL spectra and (b) Raman spectra of as-grown GaN-NTs.

The grown GaN-NTs with enhanced surface area consequently allow more photon absorption sites to generate a large number of photo-excited electron-hole pairs (EHPs). These EHPs emit high-intensity photoluminescence upon recombination which was reflected by intense near band edge emission (NBE) peak at 363.5nm in RT-PL spectra shown in Figure 3(a). The NBE peak is found to be in close agreement with bulk GaN PL emission (~364nm i.e, 3.406eV).[21] The FWHM value of NBE peak is found to be 8.51nm. The sharp NBE peak

with low FWHM value endorse the good crystalline quality of the grown GaN-NTs.[22] From PL spectra, an additional peak at 540 nm is also observed which is attributed to yellow band emission (YBE) originated due to the generated point defects during the growth.[23, 24] Furthermore, Raman scattering measurement was employed to quantify the stress in the grown structure as shown in figure 3(b). The Raman spectra exhibits, E_2-high mode of grown GaN-NTs located at 567.603 cm^{-1} which is shifted by ~0.397 cm^{-1} as compared to the E_2 (high) phonon mode of bulk GaN (568 cm^{-1}).[21] The observed red-shift in the E_2 (high) position specifies of tensile stress in the grown GaN-NTs.[25, 26] The value of tensile stress[18] is estimated to be 0.092 GPa which is significantly lower as compared to 3D GaN-NSs and 2D GaN films[27-29] which can be attributed to lower defect states. Further, it has been proposed that these defects can have an adverse impact on electrical as well as the optical property of the GaN-based UV PD.[30] So, the lower defects attributes to lower trap states, hence the grown GaN-NTs having lower defect density can significant be useful for fabrication of efficient optoelectronic devices.

Figure 4(a) displays RT current-voltage (I-V) characteristic of fabricated GaN-NTs based UV-PD under dark as well as 325nm UV-laser illumination (3mW optical power) with applied bias voltage sweep of ±3V. The schematic image of the fabricated GaN-NT array based device is shown in the figure 4(b). The difference in dark and light I-V characteristics of the fabricated device leads to photodetection aspect. As revealed by I-V-plots, the self-channelization of light current [inset figure 4(a)] without applied bias voltage (zero bias) viz. photovoltaic mode designates the fabricated detector as self-powered UV-PD. The presence of non-identical end tips of the grown GaN-NT structure lead to form non-homogeneous Au/GaN interfaces between Au-metal electrodes and GaN-NTs which exhibits two Schottky natured electrodes with different barrier heights. The variation in barrier heights creates built-in potential gradient and powered the device to respond even at zero bias voltage.[31]

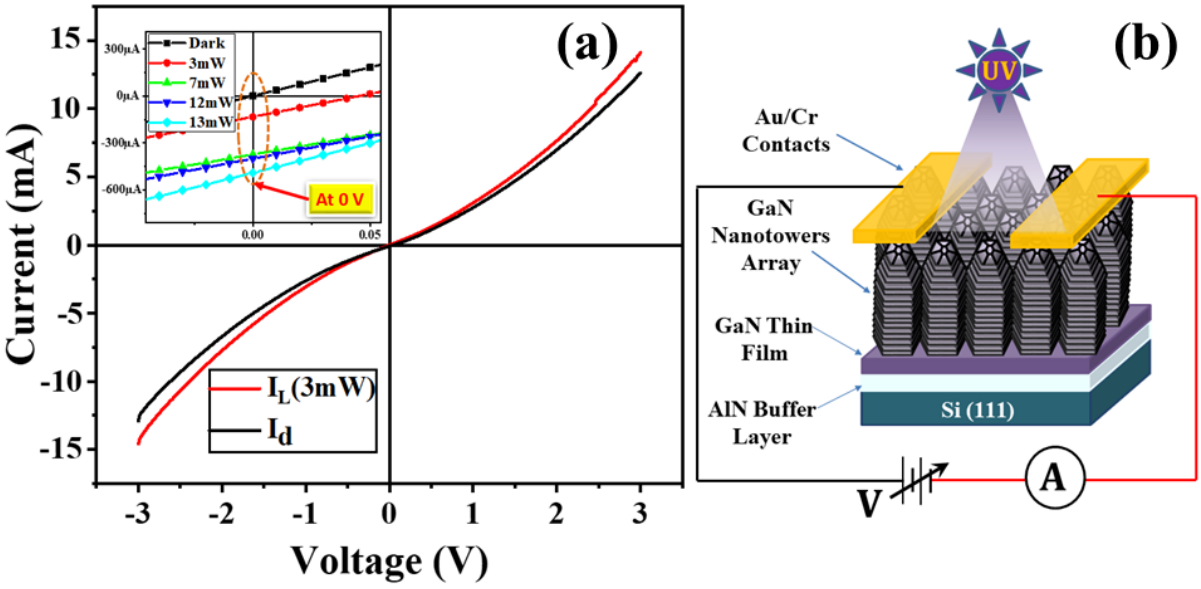


Figure 4. (a)RT I-V characteristic of fabricated GaN-NTs based UV-PD under dark and 325nm UV illumination with 3mW optical power. Inset: indicate self-powered characteristic at various optical power and bottom right indicate schematic diagram of the fabricated device, (b) Schematic image of the fabricated GaN-NT array based device.

Initially, the transient photo-response behaviour of the fabricated detector has been realized in photovoltaic mode with the variable incident optical power (from 3mW to 13mW) as shown in figure 5(a). The transient photo-response measurement has been performed by measuring inflow current with periodically switching ON (I_{Light}) and OFF (I_{Dark}) the UV laser with variable optical power. The maximum value of I_{Light} / I_{Dark} ratio was recorded to be $>10^4$ under 13mW optical power. The analysis demonstrates a very low dark current ~12nA, wherein the fabricated UV-PD shows a stable and significant enhancement in light current value from 67.5µA to 545µA as function of optical power varying from 3mW to 13mW. Further to explore the capability of the detector to detect low power signal, the optical power is reduced from 0.7mW to 180µW correspondingly the light current values were observed to be varying from 3.4µA to 70.8nA (Figure S3, supplementary information). Conclusively, the fabricated GaN-NTs based UV-PD has shown remarkable enhancement of ~7700× in the photocurrent value from 70.8nA to 545 µA by varying optical power from 180µW to13mW in the photovoltaic mode (self-power) of operation.

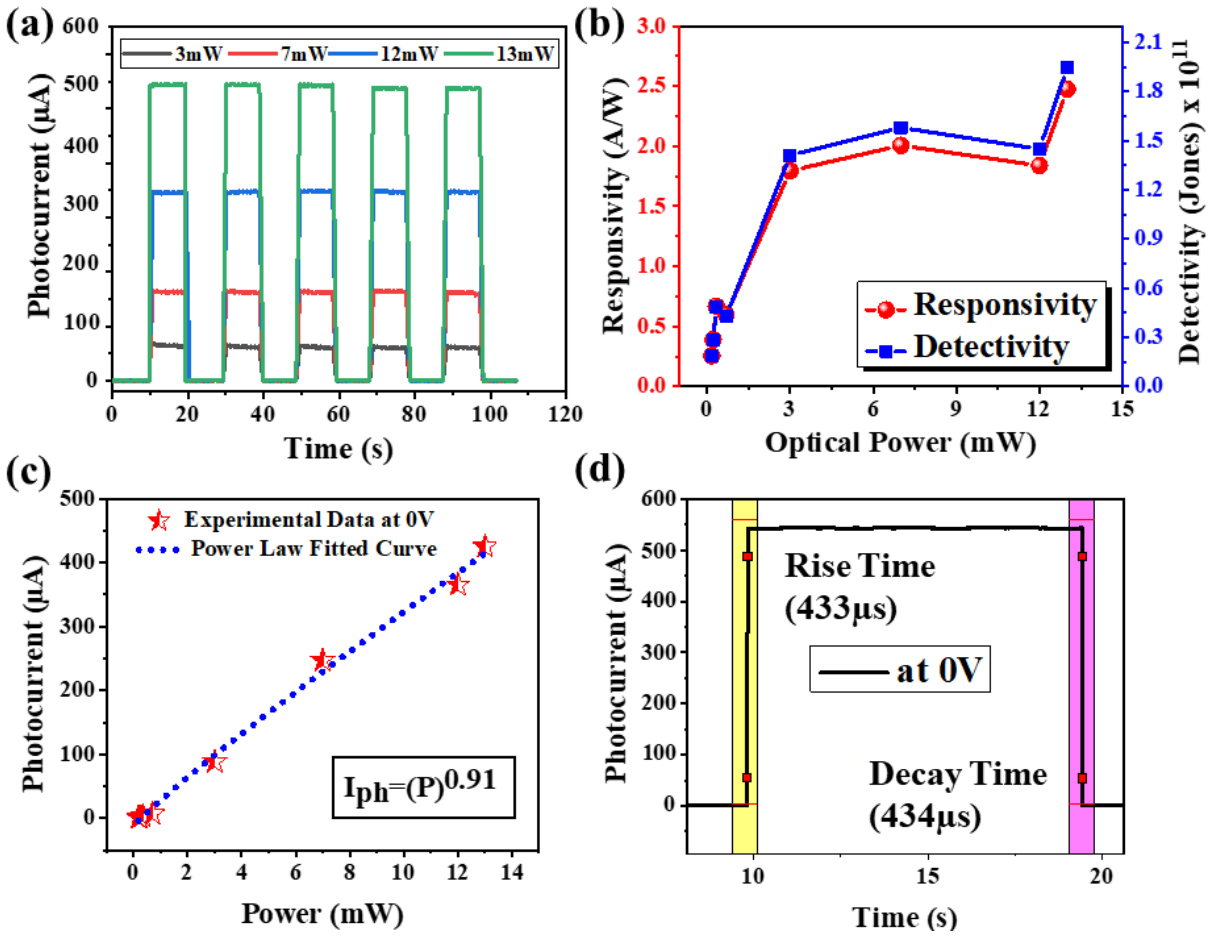


Figure 5 Incident optical power-dependent performance of GaN-NTs UV-PD in photovoltaic mode (a) time-correlated transient photoresponse, (b) Correlation of responsivity and detectivity with varied optical power, (c) Power-law fitted curve in the experimental data of photocurrent, (d) Reversibility (Response time) of the detector under constant 13mW optical power illumination.

In addition the performance of the fabricated device has also been investigated based on their well-known performance parameters like responsivity (R), noise equivalent power (NEP), detectivity (D) and external quantum efficiency (Q.E: η) and reversibility. using the following equations [32] :

$$R = \frac{I_{ph}}{P_{inc} \times A_0} \quad \text{... (1)}$$

$$NEP = \frac{\sqrt{2eI_d}}{R} \quad \text{... (2)}$$

$$D = \frac{R \times \sqrt{A_0}}{NEP} \quad \text{... (3)}$$

$$\eta = \frac{R \times h \times c}{e \times \lambda} \quad \text{... (4)}$$

Where P_{inc} stands for incident UV photons optical power illuminated on the detector active (A_o), I_{ph} and I_d is the measured photocurrent and dark current, h is the Planck's constant, c is the speed of light, e is elementary charge and λ is incident UV photon wavelength. Figure 5(b) displays that the fabricated GaN-based detector yielded a gradual increment in responsivity from 0.255 A/W to 2.47 A/W and detectivity from 1.8×10^{10} jones to 1.95×10^{11} jones with increasing incident optical power from 180μW to 13mW. Remarkably, the determined responsivity and detectivity is found to be significantly very high compared to the existing bare GaN-based UV-PDs operated in photovoltaic or self-power mode of operation. Table 1 consolidates the comparison of performance parameters of other GaN based UV-PD operable in self-powered mode of operation. The analysis also divulges a linear relationship between photocurrent and incident optical power as demonstrated in figure 5(c). The power law has been fitted on the experimental data, $I_{Ph} \alpha P^{\theta}$, where I_{Ph} is the photocurrent and θ determines the photocurrent response w.r.t. optical power density. The θ value is determined to be 0.91, which is close to unity reflecting that the carrier trap states between the conduction band edge and the fermi level are very low.[33] Moreover, to test the reversibility of the detectors, the time-correlated transient photoresponse of the fabricated GaN-based UV-PD has been analysed. The rise time (as measured from 10% to 90%) and decay time (from 90% to 10%) under UV on/off condition was measured to be ~433μs and ~434μs, respectively as shown in figure 5(d). Table 1 also demonstrates a rapid rise/decay time from the fabricated GaN-NTs/AlN/Si heterostructure based UV PD as compared to other reported bare GaN-based PDs. The GaN NTs with lower AlN buffer layer thickness and low dislocation density leads to reduced resistance path for charge transport resulting in efficient photovoltaic operation.[34]

Table 1: Comparison of performance parameters of GaN-based UV-PDs.

Sample	R (mA/W) at 0V	R (mA/W) at applied bias	Rise / Decay time	QE%	Ref.
GaN-NT/Si(111)	2.47× 10^3	3.54× 10^4, -3V	433µs / 434µs	1.35×10^4	**This work**
GaN-NT/Si(111)	----	8.86× 10^4, -3V	1.25s/ 1.15ms	3×10^4	[19]
GaN-NW/n-Si(111)	----	19, 5V	2.8s/3.7s	6.7	[29]
GaN/Si(111)	----	282, 15V	----	96.71	[27]
GaN-NW/n-Si(100)	----	68, 5V	2.4s/3.1s	24.4	[29]
GaN-MWA/p-Si	131	----	2ms / 2ms	49.98	[35]
GaN/Al2O3	104	----		36	[36]
GaN/Al2O3	147	----	0.11ms/0.12ms	50.7	[37]
GaN p-i-n/pattern Al2O3	190	210, -5V	----	65	[38]
GaN/Al2O3	300	----	<0.1s	----	[39]

To analyse the impact of applied bias on the performance of fabricated GaN-based UV-PD, time-correlated transient response of the detector at a fixed optical power of 3mW and variable bias voltage is acquired as shown in figure 6. The investigation revealed an increment in photocurrent values with increasing applied bias i.e., 66.6µA (0V), -0.35mA (-1V), -1.025mA (-2V) and -1.73mA (-3V) which is ascribed to enhanced thermionic field emissions. The increment in the photocurrent with respect to applied bias voltage is also reflected in the performance parameters of the fabricated UV-PD. Figure 7(a) displays the increasing trend of responsivity of the device with an increasing value of applied bias voltage from 0V to -3V. The attained responsivity values are 1.79 A/W and 35.4 A/W at 0V and -3V applied bias respectively by exposing the device using 325nm UV source at 3mW optical power. The responsivity value of 35.4 A/W is significantly higher as compared to other bare GaN-based UV-PD as tabulated in Table 1. However, the recently reported GaN-NT based UV-PD [19] demonstrated higher responsivity of 88.6 A/W at -3V, moreover the performance of the devices is limited to photo-conductive mode only with ~300x higher response time. In addition, the present analysis reports optimised AlN buffer with reduced defect density in GaN-NTs. Consequently, the fabricated detector displayed independence of operability in both photovoltaic and photoconductive modes with significantly higher performance with faster transient response. Besides that, another important parameter i.e., the detectivity of the fabricated device is also calculated to be 1.41×10^{11} jones at 0V as and 1.83×10^{10} jones at -3V which divulge that at applied bias, the high dark current reduce detectivity of the device. Furthermore, the quantum efficiency (QE) of the device which is

related to sensitivity and defined as the conversion rate of incident photons to generate charge carriers that contribute to current conduction have also been evaluated. Figure 7(b) demonstrates the impact of variable optical power and applied bias voltage on the quantum efficiency. In photovoltaic mode, the quantum efficiency was observed to be increased from 97.41% at 180µW optical power to 946.75% at 13mW optical power. Further, QE values have also been evaluated at a fixed optical power of 3mW and variable bias voltage (0V to -3V) condition where QE increased from 686.39% (at 0V) and reached up to 1.35×10^{4} % (at -3V). High value of quantum efficiency (> 100 %) reveals the existence of a photoconductive gain which is greater than unity.[39] Likewise, to measure the sensitivity of the fabricated detector, noise equivalent power (NEP) of the device has been determined.[40] The calculated value of NEP was 2.42×10^{-13} $WHz^{-1/2}$ (at 3mW, 0V) which elucidates that the fabricated GaN-NTs based UV-PD can detect a signal of power as low as ~ 200 femtowatt with a signal-to-noise ratio of one after the one-half second of averaging.

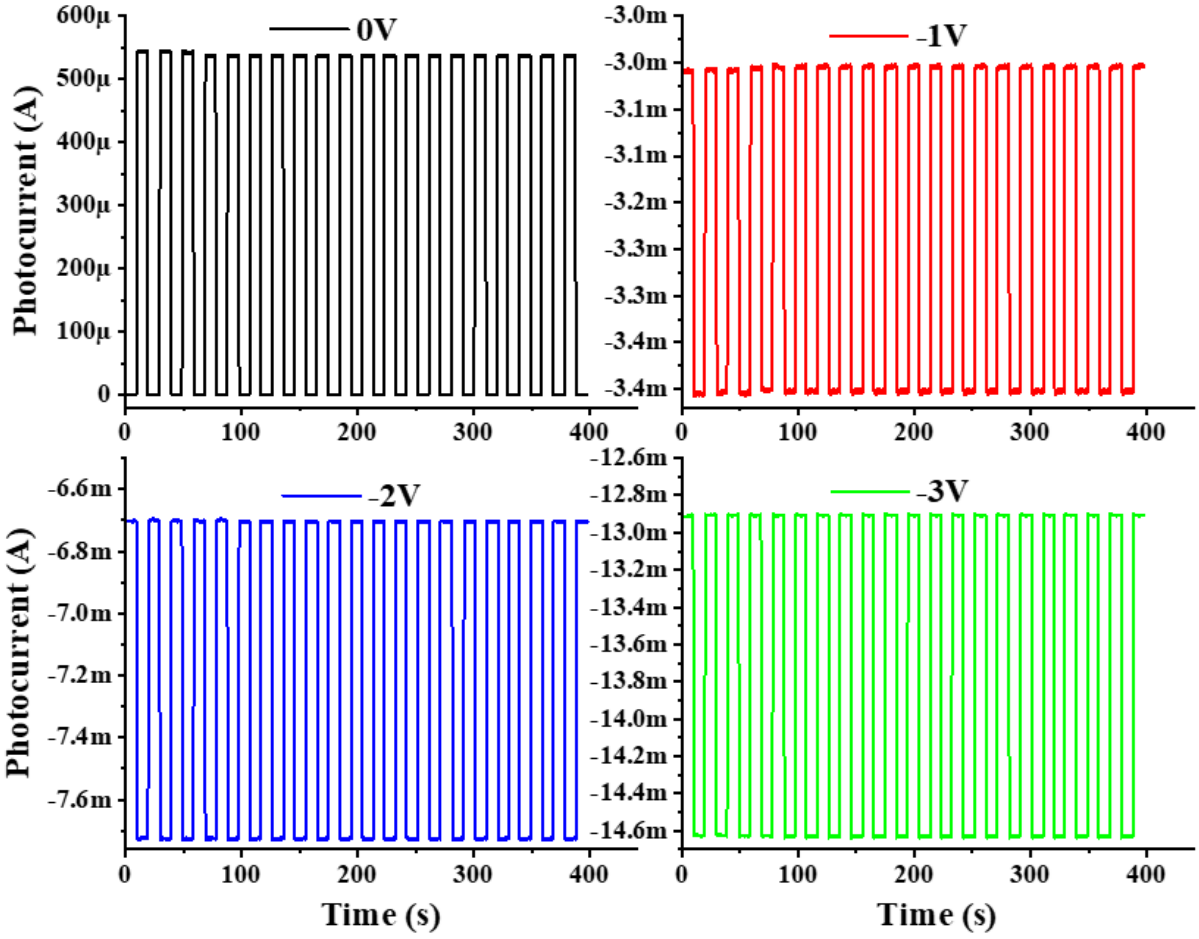


Figure 6. Bias dependent time-correlated transient photoresponse of GaN-NTs UV-PD at 3mW optical power illumination

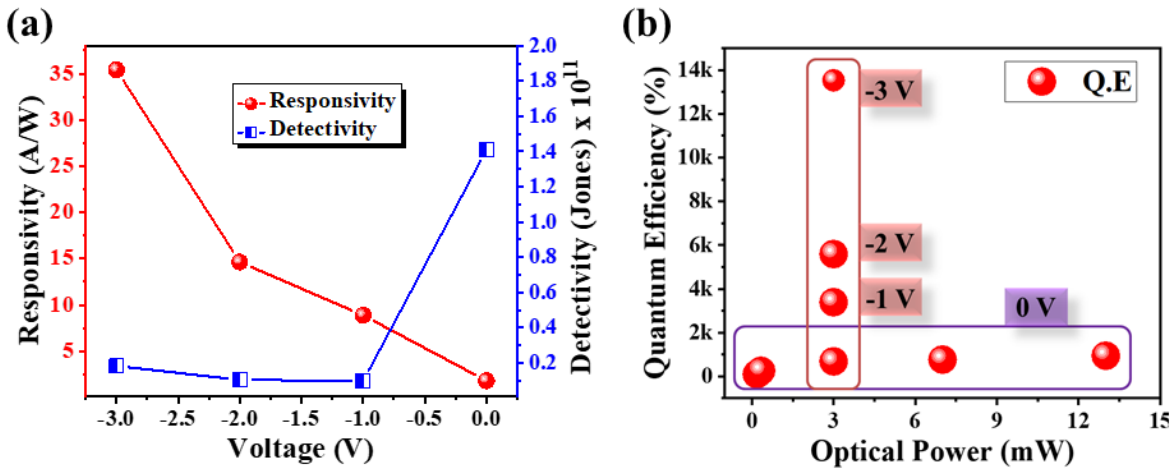


Figure 7 (a) Responsivity and detectivity graph w.r.t applied bias voltage, (b) Quantum efficiency (Q.E) of fabricated GaN-NTs UV-PD w.r.t incident optical power and bias voltage.

The spectral selectivity of the fabricated GaN-NTs based UV-PD has also been measured in the range of 200nm to 800nm at variable applied bias from -1V to -3V (figure 8). The obtained data demonstrates that the fabricated detector exhibits noteworthy UV light detection selectivity, which designate the detector to find its application in UV-A spectral range (320-400nm).

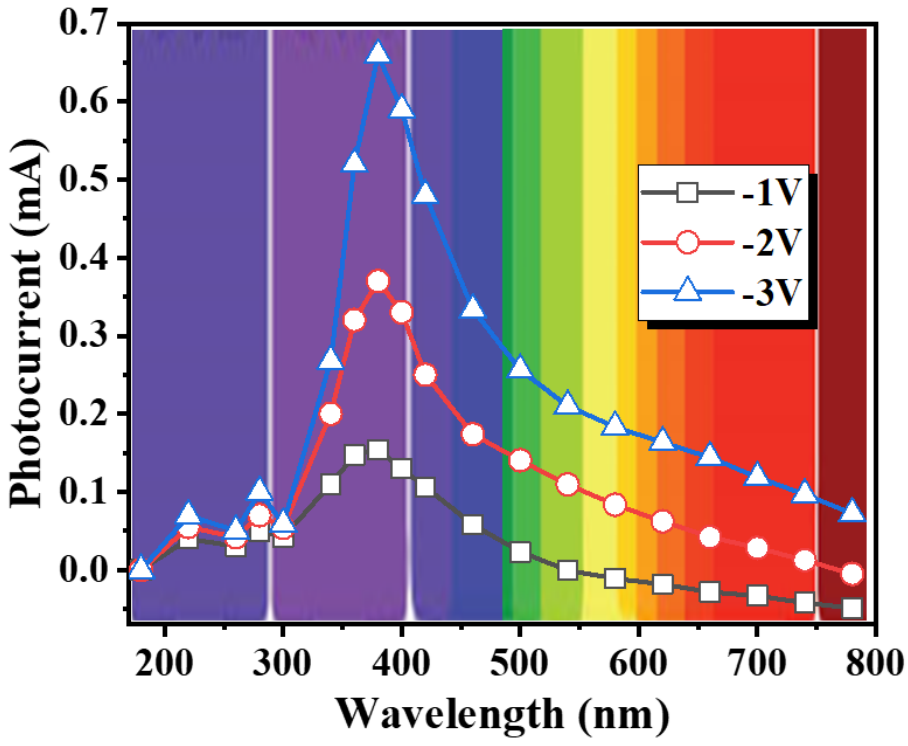


Figure 8: Spectral response characteristic of GaN-NTs based UV-PD from 200 to 800 nm under applied bias.

Conclusion

The fabrication of GaN-NT based high-performance and versatile UV-PD with independence of operability (in photovoltaic as well as a photoconductive mode) has been demonstrated where the detector yielded enhanced performance due to unique NTs tapered ended geometry. The closely packed NTs based layered structure formed by stacking of uniform hexagonal GaN layers offers the higher active surface area to the incident UV-photons. Thus, the high performance of fabricated detector using GaN NTs has been ascertained from reduce resistance path for charge transport. Thereby, in self-powered mode of operation, the fabricated UV photodetector showcased significantly higher I_{Light} / I_{Dark} ratio (>10^4), lower dark current (~12nA), remarkable enhancement in photocurrent (~7700 fold) as a function of applied optical power, highest responsivity (2.47 A/W), higher detectivity (1.95×10^{11} jones), quick response time (~433μs), very high QE (~947%) and detectability of signal as low as 200 femtowatt. Subsequently, the performance of the fabricated UV-PD has also been analysed at photoconductive mode of operation where significantly higher responsivity ~35.4 A/W at lower bias voltage of -3V is obtained. The fabricated device with efficient operation and fast transit of charge carriers can effectively be utilized for futuristic optoelectronic devices.

Acknowledgement

The authors gratefully acknowledge Director, CSIR-NPL, New Delhi, for his everlasting encouragement and support. One of the authors NA acknowledges CSIR for financial support under CSIR-Research Associateship. J Ahmed and MAM Khan extend their sincere appreciation to Deanship of Scientific Research at King Saud University for funding this Research Group (RG-1439-087).

Conflict of Interest: The authors declare that they have no conflict of interest.

Reference

[1] F. Medjdoub, *Gallium nitride (GaN): physics, devices, and technology*: CRC Press, 2017.

[2] E. Monroy, F. Omnès, and F. Calle, "Wide-bandgap semiconductor ultraviolet photodetectors," *Semiconductor Science and Technology,* vol. 18, p. R33, 2003.

[3] A. Osinsky, S. Gangopadhyay, R. Gaska, B. Williams, M. A. Khan, D. Kuksenkov, and H. Temkin, "Low noise p-π-n GaN ultraviolet photodetectors," *Applied Physics Letters,* vol. 71, pp. 2334-2336, 1997.

[4] D. Walker, A. Saxler, P. Kung, X. Zhang, M. Hamilton, J. Diaz, and M. Razeghi, "Visible blind GaN pin photodiodes," *Applied Physics Letters,* vol. 72, pp. 3303-3305, 1998.

[5] E. Monroy, E. Munoz, F. Sánchez, F. Calle, E. Calleja, B. Beaumont, P. Gibart, J. Munoz, and F. Cussó, "High-performance GaN pn junction photodetectors for solar ultraviolet applications," *Semiconductor Science and Technology*, vol. 13, p. 1042, 1998.
[6] M. Lee, J. Sheu, W. Lai, S.-J. Chang, Y.-K. Su, M. Chen, C. Kao, G.-C. Chi, and J. Tsai, "GaN Schottky barrier photodetectors with a low-temperature GaN cap layer," *Applied Physics Letters*, vol. 82, pp. 2913-2915, 2003.
[7] O. Katz, G. Bahir, and J. Salzman, "Persistent photocurrent and surface trapping in GaN Schottky ultraviolet detectors," *Applied Physics Letters*, vol. 84, pp. 4092-4094, 2004.
[8] D. Walker, E. Monroy, P. Kung, J. Wu, M. Hamilton, F. Sanchez, J. Diaz, and M. Razeghi, "High-speed, low-noise metal–semiconductor–metal ultraviolet photodetectors based on GaN," *Applied Physics Letters*, vol. 74, pp. 762-764, 1999.
[9] J. Carrano, T. Li, P. Grudowski, C. Eiting, R. Dupuis, and J. Campbell, "Comprehensive characterization of metal–semiconductor–metal ultraviolet photodetectors fabricated on single-crystal GaN," *Journal of Applied Physics*, vol. 83, pp. 6148-6160, 1998.
[10] S. Butun, M. Gökkavas, H. Yu, and E. Ozbay, "Low dark current metal-semiconductor-metal photodiodes based on semi-insulating GaN," *Applied Physics Letters*, vol. 89, p. 073503, 2006.
[11] S. Tomiya, K. Funato, T. Asatsuma, T. Hino, S. Kijima, T. Asano, and M. Ikeda, "Dependence of crystallographic tilt and defect distribution on mask material in epitaxial lateral overgrown GaN layers," *Applied Physics Letters*, vol. 77, pp. 636-638, 2000.
[12] P. Tchoulfian, F. Donatini, F. Levy, A. Dussaigne, P. Ferret, and J. Pernot, "Direct imaging of p–n junction in core–shell GaN wires," *Nano letters*, vol. 14, pp. 3491-3498, 2014.
[13] F. Qian, H. Wang, Y. Ling, G. Wang, M. P. Thelen, and Y. Li, "Photoenhanced electrochemical interaction between Shewanella and a hematite nanowire photoanode," *Nano letters*, vol. 14, pp. 3688-3693, 2014.
[14] M. Yoshizawa, A. Kikuchi, M. Mori, N. Fujita, and K. Kishino, "Growth of self-organized GaN nanostructures on Al2O3 (0001) by RF-radical source molecular beam epitaxy," *Japanese journal of applied physics*, vol. 36, p. L459, 1997.
[15] L. Goswami, N. Aggarwal, M. Singh, R. Verma, P. Vashishtha, S. K. Jain, J. Tawale, R. Pandey, and G. Gupta, "GaN Nanotowers Grown on Si (111) and Functionalized with Au Nanoparticles and ZnO Nanorods for Highly Responsive UV Photodetectors," *ACS Applied Nano Materials*, 2020.
[16] L. Goswami, N. Aggarwal, R. Verma, S. Bishnoi, S. Husale, R. Pandey, and G. Gupta, "Graphene quantum dots sensitized ZnO-Nanorods/GaN-Nanotowers heterostructure based high performance UV Photodetector," *ACS applied materials & interfaces*, 2020.
[17] N. Aggarwal, S. Krishna, A. Sharma, L. Goswami, D. Kumar, S. Husale, and G. Gupta, "A highly responsive self-driven UV photodetector using GaN nanoflowers," *Advanced Electronic Materials*, vol. 3, p. 1700036, 2017.
[18] N. Aggarwal, S. T. Krishna, L. Goswami, M. Mishra, G. Gupta, K. Maurya, S. Singh, N. Dilawar, and M. Kaur, "Extenuation of stress and defects in GaN films grown on a metal–organic chemical vapor deposition-GaN/c-sapphire substrate by plasma-assisted molecular beam epitaxy," *Crystal Growth & Design*, vol. 15, pp. 2144-2150, 2015.
[19] L. Goswami, N. Aggarwal, M. Singh, R. Verma, P. Vashishtha, S. K. Jain, J. Tawale, R. Pandey, and G. Gupta, "GaN Nanotowers Grown on Si (111) and Functionalized with Au Nanoparticles and ZnO Nanorods for Highly Responsive UV Photodetectors," *ACS Applied Nano Materials*, vol. 3, pp. 8104-8116, 2020.
[20] Y. Liu, X. Wang, L. Fu, and D. Zhu, "Tower-like structure of ZnO nanocolumns," *Chemical communications*, pp. 1304-1305, 2003.
[21] A. Denis, G. Goglio, and G. Demazeau, "Gallium nitride bulk crystal growth processes: A review," *Materials Science and Engineering: R: Reports*, vol. 50, pp. 167-194, 2006.
[22] Y. Inoue, T. Hoshino, S. Takeda, K. Ishino, A. Ishida, H. Fujiyasu, H. Kominami, H. Mimura, Y. Nakanishi, and S. Sakakibara, "Strong luminescence from dislocation-free GaN nanopillars," *Applied Physics Letters*, vol. 85, pp. 2340-2342, 2004.
[23] J. Hsu, F. F. Schrey, and H. Ng, "Spatial distribution of yellow luminescence related deep levels in GaN," *Applied Physics Letters*, vol. 83, pp. 4172-4174, 2003.

[24] G. You, W. Guo, C. Zhang, P. Bhattacharya, R. Henderson, and J. Xu, "Excitation dependent two-component spontaneous emission and ultrafast amplified spontaneous emission in dislocation-free InGaN nanowires," *Applied Physics Letters*, vol. 102, p. 091105, 2013.
[25] K. Jeganathan, R. Debnath, R. Meijers, T. Stoica, R. Calarco, D. Grützmacher, and H. Lüth, "Raman scattering of phonon-plasmon coupled modes in self-assembled GaN nanowires," *Journal of Applied Physics*, vol. 105, p. 123707, 2009.
[26] L. Zhang, K. Cheng, S. Degroote, M. Leys, M. Germain, and G. Borghs, "Strain effects in GaN epilayers grown on different substrates by metal organic vapor phase epitaxy," *Journal of Applied Physics*, vol. 108, p. 073522, 2010.
[27] L. Ravikiran, K. Radhakrishnan, N. Dharmarasu, M. Agrawal, Z. Wang, A. Bruno, C. Soci, T. Lihuang, and A. K. Siong, "Responsivity drop due to conductance modulation in GaN metal-semiconductor-metal Schottky based UV photodetectors on Si (111)," *Semiconductor Science and Technology*, vol. 31, p. 095003, 2016.
[28] T.-H. Chou, "GaN Thin Films Deposited on n-Si (111) Substrate with a Metal Organic Chemical Vapor Deposition System for Sensing Ultraviolet Light," *Sensors and Materials*, vol. 28, pp. 689-694, 2016.
[29] K. Saron, M. Hashim, N. Naderi, and N. K. Allam, "Enhanced light sensing characteristics of nanostructured gallium nitride/silicon heterojunctions: Interface matters," *Journal of Applied Physics*, vol. 114, p. 134510, 2013.
[30] M. Hanzaz, A. Bouhdada, P. Gibart, and F. Omnes, "Impact of the defects on the electrical and optical properties of AlGaN ultraviolet photodetectors," *Journal of Applied Physics*, vol. 92, pp. 13-18, 2002.
[31] L. Goswami, R. Pandey, and G. Gupta, "Ultra-thin GaN nanostructures based self-powered ultraviolet photodetector via non-homogeneous Au-GaN interfaces," *Optical Materials*, vol. 102, p. 109820, 2020.
[32] Z. Alaie, S. M. Nejad, and M. Yousefi, "Recent advances in ultraviolet photodetectors," *Materials Science in Semiconductor Processing*, vol. 29, pp. 16-55, 2015.
[33] S.-C. Kung, W. E. van der Veer, F. Yang, K. C. Donavan, and R. M. Penner, "20 μs photocurrent response from lithographically patterned nanocrystalline cadmium selenide nanowires," *Nano letters*, vol. 10, pp. 1481-1485, 2010.
[34] Y. Huang, X. Chen, S. Fung, C. Beling, and C. Ling, "Experimental study and modeling of the influence of screw dislocations on the performance of Au/n-GaN Schottky diodes," *Journal of Applied Physics*, vol. 94, pp. 5771-5775, 2003.
[35] W. Song, X. Wang, H. Chen, D. Guo, M. Qi, H. Wang, X. Luo, X. Luo, G. Li, and S. Li, "High-performance self-powered UV-Vis-NIR photodetectors based on horizontally aligned GaN microwire array/Si heterojunctions," *Journal of Materials Chemistry C*, vol. 5, pp. 11551-11558, 2017.
[36] X. Sun, D. Li, Z. Li, H. Song, H. Jiang, Y. Chen, G. Miao, and Z. Zhang, "High spectral response of self-driven GaN-based detectors by controlling the contact barrier height," *Scientific reports*, vol. 5, p. 16819, 2015.
[37] W. Mou, L. Zhao, L. Chen, D. Yan, H. Ma, G. Yang, and X. Gu, "GaN-based schottky barrier ultraviolet photodetectors with graded doping on patterned sapphire substrates," *Solid-State Electronics*, vol. 133, pp. 78-82, 2017.
[38] G. Wang, H. Lu, D. Chen, F. Ren, R. Zhang, and Y. Zheng, "High quantum efficiency GaN-based pin ultraviolet photodetectors prepared on patterned sapphire substrates," *IEEE Photonics Technology Letters*, vol. 25, pp. 652-654, 2013.
[39] M. Peng, Y. Liu, A. Yu, Y. Zhang, C. Liu, J. Liu, W. Wu, K. Zhang, X. Shi, and J. Kou, "Flexible self-powered GaN ultraviolet photoswitch with piezo-phototronic effect enhanced on/off ratio," *ACS nano*, vol. 10, pp. 1572-1579, 2016.
[40] S. Rathkanthiwar, A. Kalra, S. V. Solanke, N. Mohta, R. Muralidharan, S. Raghavan, and D. N. Nath, "Gain mechanism and carrier transport in high responsivity AlGaN-based solar blind metal semiconductor metal photodetectors," *Journal of Applied Physics*, vol. 121, p. 164502, 2017.

RJC

RASĀYAN *J. Chem.*
Vol. 13 | No. 4 |2603-2608| October - December | 2020
ISSN: 0974-1496 | e-ISSN: 0976-0083 | CODEN: RJCABP
http://www.rasayanjournal.com
http://www.rasayanjournal.co.in

FLUORIDE REMOVAL FROM DRINKING WATER BY PRISTINE PUMICE AS ADSORBENT

S. K. Singh, S. Shweta, Mansi Shrivastava*and Shivankar Aggarwal
Department of Environmental Engineering, Delhi Technological University, Delhi-110042, India
*E-mail: mansis1998@gmail.com

ABSTRACT

Higher levels of Fluoride concentration in drinking water is a matter of serious concern in the present scenario. Controlling the Fluoride levels within the permissible range is of utmost importance to Human beings. The potential of pristine pumice to be used as an adsorbent for the elimination of fluoride from the drinking water is examined. The experiments are performed in batch equilibrium techniques. The influence of varying sorbent concentration, contact time and initial concentration of fluoride was also studied. It has been observed that the fluoride removal is directly proportional to the contact time and sorbent dose. The increase in the reaction period and the sorbent dose improved the quantity of fluoride adsorbed. The present study has indicated that an increase in fluoride initially present results in low efficacy of percentage removal. The adsorption process was found to follow the Langmuir Isotherm.

Keywords: Pumice, Adsorption, Fluoride Removal, Batch Techniques

© RASĀYAN. All rights reserved

INTRODUCTION

Water is an indispensable resource for both biotic as well as abiotic components. About 70% of Earth's surface is covered with water resources out of which 96% falls under Oceans and Seas and is rendered unfit for drinking due to the high salinity. Ice caps, glaciers, and permanent snow constitute approximately 1.7% of the total water. Only a small portion of water is available as groundwater which can be accessed by human beings. In India, 67% of the total population resides in the rural area and groundwater forms the basic source for satisfying their drinking as well as agricultural needs[1]. But nowadays it has become problematic as the water has become severely contaminated due to natural and anthropogenic factors. Water consists of numerous ions like SO_4^{2-}, Mg^{2+}, Ca^{2+}, F^- etc. Out of this, Fluoride tends to have both a positive as well as a detrimental effect on human health. The prevalence of fluoride in water aids the formation of enamel and bones in children and animals. Fluoride content up to 1mg/litre is considered beneficial for humans as prescribed by the Bureau of Indian Standards (BIS)[2]. World Health Organisation (WHO) has laid down a limitation on the concentration of fluoride at 1.5mg/litre in potable water[3]. Consumption of water with fluoride content greater than the permissible limit results in the development of white patches which further leads to dental fluorosis characterized by brown staining followed by pitting of teeth. The prolonged consumption of contaminated water with fluoride concentration exceeding 8 mg/litre results in pain in the joints and brittleness of bones. The intake of fluorides more than 20 mg/day over 20 years or more might result in chronic fluorosis[4].Dental and skeletal fluorosis are encountered due to the affinity of positively charged calcium ions in bones and teeth to the highly electronegative fluoride ions. Nausea, alterations in the functioning of the kidney, digestive system, central nervous system, nervousness, mottling of teeth, dwarfishness, reduced immunity are the other problems associated with the intake of fluoride-rich water. The severity of the problem depends on various factors like contact time of water with the aquifers, pH, climatic conditions, solubility of fluoride minerals, evaporation rate.

Out of all the elements in the periodic table, Fluorine is known for its highest electronegativity and reactivity. It is the thirteenth most abundant element in the earth's crust. On the virtue of its high reactivity, it does not occur in its natural state in groundwater. The minerals containing fluoride are Apatite $[Ca_5(PO_4)_3(Cl,F,OH)]$, Topaz $[Al_2F_2(SiO_4)]$, Cryolite (Na_3AlF_6), Fluorspar (CaF_2), Villiaumite(NaF). These minerals are found to be almost insoluble in water. Therefore, the traces of these minerals will be present in groundwater only when the conditions are favourable[6].

CrossMark

Table-1: Effect of Fluoride on Public Health[5]

Fluoride (F^-) Concentration (mg/litre)	Health Effect
< 0.5	Dental carries
0.5 - 1.5	Optimum Dental Health
1.5 - 4.0	Dental fluorosis
4.0 – 10	Dental and Skeletal fluorosis
>10	Crippling fluorosis

The occurrence of fluoride is mainly on the account of seepage of rainwater through the ground surface. The presence of fluoride is primarily on the account of the weathering of fluoride-rich rocks due to the rock-water interaction which often depends on the velocity of water, residence time and the solubility of fluoride minerals in the water. Consequently, fluoride from the rocks is leached out and augments the fluoride content in groundwater. Seawater has a fluoride concentration of roughly 1mg/litre. Groundwater quality in many areas is adversely affected by the fluoride and is presently 67mg/litre whereas surface water has a fluoride concentration of less than 0.1mg/litre. Foodstuff like fish and tea also contains fluoride to some extent. Although foodstuffs contain trace quantities of fluoride, it is the drinking water that accounts for 60% of the total intake of fluoride and makes fluorosis pandemic[7]. The other sources contributing to the deleterious effects of fluoride are toothpaste, phosphate fertilizers, volcanic activity, etc.

Globally, it was estimated that 260 million people were exposed to water consisting of fluoride greater than 1 mg/litre and the greatest problem was reported in China, India, Sri Lanka and few parts of Africa[8]. In India, Andhra Pradesh was the first state to face the issue of excessive fluoride in potable water in the year 1937[9]. Approximately 65% of India's villages and 62 million people are suffering from fluorosis[10].The contamination was adverse in Andhra Pradesh (1.8-8.4 mg/litre), Haryana (1.5-17 mg/litre), Punjab (0.44-6 mg/litre), Rajasthan (1.54-11.3 mg/litre) , Gujarat (1.6-6.8 mg/litre), Tamil Nadu (1.5-3.8 mg/litre) and Uttar Pradesh (1.5-3.11 mg/litre) as per the groundwater testing carried out in the districts in 2201[11].

Due to the deterioration of water quality and subsequent health implications, it is crucial to get rid of fluoride from drinking water before its utilization. For this purpose, various techniques have been adopted like the Nalgonda technique, Adsorption, Electrocoagulation, Membrane Filtration, etc. Adsorption is also known to be effective in the removal of heavy metals. For example, Nano adsorbents aid in the extraction of heavy metals from water[12]. A study reported satisfactory results on the removal of heavy metals and dyes using Nano adsorbents and plant residue[13]. The application of these techniques has been limited by its shortcomings.

In this paper use of Pumice, which is a lightweight volcanic rock has been discussed as an adsorbent for fluoride removal from water. Pumice has been used in the past by researchers also for fluoride removal and has been established as a potent adsorbent because of its large surface area, 90% porosity and ease of availability in nature. Pumice is known to have a low density as a result of the formation of cavities and cells due to gas expansion which results in the release of pressure. The removal efficiency of fluoride from drinking water by using pristine pumice as an adsorbent and its variation with sorbent dose, contact time and initial fluoride concentration has been discussed.

EXPERIMENTAL

Chemicals

The chemicals used in this study were of analytical grade. The experiment was conducted at 25°C with a steady agitation speed of 120 rpm. In order to achieve a stock solution of 100 mg/litre, 221 mg of NaF was added to 1 litre of deionized water. Standard solutions of appropriate concentrations were obtained by serially diluting the stock solution. Pumice Stone powder was procured from a local drug store in Delhi, India. The composition of Pumice is given in Table-2.

Preparation of Pumice

The Pumice Stone powder was washed with distilled water to remove dust. Further, the mixture was heated at 105°C for 24 hours in a hot air furnace to achieve a constant weight.

Table-2: Composition of Pumice[14]

Constituent	Percentage (%)
SiO_2	61.5
Al_2O_3	15.49
CaO	5.9
MgO	2.65
Fe_2O_3	8.4
K_2O	1.65
LOI (lost of ignition)	1.59

Batch Adsorption Study

Variation of Fluoride with Time

The effect of time on the amount of fluoride removal was assessed by making use of a fluoride spiked solution of 11mg/litre. Pumice in the dosage of 2g/100mL at a neutral pH was added to the solution. The resultant solution was kept in a shaker at 120 rpm and the fluoride concentration was recorded at an interval of 30 minutes.

Variation of Fluoride with Sorbent Dosage

The impact of sorbent dose on fluoride concentration was carried out by varying the amount of Pumice mixed with the standard solution of 11 mg/litre. The quantity of Pumice powder was kept as 2g/100mL, 3g/100 mL, 4g/100 mL, and 5g/100mL for 150 minutes at pH 7.

Variation with a Concentration of Fluoride

The effect on removal efficiency due to the fluoride initially present in water was analyzed by preparing standard solutions of working concentration (8, 9, 10, 11 and 12 mg/litre). The fluoride in resultant solutions was examined at a neutral pH after a contact time of 150 minutes.

$$\text{Removal Efficiency (E)} = \frac{[Initial\ fluoride\ (ppm) - Fi \quad fluoride\ (ppm)] * 100}{Initial\ fluoride\ (ppm)} \qquad (1)$$

RESULTS AND DISCUSSION

Impact of Contact Time on Removal Efficiency

The variance of the amount of fluoride with the residence time is depicted in Fig.-1. It has been found that with an increase in the contact time, the amount of fluoride adsorbed has increases when the pumice dosage was kept constant and agitation speed was set at 120 rpm. There was a gradual increase in efficiency from 19.09% at 30 minutes to 25.45% at 150 minutes. The advancement in removal efficiency was on the virtue of access to the adsorbent spots and dispersion of adsorbent particles in the solution. There was a hike in efficacy at 30 minutes but as time progressed; the performance was not significant due to the saturation of active sites.

Impact of the Sorbent Dosage on Removal Efficiency

The outcome of Pumice dosage on fluoride spiked solution is illustrated by Fig.-2. The results demonstrated that the fluoride ingested by Pumice is directly proportional to the amount of Pumice in the solution. The increasing adsorbent dosage contributes to a greater number of active sites in the solution which further induces a change in the removal efficiency[15]. When the Pumice dosage was varied from 2g/100mL to 5g/100mL, there was an upsurge in removal efficiency from 25.45% to 42.72%. However, the increase in percentage removal after 4g/100mL is lower as compared to the increase in percentage removal when the adsorbent dose exceeded 3g/100mL. This is on the account of the formation of a monolayer of fluoride on the adsorbent surface, shielding affect and agglomeration of adsorbent particles.[16]

Impact of the Initial Concentration of Fluoride on Removal Efficiency

There was an indirect relationship between the initial concentration of fluoride and the removal percentage, which is portrayed in Fig.-3. As the initial concentration of fluoride began to rise, the removal performance

declined with constant pH, temperature, agitation speed and Pumice dose. When the initial strength was 8 mg/litre, the percentage of extracted fluoride was found to be 30% and when the initial dosage of fluoride rose to 12 mg/litre, the efficacy depreciated to 20%. At lower fluoride concentration, access to a greater number of active spots on the adsorbent leads to a higher removal percentage of fluoride. The fall in efficiency is attributed to increasing fluoride concentration. At a higher initial concentration, the fluoride ions compete for the active spots which leads to a fall in the removal efficiency.

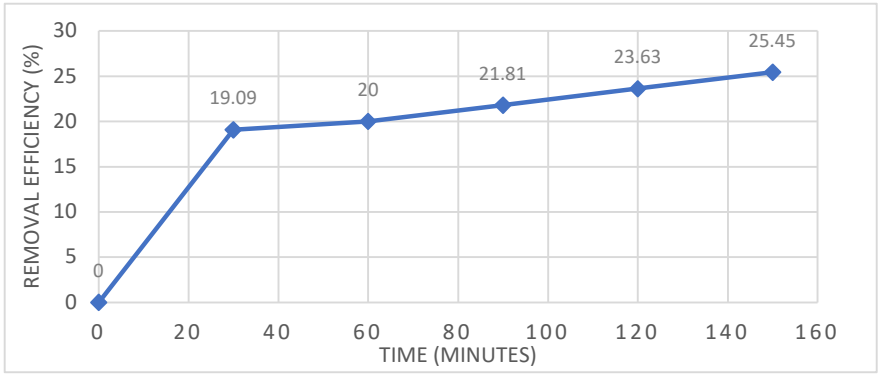


Fig.-1: Impact of Time on Removal Efficiency (11mg/litre Fluoride Solution, Temperature 25°C, Neutral pH, 2g/100mL Sorbent Dose)

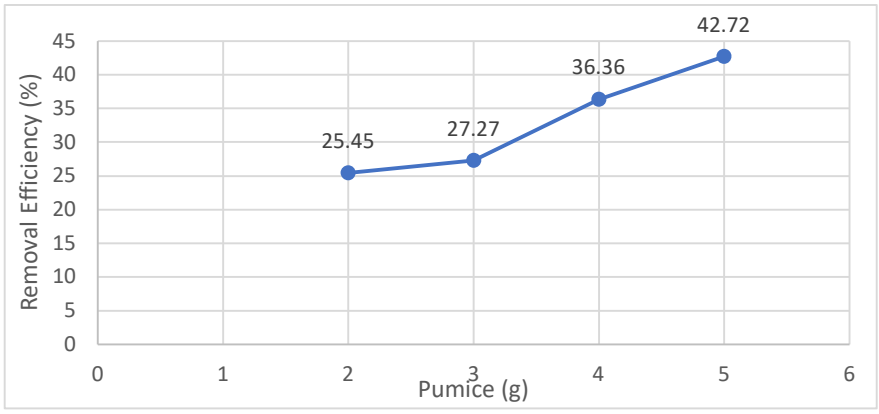


Fig.-2: Impact of Sorbent Dosage on Removal Efficiency (11mg/litre Fluoride Solution, Temperature 25°C, Neutral pH)

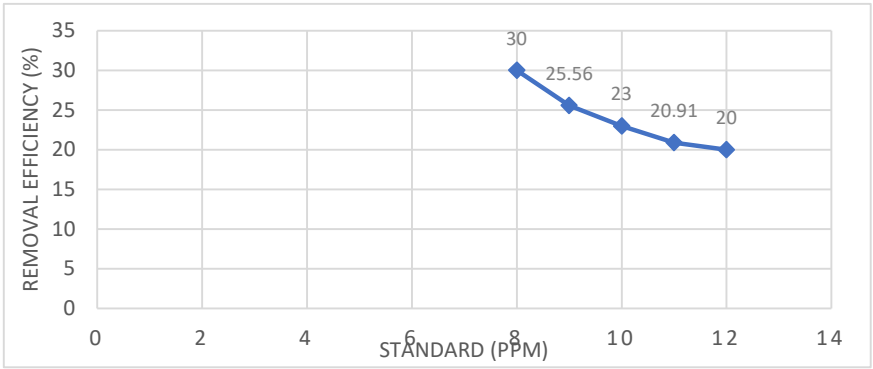


Fig.-3: Impact of Initial Fluoride Concentration on Removal Efficiency (2g/100mL Sorbent Dose, Temperature 25°C, Neutral pH)

Adsorption Isotherms

In this study, Langmuir and Freundlich's Isotherms were plotted. By plotting a graph between $\frac{C_e}{q_e}$and C_e, the value of R^2 was determined. The general equation of Langmuir Isotherm is given by:

$$\frac{C_e}{q_e} = \left(\frac{1}{q_m K_L}\right) + Ce\left(\frac{1}{q_m}\right) \quad (2)$$

The R^2 value was maximum for Langmuir Isotherm, which was 0.9869. Hence, it was inferred that the adsorption of fluoride on Pumice follows Langmuir Isotherm. Separation factor, R_L in the case of Langmuir Isotherm indicates whether the isotherm is favourable or not favourable. It is represented by the equation-3.[17]

$$R_L = \frac{1}{1+K_L C_o} \quad (3)$$

Table-3: Values of Separation Factor

Value of R_L	Isotherm
$R_L > 1$	Not Favorable
$R_L = 1$	Linear
$0 < R_L < 1$	Favorable
$R_L = 0$	Irreversible

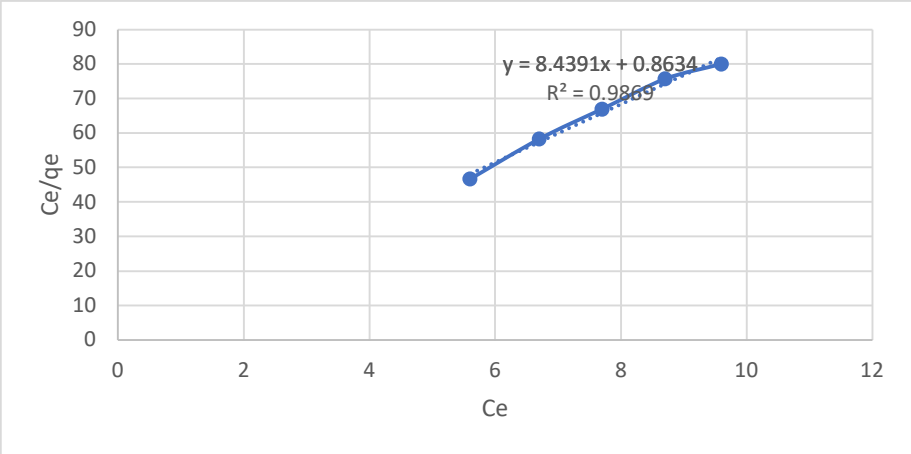


Fig.-4: Langmuir Adsorption Isotherm

CONCLUSION

Following conclusions are drawn from the study:

- Pristine Pumice can be utilized to extract fluoride effectively from the fluoride contaminated solution on the virtue of its ease of availability and low cost.
- The effectiveness of Pumice as an adsorbent is determined by the actual fluoride concentration in the water, sorbent dosage and the duration for which Pumice is in contact with the solution.
- A direct relationship between contact time and the percentage of fluoride removal was witnessed. As the contact time increased, the ability of Pumice to adsorb fluoride was enhanced. When the reaction time was varied from 30 minutes to 150 minutes, the removal capacity gradually increased from 19.09% to 25.45%. The removal efficiency was also a function of sorbent dosage. The efficiency experienced a rise when the sorbent dosed was increased. The removal percentage advanced from 25.45% to 42.72% when the sorbent dosed was varied from 2g/100mL to 5g/100mL.
- With a rise in the initial concentration of fluoride in water, the removal percentage was observed to decrease. The reduction rate fell from 30% to 20%as the strength of fluoride solution increased from 8mg/liter to 12mg/liter.
- R_L value for Langmuir Isotherm was 0.013 which indicated that the isotherm was favorable.

ACKNOWLEDGEMENT

The authors are thankful to the Head of Department of Environmental Engineering and Administration, Delhi Technological University for permitting us to conduct the above study.

REFERENCES

1. https://www.business-standard.com/article/economy-policy/70-indians-live-in-rural-areas-census-111071500171_1.html
2. Drinking water- specification IS 10500, Bureau of Indian Standards, New Delhi, (1991).
3. https://www.who.int/water_sanitation_health/publications/2011/dwq_guidelines/en/
4. A. Jain, S.K. Singh, *Open Journal of Water Pollution and Treatment,* **1(2)**,50(2014), **DOI:** 10.15764/wpt.2014.02006
5. M. Mohapatra, S. Anand, B. Mishra, D. Giles and P. Singh, *Journal of Environmental Management*, **91(1)**,67(2009), **DOI:**10.1016/j.jenvman.2009.08.015
6. A. Jain, S.K. Singh, *International Journal of Engineering Science and Innovative Technology*, **3**, 591(2013).
7. S. Jagtap, M. Yenkie, N. Labhsetwar and S. Rayalu, *Chemical Reviews*, **112(4)**, 2454(2012), **DOI:** 10.1021/cr2002855k
8. J.E.Frencken, Endemic Fluorosis in Developing Countries: Causes, Effects and Possible Solutions(1992).
9. S. Ayoob and A. Gupta, *Critical Reviews in Environmental Science and Technology*, **36(6)**,433(2006), **DOI:**10.1080/10643380600678112
10. S. Gitte, R. Sabat and K. Kamble, *International Journal of Medicine And Public Health*, **5(3)**, 232(2015), **DOI:**10.4103/2230-8598.161539
11. A. Susheela, *Current Science,* **77(10)**,1250(1999).
12. A. Kaushal, SK Singh, *Journal of Environmental and Biotechnology Research*, **6(1)**, 96(2017)
13. N. Sharma, D.P. Tiwari, S.K. Singh, *Rasayan Journal of Chemistry*, **13(2)**, 1063(2020), **DOI:** 10.31788/RJC.2020.1325405
14. M. Malakootian, M. Moosazadeh, N. Yousefi, A. Fatehizadeh, *Journal of Environmental Science and Technology,* **5(4)**(2011).
15. P. Mondal, C. Majumder and B. Mohanty, *Journal of Hazardous Materials*, **150(3)**, 695(2008).
16. R. Rajumon, J. Anand, A. Ealias, D. Desai, G. George and M. Saravanakumar, *Journal of Environmental Chemical Engineering*, **7(6)**,103479 (2019).
17. M. Salim & Y. Munekage, *International Journal of Chemistry*, **1(1)**, (2009), **DOI:**10.5539/ijc.v1n1p23

[RJC-5967/2020]

Footer Voltage Controlled Dual Keeper Domino Logic with Static Switching Approach

Chirag PARASHAR, Avijeet Kumar TRIVEDI, Aman AGARWAL, Neeta PANDEY

Department of Electronics and Communication Engineering, Delhi Technological University, Bawana Road, Shahbad Daulatpur Village, Rohini, 110042 Delhi, India

mechiragp1997@gmail.com, avijeet.trivedi@gmail.com, amanagarwal702@gmail.com, n66pandey@rediffmail.com

DOI: 10.15598/aeee.v18i4.3794

Abstract. *In this paper, two circuits, namely Footer Voltage Controlled Dual Keeper domino logic (FVCDK) and Footer Voltage Controlled Dual Keeper with Static Switching domino logic (FVCDK-SS) are presented, in order to achieve high speed, low power consumption and robustness. The dual keeper arrangement helps in reducing the loop gain of the feedback circuitry, which leads to lower delay variability. The keeper circuitry is controlled using the footer voltage to reduce the contention current in the initial evaluation phase, and thus providing enhanced speed. In FVCDK-SS domino logic, unwanted transients at the output are reduced by incorporating pseudo-dynamic buffer in the proposed FVCDK domino logic. This further reduces the dynamic power consumption. The results of the logic presented here are validated by comparing them to a wide range of existing domino logic circuits for a variety of performance metrics such as delay, power, power-delay product and unity noise gain. To effectively gauge the wide fan-in capabilities of the proposed logic, results are shown for the various fan-in OR gate. The simulations of the circuits are carried out using industry standard full-suite Cadence tools using 45 nm technology library.*

Keywords

Contention current, corner analysis, delay variability, domino logic, static switching.

1. Introduction

The low-power and high-speed requirement for wide fan-in applications such as SRAM, pre-encoders, OR gates and tag comparators [1] have been effectively accomplished by the domino logic circuit design. A conventional domino logic, shown in Fig. 1, consists of a pre-charge transistor (M_{Pre}), Controlled by Clock (CLK), and a Pull-Down Network (PDN). The low number of transistors needed for the domino logic make it an attractive choice for wide fan-in circuits, as only two additional transistors are needed in addition to the evaluation network. Moreover, this logic only needs either a Pull-Up Network (PUN) or PDN, which leads to its immensely small footprint in comparison to the static CMOS logic. When CLK is LOW, M_{Pre} turns ON and charges the dynamic node (Dyn_{node}) to the Supply Voltage (VDD). This is known as the pre-charge phase. When CLK goes HIGH, M_{Pre} turns OFF and the Dyn_{node} is conditionally discharged to Ground (GND), if the PDN is evaluated to be TRUE. This is known as the evaluation phase.

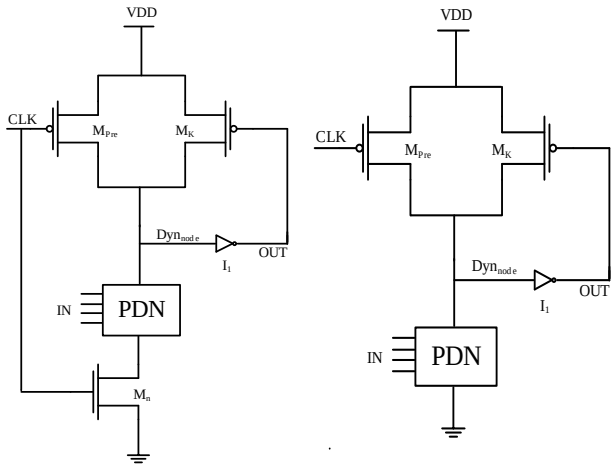
The presence of footer NMOS M_n is optional, giving two distinguished forms of conventional domino logic, namely, Footed Domino Logic (FDL) (Fig. 1(a)) and Footless Domino Logic (FLDL) (Fig. 1(b)). The footless domino logic is faster than the footed domino logic, because there is no stacking effect. However, this speed improvement comes at the cost of increased power consumption and leakage current in comparison to the footed domino logic.

In order to prevent the Dyn_{node} from being in the high impedance state during the evaluation phase, in case the PDN evaluates to be FALSE, a keeper transistor (M_K) is used. It counteracts the charge leakage at the dynamic node during the evaluation phase in such scenario [2], [3], [4] and [5]. If the clock frequency is low, then, in the absence of keeper transistor, voltage at the Dyn_{node} can stray from its ideal value of VDD to a lower value due to charge leakage during the elongated evaluation phase. The magnitude to which the Dyn_{node} falls below VDD can be taken as a mea-

sure of robustness for the circuit, with values closer to VDD pertaining to the circuits that are more robust. It can be enhanced by increasing the aspect ratio of the keeper, but at the cost of speed. This is because if the PDN evaluates to be TRUE in the evaluation phase, it tries to discharge the node, while M_K tries to keep it at VDD. Several modifications have been made in the conventional domino logic to reduce the contention current for increased speed and lower power.

In this paper, two novel domino structures, FVCDK and FVCDK-SS have been introduced. Both circuits take advantage of a controlled dual keeper arrangement to minimize the contention current at the early stage of evaluation and a controlled discharge path. These two techniques working in parallel lead to improvement in speed, power and power-delay product.

The organization of this paper is as follows: Section 2. presents an overview of the previous studies in the field of domino logic circuits. Section 3. provides explanation of the proposed circuits and design methodology. Section 4. presents the simulation results obtained for the proposed logic and its comparison with existing architectures. Section 5. concludes the discussion.



(a) Footed. (b) Footless.

Fig. 1: Conventional Domino Logic.

2. Related Work

The study on domino logic focuses on one or more of the following:

- Counteracting the leakage current.
- Low power consumption.
- Speed enhancement.
- Reduced unwanted switching at the output.
- Increased noise margin.

These enhancements are achieved by either modifying keeper circuitry (to reduce contention current), or adding static switching mechanism at the output node (to reduce unwanted transients), or providing an additional discharge path (for improving speed). These techniques are described in the following sub-sections.

2.1. Dual Keeper Modification

In the conventional domino logic shown in Fig. 1(b), the positive feedback loop formed by the output inverter I_1 and keeper M_K increases the delay variability produced due to variation in various process parameters, such as C_{ox}, t_{ox} etc. This loop gain is given by Eq. (1).

$$A_{loop} = A_{inv} \cdot g_{m/keeper} \cdot Z_{dyn}, \tag{1}$$

where A_{inv} is the gain of I_1, $g_{m/keeper}$ is the transconductance of keeper M_K, and Z_{dyn} is the impedance at the dynamic node.

The Grounded PMOS Keeper (GPK) domino logic [6] shown in Fig. 2(a) divides the keeper M_K into M_{K1} and M_{K2}, such that the sum of their length is same as M_K. This reduces the loop gain of the output node, thus reducing the delay variability. However, since the keeper M_{K2} always remains ON, this leads to a high contention current, and hence reduced speed.

This contention current can be reduced by making modification in the dual keeper circuitry and turning it OFF during the initial stage of the evaluation phase. In the Clock Delayed Dual Keeper (CDDK) domino logic [7] depicted in Fig. 2(b), the keeper circuitry is enabled after a delay produced by the inverter I_1 during the initial part of the evaluation phase. This leads to the truncation of contention current and hence enhanced speed.

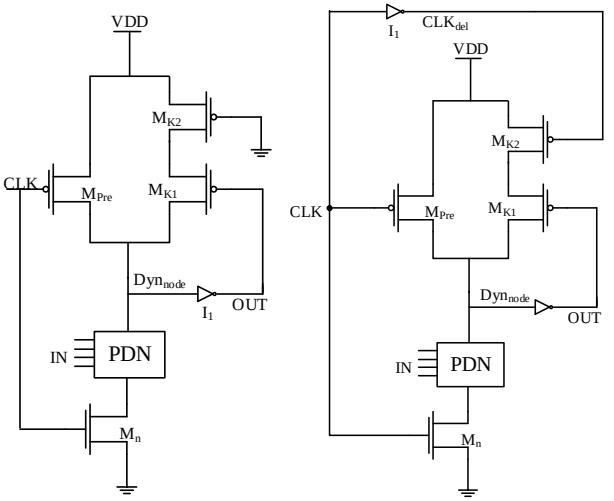
2.2. Static Switching Mechanism

Unwanted switching transients during the pre-charge phase at the output node elevates the problem of dynamic power consumption in conventional domino logic circuits. This is because during the evaluation phase, if PDN evaluates TRUE, Dyn_{node} is discharged to GND, hence OUT is charged to VDD. In the subsequent pre-charge phase, Dyn_{node} is charged to VDD by M_{Pre}, causing OUT to get discharged to GND. This transient in the output node is undesired as output would again be charged to VDD in the next evaluation phase, if the inputs to the PDN remains unchanged.

Figure 3(a) and Fig. 3(b) show the domino logic using Pseudo-Dynamic Buffer (PDB) [8] and Clock Delayed Dual Keeper with Static Switching (CDDK-SS) mechanism [9] respectively, that reduce this unwanted transient at the output node. In both the logics, source terminal of the NMOS M_{n1} of the output inverter is connected to V_{foot}, instead of GND. If inputs to the PDN remain HIGH for two consecutive cycles, charging of Dyn_{node} to VDD by M_{Pre} in the pre-charge phase will not cause the discharging of OUT to GND. This is because M_n remains OFF, and since PDN is TRUE, voltage at V_{foot} rises, thus preventing M_{n1} from discharging the output node.

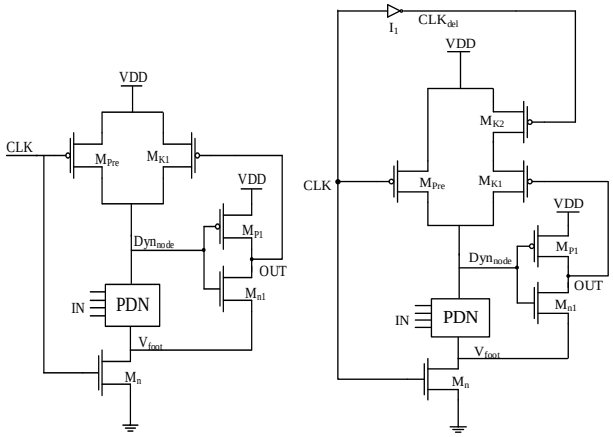
2.3. Dual Keeper with Additional Discharge Path

In the evaluation phase, if PDN evaluates to be TRUE, then the time required to discharge Dyn_{node} to GND depends upon the path delay offered by the PDN and the footer transistor M_n. In the Clock Delayed Dual Keeper domino logic with Additional Discharge Path (CDDK-ADP) [10] shown in Fig. 4, an additional discharge path for the Dyn_{node} is available through M_{n1} and M_{n2}. During the pre-charge phase, Dyn_{node} is charged to VDD. Since CLK is LOW, M_{n2} is in cut-off region, and hence the additional discharge path remains OFF. In the initial part of the evaluation phase, if the PDN evaluates to be TRUE, voltage at V_{foot} rises, which turns M_{n1} ON. And since CLK is high during this period, M_{n2} also turns ON. This enables the discharging of Dyn_{node} through the additional discharge path, comprising of M_{n1} and M_{n2}, and hence reduces the delay.



(a) Grounded PMOS Keeper (GPK) [6]. (b) Clock Delayed Dual Keeper (CDDK) [7].

Fig. 2: Footed Dual Keeper Domino Logic.



(a) Pseudo-Dynamic Buffer (PDB) [8]. (b) Clock Delayed Dual Keeper with Static Switching (CCDK-SS) [9].

Fig. 3: Domino Logic with Static Switching.

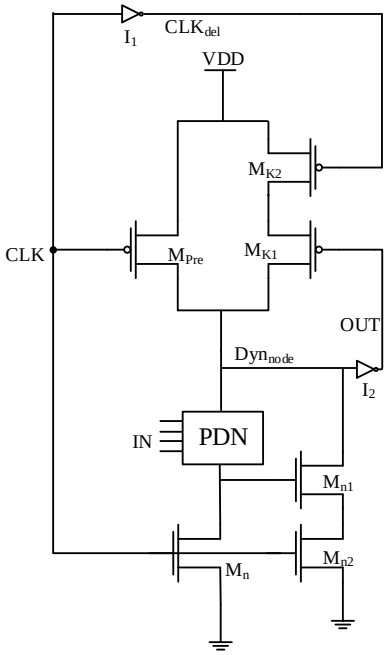
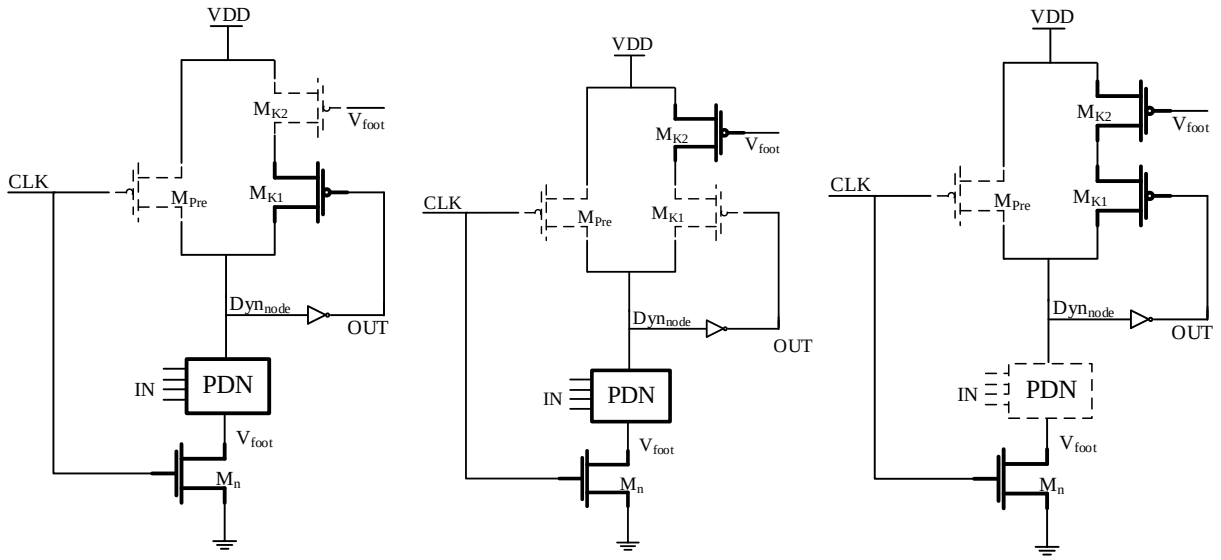


Fig. 4: Clock Delayed Dual Keeper with Additional Discharge Path (CDDK-ADP) [10].

3. Proposed Work

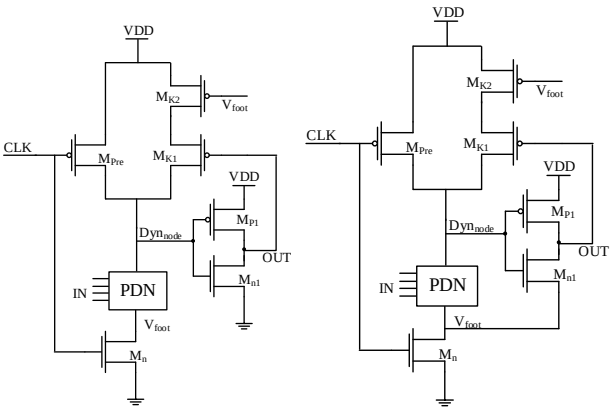
Figure 6(a) and Fig. 6(b) show the structures of the proposed FVCDK domino logic and FVCDK-SS domino logic respectively. Both of these logics make use of an additional keeper transistor M_{K2} along with the conventional transistor M_{K1}. The transistor M_{K2}



(a) During initial evaluation phase if PDN is TRUE. (b) During evaluation phase if PDN is TRUE. (c) During evaluation phase if PDN is FALSE.

Fig. 5: Phases of operation in FVCDK.

is controlled by applying the footer voltage (V_{foot}) at its gate terminal.



(a) Footer Voltage Controlled Dual Keeper (FVCDK). (b) Footer Voltage Controlled Dual Keeper with Static Switching (FVCDK-SS).

Fig. 6: Proposed Domino Circuits.

The operation of the proposed FVCDK circuit is elucidated in Fig. 5 for the evaluation phase when PDN evaluates to be either TRUE or FALSE. Bold and dotted lines are used to represent the state of transistors as ON and OFF respectively. During the initial evaluation phase, if PDN evaluates to be TRUE, voltage at V_{foot} increases. The extent of this increase depends upon the relative sizing of the PDN and the footer transistor (M_n). An increase in this extent can be obtained by increasing the aspect ratio of the transistors in the PDN. This increase in V_{foot} reduces the driving capability of the keeper M_{K2}, and hence reduces contention current. This in turn speeds up the discharging of the Dyn_{node}. After the Dyn_{node} is discharged sufficiently, even if M_{K2} turns ON, the keeper M_{K1} would remain OFF, leading to this topology yielding a lower contention current throughout the discharging of the Dyn_{node}. Therefore it may be inferred that if the PDN is TRUE, either of the two keepers is in the OFF state, thus reducing the contention current.

The FVCDK-SS works in a similar way to FVCDK, but in order to eliminate the unnecessary switching at the output node and lower the dynamic power consumption of the circuit, PDB is used. The advantages of PDB are described in Subsec. 2.2.

4. Results

The simulations of the circuits are carried out using industry standard full-suite Cadence® tools using 45 nm technology library. We have analyzed various footed domino circuit topologies using metrics like power consumption, delay, Power-Delay Product (PDP) and Unity Noise Gain (UNG). For comparison purpose, all the transistors are kept at minimum size, i.e., channel width and length are set to 120 nm and 45 nm respectively, with supply voltage and temperature set to 1.2 V and 300 K, respectively. Footless topologies are not considered for this comparison, as those would lead to high power consumption due to excessive leakage at 45 nm technology node.

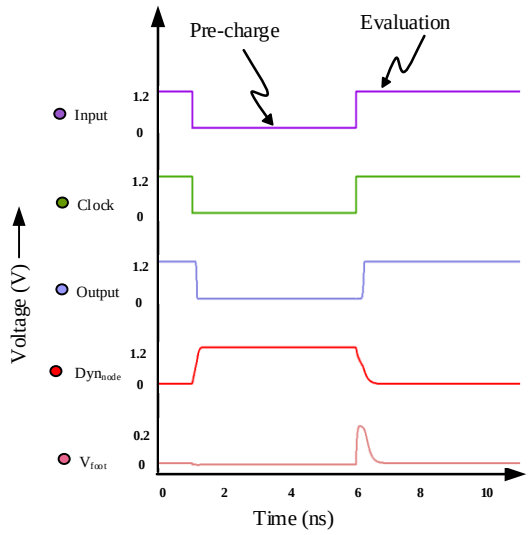


Fig. 7: Output waveform of the proposed circuit during pre-charge and evaluation phase.

Figure 7 shows the output waveform of the proposed FVCDK circuit during the pre-charge and evaluation phase. The delay is calculated by making one of the inputs HIGH as the evaluation phase starts and calculating the time taken for the OUT signal to rise from LOW to VDD/2. The power is determined under the same simulation environment by calculating the average power consumption in a period of time. UNG readings are taken by applying a noise pulse of 100 ps duration to all the inputs of the PDN during the evaluation phase [11]. The amplitude of the noise pulse is varied, until the output node reaches the same amplitude. The UNG is a measure of DC robustness of the circuits.

Tab. 1: Power, delay, PDP and UNG for 128 input OR gate.

Topologies	Delay (e^{-10} s)	Power (µW)	PDP (e^{-16} J)	UNG (mV)
GPK [6]	2.627	2.024	5.317	945
CDDK [7]	2.619	2.064	5.404	944
CDDK-SS [9]	1.888	1.739	3.282	831
CDDK-ADP [10]	2.638	2.077	5.480	945
FVCDK	2.313	1.873	4.333	879
FVCDK-SS	1.799	1.658	2.983	817

Table 1 shows the power, delay, PDP and UNG values for 128 fan-in OR gate designed using various footed domino topologies, including the proposed FVCDK and FVCDK-SS logic. It can be seen from Fig. 8 that FVCDK-SS offers the smallest rise time delay and the lowest power consumption, hence lowest PDP. On the other hand, FVCDK offers improvement in all three metrics when compared to the pre-existing domino logic circuits, with the exception of CDDK-SS. While the speed and power consumption of FVCDK are slightly worse than that of CDDK-SS, the logic makes up for it by having a larger UNG value, as depicted in Fig. 9, and thus resulting in a more robust circuit.

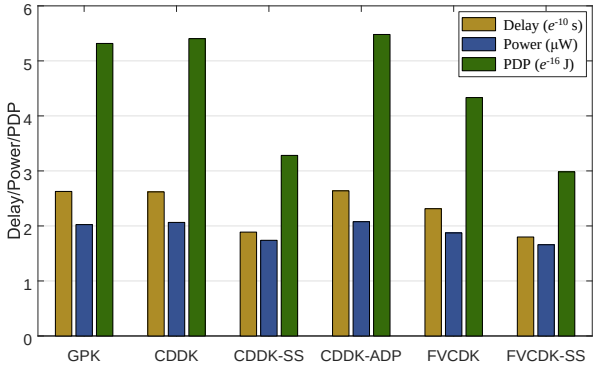


Fig. 8: Comparison of power, delay and PDP.

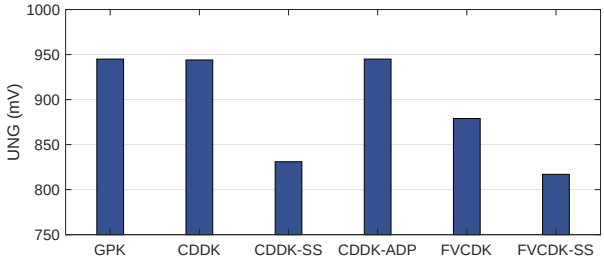


Fig. 9: Comparison of UNG.

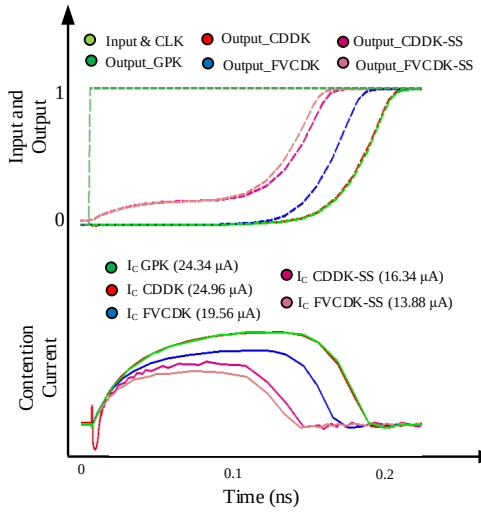
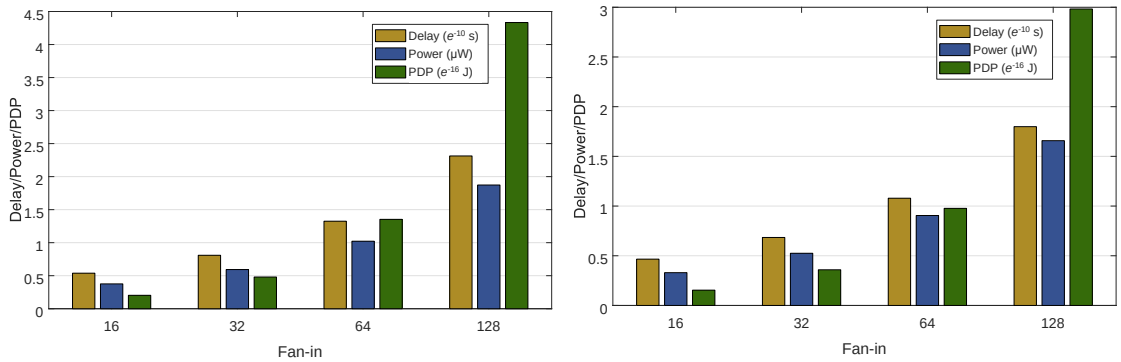


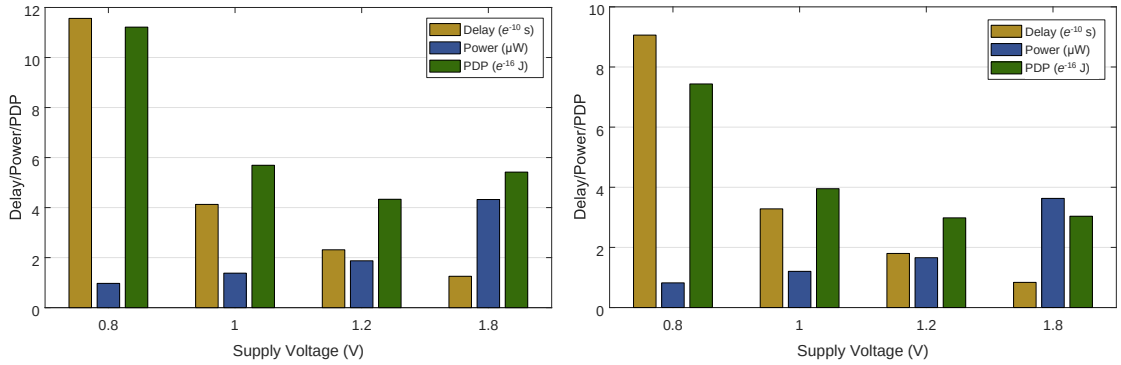
Fig. 10: Contention current obtained for GPK, CDDK, CDDK-SS, FVCDK and FVCDK-SS.

Figure 10 shows the plot for the contention current obtained for GPK, CDDK, CDDK-SS, and proposed FVCDK and FVCDK-SS logic, along with their maximum value. It can be noticed that speed improvement of FVCDK-SS and FVCDK domino circuits from



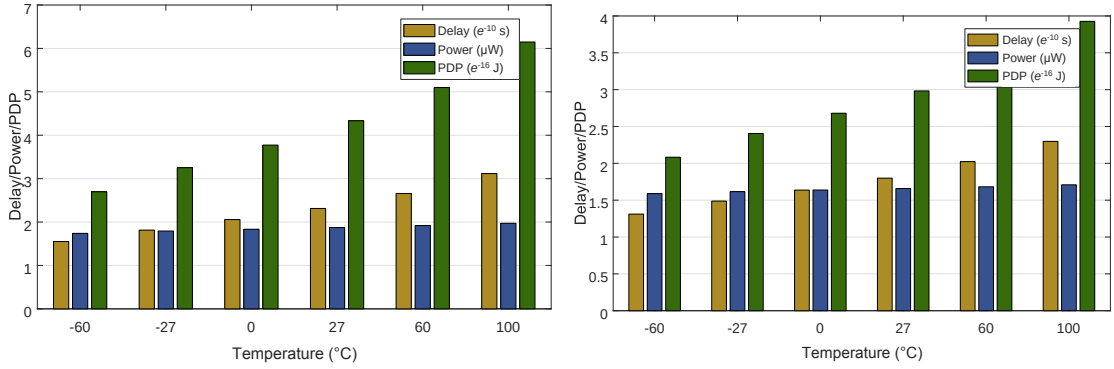
(a) (b)

Fig. 11: Power, delay and PDP variation with fan-in for (a) FVCDK (b) FVCDK-SS.



(a) (b)

Fig. 12: Power, delay and PDP variation with supply voltage for (a) FVCDK (b) FVCDK-SS.



(a) (b)

Fig. 13: Power, delay and PDP variation with temperature for (a) FVCDK (b) FVDK-SS.

CDDK-SS and CDDK respectively is attributed to the reduced contention current in the proposed circuits. There is no contention reduction mechanism in GPK, which leads to the largest delay for the GPK domino logic.

Figure 11 shows the effect of fan-in on delay, power consumption and PDP for OR gate designed using FVCDK and FVCDK-SS. With higher fan-in, capacitance at the dynamic node increases, leading to an increased power consumption, as well as increased delay. Figure 12 and Fig. 13 depict the effect of change in supply voltage and temperature respectively, on power, delay and PDP values. With a rise in VDD, power starts to increase because of the increase in leakage current. This is accompanied by a reduction in delay because of the increased driving capacity of the circuit. Both power and delay increase with a rise in temperature.

Tab. 2: Circuit designing using MTCMOS devices: comparison of power, delay, PDP and UNG.

Topologies	Delay (e^{-10} s)				Power (µW)			
	CDDK	FVCDK	CDDK-SS	FVCDK-SS	CDDK	FVCDK	CDDK-SS	FVCDK-SS
SS	3.924	3.370	2.883	2.739	2.170	1.927	1.803	1.708
SF	3.752	3.121	2.334	2.200	2.484	2.178	1.920	1.822
TT	2.619	2.313	1.888	1.799	2.064	1.873	1.739	1.658
FS	2.016	1.844	1.584	1.528	1.791	1.655	1.587	1.516
FF	1.931	1.746	1.329	1.284	2.035	1.875	1.710	1.638

Topologies	PDP (e^{-16} J)			
	CDDK	FVCDK	CDDK-SS	FVCDK-SS
SS	8.514	6.496	5.199	4.678
SF	9.319	6.796	4.481	4.008
TT	5.404	4.333	3.282	2.983
FS	3.611	3.053	2.513	2.317
FF	3.930	3.274	2.273	2.103

To obtain better performance, transistors in the PDN and the PMOS transistor in the output inverter were replaced with their lower V_{th} counterpart. Moreover, to achieve lower power consumption, all the other transistors were replaced with their higher V_{th} components. As the high V_{th} transistors lie on the non-critical delay path, the performance of the circuit is not affected. The results obtained for 128-input OR gate using this approach are shown in Tab. 3. A speed improvement of 47 % and power reduction of 14 % is observed for FVCDK-SS from this technique.

Tab. 3: Circuit designing using MTCMOS devices: comparison of power, delay, PDP and UNG.

Topologies	Delay (e^{-10} s)	Power (µW)	PDP (e^{-16} J)	UNG (mV)
FDL	1.580	1.799	5.130	707
FDL-SS	1.110	1.524	1.692	626
FVCDK	1.089	1.489	1.622	617
FVCDK-SS	0.959	1.427	1.368	593

In order to verify the robustness of the proposed circuits, corner case analysis was conducted. In this analysis, circuits were simulated using transistors with extreme fabrication parameters. The delay and temperature obtained for the circuit in these corner conditions tend to deviate from their typical corner value. A circuit is said to have an inadequate design margin, if it does not function correctly at any of these process extremes. The corner cases used for the analysis are Slow NMOS-Slow PMOS (SS), Slow NMOS-fast PMOS (SF), fast NMOS-Slow PMOS (FS) and fast NMOS-fast PMOS (FF). Table 2 presents the value of delay, power and PDP for these corners, including the results for the Typical NMOS-Typical PMOS (TT) case. The topologies used were CDDK, FVCDK, CDDK-SS and FVCDK-SS. A 128 input OR gate was taken up for this analysis.

Also, the Monte-Carlo Simulations were performed over 1000 points in order to obtain the delay variability for the compared topologies. The delay variability was obtained by dividing the standard deviation value of the delay with the mean value. Table 4 shows the values obtained for both the standard deviation, as well as the variability factor of the delay. The variability factors obtained for FVCDK and FVCDK-SS were 13.56 % and 9.29 %, respectively, down from 19.04 % and 10.50 % for CDDK and CDDK-SS, respectively. This illustrates the better delay invariability offered by the proposed domino circuits from the CDDK variant.

5. Conclusion

The novel FVCDK and FVCDK-SS domino circuits were proposed in this paper. Both the circuits are designed for the contention current reduction by switching OFF the keeper arrangement during the initial part of the evaluation phase. This leads to a reduced power consumption and increased speed at the cost of a little degradation in the UNG value, as compared to the existing dual keeper footed domino circuits. In addition, FVCDK-SS domino circuit offers even better speed and power efficiency than the FVCDK by trading off the UNG value. This is due to the reduction in the unnecessary switching of the output node.

The simulation for all the process corners validates that both circuits are process variation tolerant. Standard deviation of delay for all the corner cases about the typical values illustrates that the proposed circuits' performance is more invariant to the process change as compared to the CDDK and CDDKS-SS.

Tab. 4: Standard deviation for 128 input OR gate.

Delay (e^{-10} s)	CDDK	FVCDK	CDDK-SS	FVCDK-SS
SD (σ)	0.516	0.321	0.200	0.171
Mean (μ)	2.712	2.368	1.909	1.835
Delay Variability ($\sigma \cdot \mu^{-1}$) (%)	19.04	13.56	10.50	9.29

The simulations are also carried out for a range of temperature and supply voltage variations to validate the robustness of the circuit. Also, the use of MTCMOS results in a 53 % and 47 % speed improvement in FVCDK and FVCDK-SS, respectively. However, this improvement comes at the cost of increased complexity in terms of fabrication, as both the lower V_{th} and higher V_{th} counterparts for the same MOS needs to be fabricated on the same die. By comparing the proposed circuits with the conventional footed domino circuit, where all are designed using MTCMOS, we note that the proposed circuits outperform the latter in speed, power and PDP metrics by a significant margin.

References

[1] DING, L. and P. MAZUMDER. On circuit techniques to improve noise immunity of CMOS dynamic logic. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*. 2004, vol. 12, iss. 9, pp. 910–925. ISSN 1557-9999. DOI: 10.1109/tvlsi.2004.833668.

[2] CHATTERJEE, B., M. SACHDEV and R. KRISHNAMURTHY. Leakage control techniques for designing robust, low power wide-OR domino logic for sub-130nm CMOS technologies. In: *International Symposium on Signals, Circuits and Systems. Proceedings*. San Jose: IEEE, 2004, pp. 415–420. ISBN 0-7695-2093-6. DOI: 10.1109/isqed.2004.1283709.

[3] MORADI, F., T. V. CAO, E. I. VATAJELU, A. PEIRAVI, H. MAHMOODI and D. T. WISLAND. Domino logic designs for high-performance and leakage-tolerant applications. *Integration*. 2013, vol. 46, iss. 3, pp. 247–254. ISSN 0167-9260. DOI: 10.1016/j.vlsi.2012.04.005.

[4] OKLOBDZIJA, V. G. and R. K. MONTOYE. Design-performance trade-offs in CMOS-domino logic. *IEEE Journal of Solid-State Circuits*. 1986, vol. 21, iss. 2, pp. 304–306. ISSN 0018-9200. DOI: 10.1109/jssc.1986.1052519.

[5] PEIRAVI, A. and M. ASYAEI. Robust low leakage controlled keeper by current-comparison domino for wide fan-in gates. *Integration*. 2012, vol. 45, iss. 1, pp. 22–32. ISSN 0167-9260. DOI: 10.1016/j.vlsi.2011.07.002.

[6] PALUMBO, G., M. PENNISI and M. ALIOTO. A simple keeper topology to reduce delay variations in nanometer domino logic. *IEEE International Symposium on Circuits and Systems*. 2012, vol. 59, iss. 10, pp. 2292–2300. ISSN 1549-8328. DOI: 10.1109/iscas.2012.6271554.

[7] ANGELINE, A. A. and V. S. K. BHAASKARAN. High Performance Domino Logic Circuit Design by Contention Reduction. *Lecture Notes in Electrical Engineering VLSI Design: Circuits, Systems and Applications*. 2018, vol. 469, pp. 161–168. ISBN 978-981-10-7250-5 DOI: 10.1007/978-981-10-7251-2_18.

[8] TANG, F., A. BERMAK and Z. GU. Low power dynamic logic circuit design using a pseudo dynamic buffer. *Integration*. 2012, vol. 45, iss. 4, pp. 395–404. ISSN 0167-9260. DOI: 10.1016/j.vlsi.2011.08.003.

[9] VARGHESE, A., S. R. ANUSHA, A. A. ANGELINE and V. S. K. BHAASKARAN. Clock Delayed Dual Keeper Domino-Logic Design with Reduced Switching. *International Journal of Engineering and Advanced Technology Special Issue*. 2019, vol. 9, iss. 1S3, pp. 397–402. ISSN 2249–8958. DOI: 10.35940/ijeat.a1072.1291s319.

[10] ANGELINE, A. A. and V. S. K. BHAASKARAN. Speed enhancement techniques for Clock-Delayed Dual Keeper Domino logic style. *International Journal of Electronics*. 2020, vol. 107, iss. 8, pp. 1239–1253. ISSN 0020-7217 DOI: 10.1080/00207217.2020.1726486.

[11] MEIMAND, H. M. 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*. 2004, vol. 51, iss. 3, pp. 495–503. ISSN 1549-8328. DOI: 10.1109/tcsi.2004.823665.

About Authors

Chirag PARASHAR was born in Delhi, India in 1997. He received his B.Tech. in Electronics and Communication Engineering from Delhi Technological University (formely Delhi College of Engineering), Delhi in 2020. His research interests include low power electronics, VLSI design, and computer architecture.

Avijeet Kumar TRIVEDI was born in Delhi, India in 1996. He received his B.Tech. in Electronics and Communication Engineering from Delhi Technological University (formely Delhi College of Engineering), Delhi in 2020. His main research interests include hardware-software co-design, digital design, computer architecture and VLSI design.

Aman AGARWAL was born in Pakur, India in 1998. He received his B.Tech. in Electronics and Communication Engineering from Delhi Technological University (formely Delhi College of Engineering), Delhi in 2020. His main research interests include digital circuit design, Very Large Scale Integrated (VLSI) circuit design using CMOS sub-nm technologies, and low power FPGA designing.

Neeta PANDEY is currently a Professor in Department of Electronics and Communication Engineering, Delhi Technological University. She received her M.E. in Microelectronics from Birla Institute of Technology and Sciences, Pilani and Ph.D. from Guru Gobind Singh Indraprastha University Delhi. Her research interests are in Analog and Digital VLSI Design.

Int. j. inf. tecnol.
https://doi.org/10.1007/s41870-020-00603-9

ORIGINAL RESEARCH

Framework for agricultural e-trading platform adoption using neural networks

Sanjay Chaudhary[1,2] · **P. K. Suri**[2]

Received: 24 June 2020 / Accepted: 26 December 2020
© Bharati Vidyapeeth's Institute of Computer Applications and Management 2021

Abstract The digital e-trading platform is an ambitious and priority intervention in the agricultural sector. The research study is among the very few that identifies and prioritizes the predictors of wholesale electronic trading platform adoption in agricultural marketing. The research has been conducted in two phases—identification of predictors based on a scholarly articles' review and generating the predictive function between the identified predictors and adoption using the multilayer perceptron neural networks method. A case study of the electronic national agriculture market platform is undertaken to understand the predictors in the context of a real project. The study has revealed significant predictors of priority (high to low) as fast transaction cycles, higher prices, easier to use, infrastructure availability, trust, social influence, low transaction costs and customer care. Post-adoption, the user expects to get benefits like transparency, quick trade settlement, reduced transaction cost, expanding the market reach, and increase in the product price realized by the farmer. The improvement in the critical predictors of the electronic trading platform may help the de-facto national initiative to succeed. In new markets, the prediction framework using the neural network may be used to identify users with the propensity to adopt e-commerce or digital initiatives in the agricultural sector and proactively approach them to get a high number of transactions—an essential ingredient of success. The digital e-trading initiative is a growth catalyst of Agriculture 4.0 and may transform the agricultural supply chain for the better.

Keywords Digitalization · E-commerce · Electronic trading · Multilayer perceptron (MLP) · National agriculture market · Neural network

1 Introduction

The agricultural sector is economically and socially significant. It is the primary source of income for about 58% of Indians. The sector is witnessing a digital transformation. The digitalization is expected to lead to an efficient marketing system that lowers the transaction costs and generate more income for farmers while reducing the environmental hazards.

Agricultural commodities trade is shifting from oral auction mode to information and communication technology (ICT) based e-trade mode. The e-trading platforms have increased the speed and extent of the procurement process. It is also expected to increase the bargaining capability of the seller by having access to a broader range of buyers. Consequently, primary sellers (farmers) are expected to get a higher price when they sell the produce on the e-trading platform rather than at the farm-gate through a limited number of regional agents [4, 63].

The wholesale e-trading adoption in the Indian agriculture sector is still at a low level. On the national agriculture market (eNAM) platform launched in one thousand wholesale markets, only fourteen per cent farmers (16.6

✉ P. K. Suri
pksuri@dtu.ac.in

Sanjay Chaudhary
one.sanjay@gmail.com

[1] Vivekananda Institute of Professional Studies, Delhi School of Business, AU Block, Outer Ring Road, Pitampura, Delhi, India

[2] Delhi School of Management, Delhi Technological University, Shahabad Daulatpur, Main Bawana Road, Delhi, India

Published online: 12 January 2021

million) have registered, and less than 50% make a regular trade on the platform. There is an urgent need to improve the adoption of eNAM among the participants (farmers, traders, corporate agents) to deliver the intended benefits [20, 47].

This research article attempts to answer the question as to how the adoption of the agricultural wholesale e-trading platform can be improved. The neural network (NN) methodology is used to generate a statistical model for predicting the adoption of the digital e-trading platform.

The research article is organized as follows: Sect. 2 presents an understanding of wholesale e-trading adoption and the eNAM project in the context of the agricultural supply chain and marketing. Section 3 represents the proposed adoption framework. Sections 4 and 5, exhibit research objectives and the methodology used to achieve them. Section 6 explains the data analysis, followed by Sect. 7, which describes the findings and suggestions after that. Finally, the article is concluded in Sect. 8, with Sect. 9 mentioning about the limitations in the research work.

2 Literature review

An integrated and digitalized supply chain with quick processing has the potential to improve the effectiveness of the agriculture and food sector significantly. Information and communication technology (ICT) is cited as a vital facilitator of efficiency, agility, and performance in the supply chain and a vital component of Agriculture 4.0 [17]. Understanding the current agricultural supply chain and eNAM e-trading initiatives helps in understanding e-trading adoption and its predictors in the right context.

2.1 Wholesale e-trading adoption in agricultural marketing

The agricultural supply chain in India is full of no value-add intermediaries. The total margin mark-up in the agricultural supply chain is approximately 75%; thus, the primary producers receive less than 26% of the consumer price. The involvement of multiple intermediaries and information asymmetry has also resulted in massive wastage of 15–25% of the consumer value [15, 24, 40].

ICT based wholesale e-trading is expected to improve the current Indian agricultural supply chain (Fig. 1). The e-trading platform reduces operational cost, fasten the process, and improve on information quality. It enhances the performance of purchasing rights as it helps them delivered with all rights, e.g. price, quality, quantity, source, and time [48, 50, 59, 64].

Initially, e-trading implementation was rarely attempted on agricultural-related websites in India. The Indian agriculture websites were limited only to information exchange. Most of the initiatives lacked e-trading of agricultural produce or service provision. Even today, most of the successful private e-trading (ITC e-Choupal, eKutir, Agribazaar) are concentrated in a small area of the country or meet the need of enterprise partners.

The e-trading system is comparatively better than other prevailing systems, e.g. open auction, manual tender, and direct sale. Open auction may have a scope of trader collusion whereas manual tendering is prone to, alteration of quotes, and entry errors. The direct sale system reduces the bargaining power of a farmer due to lack of competition. The e-trading system is transparent and fast. Due to real-time transparency, a farmer may get a higher price [4, 11, 16, 42].

The Indian government has put in its efforts as per the national policy on information and communication technology (ICT) in agricultural extension. The need is for continuous improvement, in information exchange, promoting new farming techniques, restructuring sales and marketing via national commodity e-trading platform, and quality input supply in the distribution method, e.g. mKRISHI, AGMARKNET. The improvement process also calls for a regular impact assessment [2, 41, 52, 66].

2.2 National agriculture market (NAM)

The eNAM—a national level digital initiative launched in 2016—is so far implemented in 1000 wholesale markets by May 2020, with the registration of 16.6 million farmers, 131 thousand traders, 73 thousand commission agents and one thousand farmer producer organizations. It is expected to reach 2477 APMC market-yards and 4843 APMC-regulated sub-market yards by March 2022 [29, 43].

The digital eNAM initiative is a significant shift in the current ways of marketing in India's APMC markets [53]. In eNAM, most trade process activities are online in digital form (Fig. 2). The standard tradable quality parameters are developed for 175 agricultural commodities, bought-sold-paid via online transactions [43].

After eNAM implementation, some early results indicate that farmers are increasingly becoming quality conscious. Online transactions, electronic fund transfers and online banking literacy have increased. Now, the Mandi records are computerised, and better transparency of transaction flow is achieved. It allows buyers and traders to bid virtually in real-time, while the seller keeps track of his bid through a mobile application. APMC markets' fee revenues have increased due to increased online tendering. Farmers also gain in terms of price improvements due to competition, accurate electronic scale weights connected to

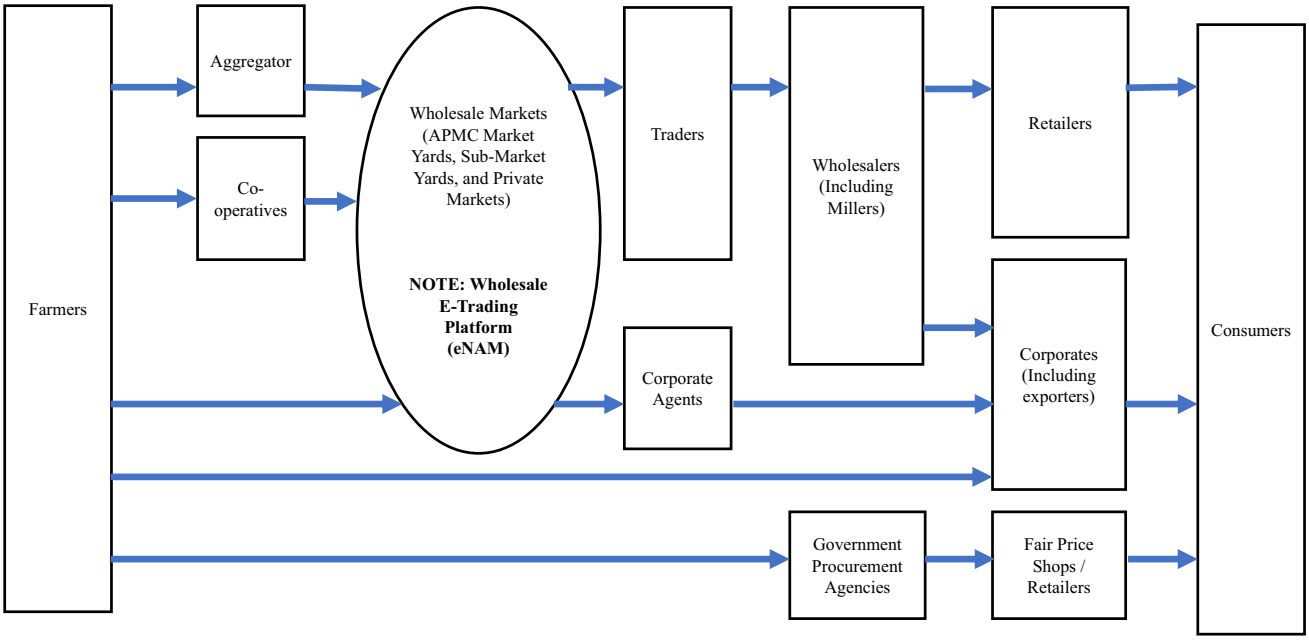


Fig. 1 Typical AGRI supply chain

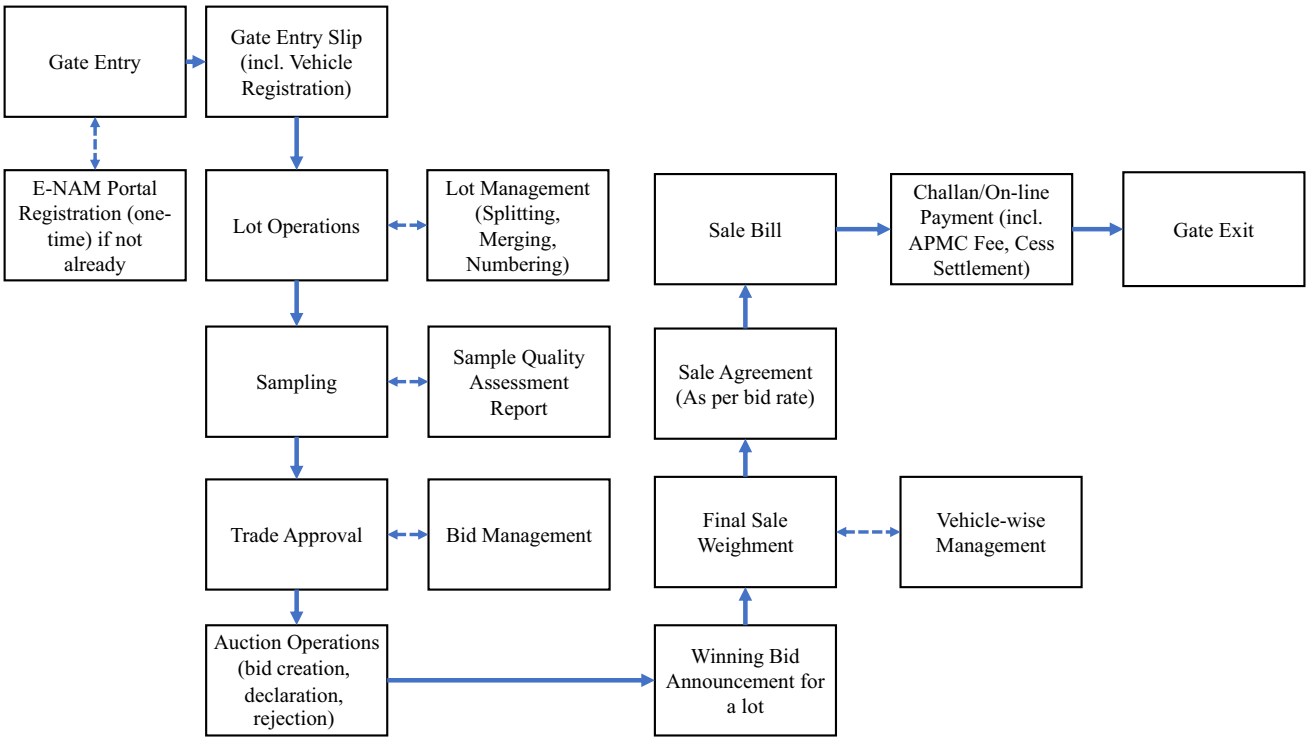


Fig. 2 eNAM trading process

eNAM portal, and reduction in inter-market marketing costs. Now, farmers get a clear sale bill of transparent e-trading, experience less trader collusion, and get the quick payment cycle of 1 day compared to the earlier duration of 1–2 weeks [33].

The eNAM project progress is, however, not at par with the targets set in its Vision Document [26, 35, 36]. There is an urgent need to promote the adoption factors and remove the emerging constraints during the implementation and adoption of eNAM through active participation in the

national agricultural market by local and state-level stakeholders [54].

3 Proposed prediction framework

The proposed wholesale e-trading adoption framework for agriculture in India (Fig. 3) is based on a review of the scholarly articles. Based on the assessment, the eight factors (constructs) are conceptualised to be significant direct or indirect determinants of wholesale e-trading adoption in Indian agricultural markets. The constructs manifestation in the respondent sample group of farmers in the agricultural sector is different from other Industry groups. The number of constructs, factor loading, structural coefficients may be different.

The adoption framework is distinct, as wholesale e-trading for agricultural marketing is different from other sectors due to lacking support infrastructure, the low literacy level of farmers, and little ICT awareness among farmers and traders. The research is conducted in a unique B2B context, wherein adoption factors are yet to be adequately addressed. The technology adoption factors are considered in a broader societal context, and due consideration is given to service-dominant logic [7].

The factors in the research framework are listed in Table 1.

4 Research objectives

The objectives of this research are:

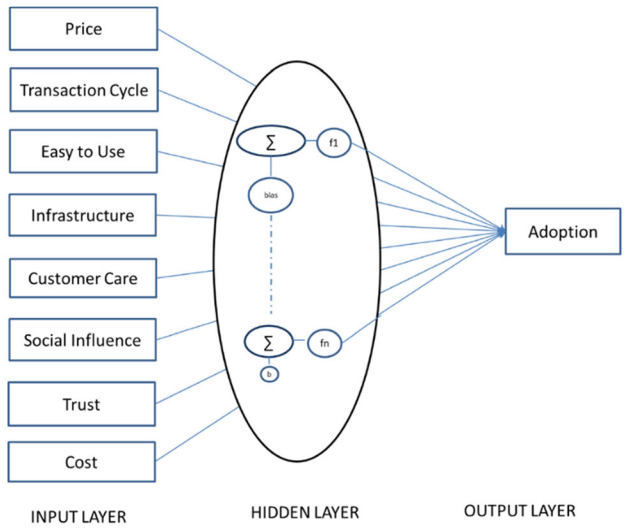


Fig. 3 Conceptual research framework

1. To validate the prediction framework for wholesale agricultural e-trading adoption in the context of National Agriculture Market.
2. Using the prediction framework, suggest possible ways to increase the wholesale agricultural e-trading adoption.

5 Methodology

The research design is exploratory-cum-descriptive. The case study of the eNAM wholesale e-trading platform and the opinion survey of its users provide a means to validate the theoretical basis of the research framework. The possible ways to increase third-party e-trading platform (eNAM) adoption are based on insights gained from the literature review and the learning from unstructured interviews.

Simple random sampling is used to draw the sample. The respondents are randomly chosen from the list of registered users on the eNAM platform. The sampling unit is a farmer who is the intended beneficiary of using eNAM for e-trading. It is expected that as a user, the respondent has an opinion about its functioning. Data were collected by conducting a customised questionnaire survey of 441 users across five wholesale markets in northern India between May and November 2019. The data for each factor (predictor variable) was collected using a Likert scale (scale of 1–7) based questionnaire. Level 1, meaning 'Negligible' to level 7, saying 'To a huge extent'. The fifty pilot questionnaires were tested for reliability with an acceptable Cronbach alpha at 0.9 [38].

The neural network (NN) is used as a classification tool to generate a statistical model for predicting the adoption of e-trading platform. The neural network creates a model for the learning patterns that can be reproduced in other instances, thus suitable in statistical analyses for this study. The neural network method is chosen because it can model non-linear relationship without imposing any restriction on input and residual distributions. The model developed in this study can be generalised and predict on unseen data in future. The classification of new data using the neural network is simple compared to factor analysis. The quantitative technique is used, due to objectivity and statistical significance, rather than qualitative technique [6, 55].

The neural network using actual data has been used in the field of agriculture mainly for crop suggestion, fertiliser advise, predict crop yield, predict maturity date, forecast price, differentiate weed from crops, forecast water resources, predict the spatial water distribution, forecast rainfall, predict nutrition level, predict frost formation,

Table 1 Variables in adoption framework

Factor	Definition	References
Price (P)	The level to which a user perceives that the e-trading platform benefits to users in terms of a better price	[4, 11, 19, 36, 42]
Transaction cycle (TC)	The level to which a user perceives that the e-trading platform benefits to users in terms of speed of the transaction	[14, 25, 37, 60, 62]
Easy to use (EU)	The ease in the learning and use of the e-trading platform for the user. The training effort is also included	[30, 57, 62]
Infrastructure (I)	The user perception that the infrastructure, including quality testing laboratories, bidding hall, and logistics support are available to support the use of the e-trading platform	[12, 46, 58, 62]
Customer care (CC)	The level of user perception that customer care and dispute resolution complement the use of the e-trading platform	[13, 58]
Social influence (SI)	The perception about the level to which the community of close friend, leading farmer, associated trader, and community leader encourages or promote the e-trading platform	[1, 56, 62]
Trust (T)	The farmer's confidence in the reliability of farmer-to-business e-trading platform concerning trade, technology, and information	[8, 9, 21, 28, 31, 36, 45, 65]
Cost (C)	The transaction cost is consisting of search cost, coordination cost, motivation cost, and commitment cost	[18, 32, 34, 51, 64, 65]

classification of disease, and estimate soil moisture [3, 22, 23, 44, 49, 61].

The multilayer perceptron (MLP) structure of the predictive model is used in the study to analyse the function of independent input variables (factors) that minimise the prediction error of the dependent variable 'adoption'. To further judge the importance of independent input variables (factors) concerning the adoption of eNAM, the sensitivity analysis is used. The data analysis tools used are Excel, R and SPSS version 20.

6 Data analysis

The valid responses (no missing value) used for analysis are 441 as per the sample summary (Table 2). The dataset is partitioned for case assignment into 67–33% among training, and testing sample.

The specifications are shown in the neural network information (Table 3) summary. In the multilayer perceptron (MLP) neural network, the dependent variable "adoption" (A), as well as the independent predictor variables, are ordinal (categorical). The eight input predictor variables are infrastructure (I), transaction cycle (TC), easy to use (EU), social influence (SI), trust (T), cost (C), customer care (CC) and price (P) (Fig. 3). The input variables in the input layer are rescaled using the standardised method, which subtracts the mean and divide by the standard deviation ((x − mean)/s).

Table 2 Sample summary

		N	Per cent
Sample	Training	295	66.9%
	Testing	146	33.1%
Valid		441	100.0%
Excluded		0	
Total		441	

The weighted connections allow data to move between input, hidden and output layer. In each layer, the node accepts data from the previous layer and calculates a weighted sum of all its net inputs. A transfer function is applied to the weighted value, to generate the node output. The hidden layer uses the hyperbolic tangent activation function (for automatic architecture), takes inputs, and convert them into the range (− 1, 1). In contrast, the SoftMax activation function in the output layer converts processed input into (0, 1).

The number of hidden layers (containing un-observable network nodes) is '1', and the number of units in this layer is seven as determined automatically by the estimation algorithm. The output dependant variable is adoption (A) with 2 number of units (0, 1) in the output layer. The option to manually increase the number of hidden layers or alter the number of neurons, as an attempt to increase accuracy [10, 27] is not exercised.

The final neural network model summary (Table 4) has information about the results of training and applying the trained network to the testing sample. The learning process is via a training method back-propagation (BP), which involves adjusting the weights associated with the transfer functions between neurons comparing neural network

Table 3 Network information

Input layer	Factors	1	Infrastructure
	Covariates	1	Transaction Cycle
		2	Price
		3	Easy to use
		4	Social influence
		5	Trust
		6	Cost
		7	Customer care
	Number of units[a]		14
	Rescaling method for covariates		Standardised
Hidden layer(s)	Number of hidden layers		1
	Number of units in hidden layer 1[a]		7
	Activation function		Hyperbolic tangent
Output layer	Dependent variables	1	Adoption
	Number of units		2
	Activation function		SoftMax
	Error function		Cross-entropy

[a]Excluding the bias unit

Table 4 Neural net model summary

Training	Cross-entropy error	65.744
	Percent incorrect predictions	6.8%
	Stopping rule used	One consecutive step(s) with no decrease in error[a]
	Training Time	0:00:00.40
Testing	Cross-entropy error	43.074
	Percent incorrect predictions	8.9%

Dependent variable: adoption

[a]Error computations are based on the testing sample

output with observed data. The intent, on this feed-forward neural network, is to minimise the cross-entropy error.

The cross-entropy error (equivalent to sum-of-squares error) based on the testing sample is low. The incorrect predictions for the testing sample are small at 8.9%. The batch training method (update the coefficient estimates only after passing all training data records) is used with the optimisation algorithm as the scaled conjugate gradient. The optimisation algorithm in the training sample stopped because the error has converged.

The classification data (Table 5) shows the practical results. For each case, the predicted response is 1 (adoption) if for that case predicted pseudo-probability is more significant than 0.5. Of the training sample cases used to create the model, 204 of the 211 people who have adopted eNAM are classified correctly. 71 of the 84 non-adopters are classified correctly. Overall, 93.2% of the cases in the training sample is classified correctly. Thus, a suitable model has been generated.

The testing sample is used to test the model; here, 91.1% of the cases are correctly classified by the model. Overall,

Table 5 Classification

Sample	Observed	Predicted		
		0	1	Percent correct
Training	0	71	13	84.5%
	1	7	204	96.7%
	Overall Percent	26.4%	73.6%	93.2%
Testing	0	40	10	80.0%
	1	3	93	96.9%
	Overall Percent	29.5%	70.5%	91.1%

Dependent variable: adoption

the model is, in fact, correct about 9 out of 10 times. Since the model is used for identifying the valid adoption '1', which the model does for the 96.9% of the cases, the predictive model meets the purpose.

The ROC curve that plots the true positive rate against the false-positive rate is shown in Fig. 4. The area under the curve for both 0 "No" and "Yes" for adoption is

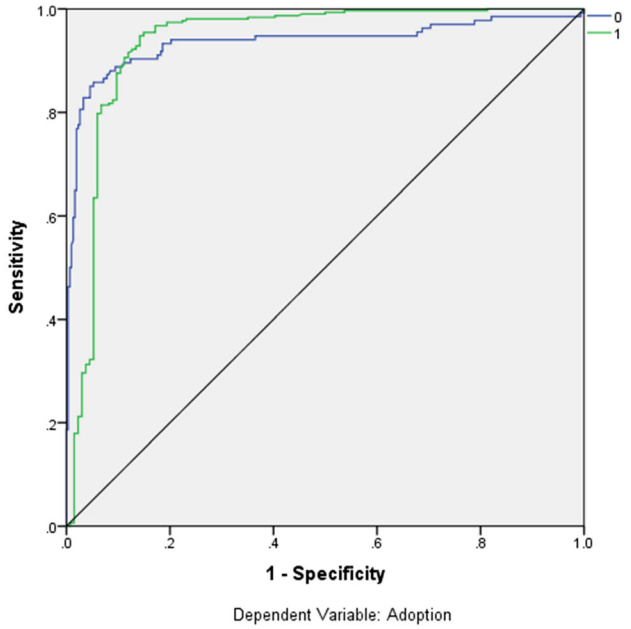


Fig. 4 ROC curve

satisfactory at 93.5% with a sensitivity of 96.9% (primary focus) and specificity of 80%.

The validation is done by comparing the proposed model [10] with other models using a second dataset, but on the same assumptions (Table 6). The proposed back-propagation multilayer perceptron neural network is found to be most accurate with 92.3% accuracy compared to a decision tree, support vector machine, and logistic regression.

7 Findings and suggestions

Using the multilayer perceptron procedure, the study has constructed a model for predicting the probability that a given user (farmer) will adopt the wholesale e-trading platform. All predictor variables are found to be relevant. The relationships with the dependent variable "adoption" have the non-zero synaptic weights.

The model may be applied to different users/new platform/market to predict user adoption based on an opinion survey questionnaire. Preferably, the predictor variables (factors) in the adoption prediction model may be improved to increase the adoption of eNAM.

To further explore the independent variables, the sensitivity analysis to check the independent variable importance is conducted. The top variables (factors) in term of their importance in determining the neural network for the adoption of eNAM (wholesale e-trading) are listed in Table 7. The eNAM market managers may put efforts to improve them in the same priority.

The synthesis of analysis conducted, and discussions held with experts, and farmers during the study at the wholesale markets have led to the following suggestions in order of their importance:

- Fast 'transaction cycle' is a critical adoption factor for farmers. The proliferation of mobile application and digital payment is expected to be the right step. The inter-market e-trade (with the dispute resolution at the trading APMC), maybe streamlined further where the total flow time of the trading process, including digital payment is reduced to less than a working day hour.
- Many of the responding farmers believe that the eNAM has increased the 'Price' realisation from their commodity sale. Besides, there may be provision made for easy access to bank credit for farmers registered and friendly credit terms made available to traders buying-selling on eNAM so that the digital system can compete against offline un-official credit system of commission agent.
- The perception 'easy to use' may be improved by making the website easy to navigate, with multi-lingual content. The mobile application is popular among users, but it needs to improve on its API interface and inter-operation with other post-harvest partner and government applications.
- The 'infrastructure' issues may be addressed by ensuring the availability of the amenities, e.g., quality assaying labs with robust mechanism, computer terminals for users, electronic weighing, and trade-rooms with broadband internet connectivity which are maintained in bare sub-optimal conditions. Also, ATMs may be made available in the market yard for encashing the online payments.
- The local agri-community engagement may improve 'Trust'. The interaction of e-trading platform officials with the volunteers and the farmer's group (farmer producer organisations, co-operatives, self-help groups) need to be enhanced as well as organising regular training camp may help. As for the inter-state trade, the dispute settlement process may be detailed and formalised.

Table 6 Comparison with other methods

Model type	Overall accuracy (percent)
Decision tree	91
Support vector machine (SVM)	92
Logistic regression	90
Neural network	92.3

Table 7 Factor importance

Rank	Factor	Importance	Normalised importance (%)
1	Transaction Cycle	0.253	100.00
2	Price	0.225	88.70
3	Easy to use	0.128	50.70
4	Infrastructure	0.103	40.70
5	Trust	0.094	37.00
6	Social influence	0.085	33.50
7	Transaction cost	0.068	26.80
8	Customer care	0.044	17.40

- The 'social influence' may be improved by promoting awareness among the farming community. Making use of influencer users and agriculture extension officers to facilitate decision making will build a favourable environment for e-trading and other similar initiatives.
- The low 'transaction cost' is a critical adoption factor for inter-market trade. The high-transaction cost, along with the existing variation between state taxes and market fees, is hindering cross-state transactions of agricultural produce. If e-trading can help in lowering the aggregate taxes and market fees with some level of uniformity across states, it may result in very high user adoption of e-trading.
- The 'customer care' may be improved through the dedicated contact centre and on-ground staff. The implementing agency may go for required user capacity building through regular training on digitalization.

8 Conclusion

The ICT based e-trading platform is an ambitious and priority intervention in the Indian agriculture sector. The digital e-trading technology is the growth catalyst and may transform the agriculture supply chain for better.

The study is among the very few that analyses the adoption of e-trading in India. It is also among the very few that extends the applicability of the neural network in predicting the wholesale e-trading adoption in the agricultural marketing area. The adoption prediction framework is expected to help service provider adopt research-based suggestions aimed at increasing adoption of digital wholesale e-trading in terms of improved organisational readiness, and implement strategies aimed at strengthening the predictors of adoption.

The fast transaction cycle, higher price, easy to use, infrastructure availability, and trust are identified as the top five significant predictors of the adoption of wholesale e-trading initiative (eNAM).

The Indian policymakers are now expanding the digital e-trading platform to 7500 wholesale markets by 2022 on a one nation one market concept. The prediction framework may be used to identify the key predictors or enablers of the e-trading (e-commerce/digital) platform into new markets or geographies, and thus rapidly improve upon them to succeed in a fast-changing situation [5]. The prediction model/framework may be used to score the opinion survey data in the new market, identify users with a propensity to adopt e-trading, and then offer them some early movers offer to ramp up the initial adoption stage. Also, the non-adopters may be approached more proactively for counselling.

The digital e-trading may be an essential precursor to other new-generation technologies used in the agricultural supply chain, e.g. blockchain, farm data analytics, artificial intelligence, and biotechnology.

The improved adoption and performance of the wholesale e-trading platform may go a long way in solving the primary problem of fragmentation and inefficiency in the agriculture supply chain. No wonder, the digital eNAM e-trading platform, is one of the critical interventions by the Indian government aimed at doubling the farmers' income by 2022.

9 Research limitations

The study is cross-sectional, as it measures perceptions at one point in time. However, opinions may change over time, with experience. Thus, a longitudinal study, using structural equation modelling (SEM) carried over a more extended period (with several perceptions, measured over time) may add some more insights on causality and inter-relationships [39, 62]. It may be explored whether 'Trust' is a predictor or mediator/moderator variables. Similarly, 'social influence' may be a moderator of the 'price', and 'easy to use' of the eNAM platform. Also, the research scope may be enhanced through a cross-country study or cross-sector, to add/drop predictors considered in this study.

Author contributions The submission is original, unpublished and not submitted simultaneously to any journal/conference/seminar for consideration of publication/presentation.

Availability of data and material We have used primary sources of information/information from our field research. We have used secondary sources of information as available in the public domain, and none of these is copyrighted.

References

1. Ajzen I (1991) The theory of planned behaviour. Organ Behav Hum Decis Process 50(2):179–211
2. Anupindi R, Sivakumar S (2007) Supply chain reengineering in agri-business. In: Lee HL, Lee CY (eds) Building supply chain excellence in emerging economies. Springer, New York
3. Banerjee T, Monalisa M, Debnath NC, Prasenjit C (2019) Implementing E-commerce model for agricultural produce: a research roadmap. Period Eng Nat Sci 7(1):302–310
4. Banker R, Mitra S, Sambamurthy V (2011) The effects of digital trading platforms on commodity prices in agricultural supply chains. MIS Q 35(3):6–34
5. Basahel A, Mohammad Y (2017) Measuring the success of e-government of Saudi Arabia. Int J Inf Technol 9:287–293
6. Betchoo H, Halkhoree R, Santally MI, Sunkur RK (2019) Assessment tool for e-services take-up (ATEST): a small island states context. Int J Inf Technol 11(2):187–201
7. Bettencourt LA, Lusch RF, Vargo SL (2014) A service lens on value creation. Calif Manag Rev 57(1):44–66
8. Bisen J, Ranjit K (2018) Agricultural marketing reforms and e-national agricultural (e-NAM) in India: a review. Agric Econ Res Rev 31:167–176
9. Casalo LV, Guinaliu FC (2011) The generation of trust in online services and product distribution: the case of Spanish electronic commerce. J Electron Commer Res 12(3):199–2015
10. Chakraborty S, Jana GC, Kumari D, Swetapadma A (2020) An improved method using a supervised learning technique for diabetic retinopathy detection. Int J Inf Technol 12(2):473–477
11. Chand R (2016) E-platform for national agricultural market. Econ Political Wkly 51(28):15–18
12. Cigolini R, Cozzi M, Perona M (2004) A new framework for supply chain management: conceptual model and empirical test. Int J Oper Prod Manag 24(1):7–41
13. Compeau DR, Higgins CA (1995) Application of social cognitive theory to training for computer skills. Inf Syst Res 6(2):118–143
14. Compeau DR, Higgins CA, Huff S (1999) Social cognitive theory and individual reactions to computing technology: a longitudinal study. MIS Q 23(2):145–158
15. CPPR (2016) Doubling farmers' income—are the claims true. Centre for Public Policy Research (CPPR). https://www.cppr.in/centre-for-comparative-studies/doubling-farmers-income-are-the-claims-true. Accessed 15 Apr 2020
16. Dey K (2016) National agricultural market-rationale, rollout, and ramifications. Econ Political Wkly 51(19):35–39
17. FAO (2020) Digital innovation for promoting Agriculture 4.0 in the Near East and North Africa. Food and Agriculture Organization of the United Nations. http://www.fao.org/3/nd262en/nd262en.pdf Accessed 10 Oct 2020
18. Garicano L, Kaplan SN (2001) The effects of business-to-business e-commerce on transaction costs. J Ind Econ 49(4):463–485
19. Goyal A (2010) Information, direct access to farmers, and rural market performance in Central India. Am Econ J Appl Econ 2(3):22–45
20. Hindu (2019) Just 14% of farmers registered on eNAM platform. The Hindu Business Line. https://www.thehindubusinessline.com/economy/agri-business/just-14-of-farmers-registered-on-enam-platform/article28363454.ece. Accessed 10 July 2019
21. Jayashanker P, Nilakanta S, Johnston WJ, Gill P (2018) IoT adoption in agriculture: the role of trust, perceived value, and risk. J Bus Ind Market 33(6):821
22. Jha K, Aalap D, Poojan P, Manan S (2019) A comprehensive review of automation in agriculture using artificial intelligence. Artif Intell Agric 2:1–12
23. Jiri S, Oldrich T (2011) Agricultural data prediction by means of neural network. Agric Econ Czech 57(7):356–361
24. Kaur T (2015) Use of e-marketing in Indian agriculture sector to manage distribution challenges. Anvesha 8(1):1–6
25. Kilpatrick J, Factor R (2000) Logistics in Canada survey: tracking the year 2000 supply chain issues and trends. Mater Manag Distrib 45(1):16–20
26. Krishnamurthy M (2019) Back-end first: a National Agenda for India's Agricultural Markets. Centre for Policy Research.https://www.cprindia.org/policy-challenge/7882/economy-the-welfare-state. Accessed 20 Oct 2019
27. Kumari D, Kilam S, Nath P, Swetapadma A (2018) Prediction of alcohol abused individuals using artificial neural network. Int J Inf Technol 10(2):233–237
28. Kuttainen C (2005) The role of trust in B2B electronic commerce: evidence from two e-marketplaces. Dissertation. Luleå University of Technology
29. MOAFW (2020) National Agriculture Market. The Government of India. http://www.enam.gov.in/NAM/home/index.html. Accessed 07 Aug 2020
30. Moore GC, Benbasat I (1991) Development of an instrument to measure the perceptions of adopting an information technology innovation. Inf Syst Res 2(3):192–222
31. Moorman C, Zaltman C, Deshpande R (1992) Relationships between providers and users of market research: the dynamics of trust within and between organisations. J Mark Res 29:314–328
32. Mustaquim M (2017) Agriculture marketing reforms: benefiting farmers. https://www.ruralmarketing.in/industry/agriculture/agriculture-marketing-reforms-benefiting-farmers. Accessed 12 Nov 2019
33. NABARD (2018) Status of marketing infrastructure under electronic national agriculture markets: a quick study. National Bank for Agriculture and Rural Development. https://www.nabard.org/auth/writereaddata/tender/0904195131 Status of Marketing Infrastructure under the Electronic National Agriculture Market. A Quick Study.pdf. Accessed 21 Oct 2019
34. Nabhani IA, Daryanto A, Machfud AR (2016) Mobile broadband for the farmers: a case study of technology adoption by cocoa farmers in Southern East Java, Indonesia. AGRIS Online Pap Econ Inform 8(2):111–120
35. Naik AR (2019) National Agri Marketplace eNAM growth slower than expected, reveals data. Inc42. https://inc42.com/buzz/national-india-agri-marketplace-enam-growth-slower-than-expected-reveals-data/. Accessed 16 Dec 2019
36. Nirmal R (2017) E-NAM: why inter-state trade is still a non-starter. Indian J Agric Market 31(3):73–81
37. Nirmal R (2019) How Suryapet has changed after eNAM. Businessline. https://search.proquest.com/docview/2263642337?accountid=173681. Accessed 07 July 2019
38. Nunnally JBL (1994) Psychometric theory. McGraw Hill Higher INC, New York

39. O'Neill M (2003) The influence of time on student perceptions of service quality: the need for longitudinal measures. J Educ Adm 41(3):310–325
40. Patnaik G (2011) Status of agricultural marketing reforms. Indira Gandhi Institute of Development Research. http://agritech.tnau.ac.in/amis/pdf/rs/Status_Agrl_Marketing_Reforms.pdf. Accessed 16 Jan 2020
41. Patnaik I, Roy S, Shah A, Burman A (2018) Diagnosing and overcoming sustained food price volatility: enabling a national market for food. https://www.nipfp.org.in/media/medialibrary/2018/07/WP_236.pdf. Accessed 5 July 2020
42. Pavithra S, Gracy CP, Saxena R, Patil GG (2018) Innovations in agricultural marketing: a case study of the e-tendering system in Karnataka. Agric Econ Res Rev 31(1):53–64
43. PIB (2020) Thirty-eight new MANDIS integrated with eNAM: The pan-India electronic AGRI-produce trading portal reaches milestone of 1000 MANDIS across 18 States and 3 UTs. Press Information Bureau. https://pib.gov.in/PressReleasePage.aspx?PRID=1624083. Accessed 15 May 2020
44. Ravichandran G, Koteshwari RS (2016) Agricultural crop predictor and advisor using ANN for smartphones. IEEE 1–6
45. Ridings C, Gefen D, Arinze B (2002) Some antecedents and effects of trust in virtual communities. J Strateg Inf Syst 11:271–295
46. Russel D, Hoag AM (2004) People and information technology in the supply chain: social and organisational influences on adoption. Int J Phys Distrib Logist Manag 34(2):102–122
47. Sajwan R (2020) COVID-19: is e-NAM portal capable of supporting farmers. DowntoEarth. https://www.downtoearth.org.in/news/agriculture/covid-19-is-e-nam-portal-capable-of-supporting-farmers–70294. Accessed 08 Apr 2020
48. Shirzad SR, David B (2013) A systematic literature review of flexible e-procurement marketplace. J Theor Appl Electron Commer Res 8(2):49–70
49. Saurabh S, Sindhu D (2017) Information dissemination using computer and communication technologies for improving agriculture productivity. Int J Emerg Trends Technol Comput Sci 6(6):143–155
50. Smart A, Harrison A (2003) Online reverse auctions and their role in buyer–supplier relationships. J Purch Supply Manag 9(5–6):257–268
51. Subramaniam C, Shaw MJ (2002) A study of the value and impact of B2B e-commerce: the case of web-based procurement. Int J Electron Commer 6(4):19–40
52. Suri PK (2005) NICNET based Agricultural Marketing Information Network—A Farmers Centric Portal on Agricultural Marketing in India and a Step towards Globalising Indian Agriculture. Agric Market J 45(4):2–11
53. Suri PK, Sushil (2017) Strategic planning and implementation of e-governance. Springer Science and Business Media, Singapore
54. Suri PK (2018) Towards an effective agricultural e-trading system in India. In: Connell J, Agarwal R, Sushil DS (eds) Global value chains, flexibility and sustainability. Springer, Singapore
55. Tafeit E, Moller R, Karl S, Gilbert R (2000) Artificial neural networks compared to factor analysis for low-dimensional classification of high-dimensional body fat topography data of healthy and diabetic subjects. Comput Biomed Res 33(5):365–374
56. Taylor S, Todd PA (1995) Understanding information technology usage: a test of competing models. Inf Syst Res 6(4):144–176
57. Thompson RL, Higgins CA, Howell JM (1991) Personal computing: toward a conceptual model of utilisation. MIS Q 15(1):124–143
58. Tomar G, Chauhan GS, Panigrahi PK (2016) Feasibility of m-governance in agriculture: insights from a multimodal study in rural India. Transform Gov People Process Policy 10(3):434–456
59. Tripathy S, Aich S, Chakraborty A, Lee GM (2016) Information technology is an enabling factor affecting supply chain performance in Indian SMEs: a structural equation modelling approach. J Model Manag 11(1):269–287
60. Tyndall G (1998) Supercharging supply chains: new ways to increase value through global operational excellence. Wiley, New York
61. Varun R, Neema N, Sahana HP, Sathvik A, Muddasir M (2019) Agriculture commodity price forecasting using Ml techniques. Int J Innov Technol Explor Eng 9(25):729–731
62. Venkatesh V, Morris MG, Davis GB, Davis FD (2003) User acceptance of information technology: toward a unified view. MIS Q 27(3):425–478
63. Verma S, Chaudhuri R (2008) Creating customer satisfaction and profitable value chain with e-commerce in rural India. Int J Bus Insights Transform 2(1):51–63
64. White AEDM, Daniel EM, Mohdzain M (2005) The role of emergent information technologies and systems in enabling supply chain agility. Int J Inf Manag 25(5):396–410
65. Wiyada T, Vatcharaporn E, Donyaprueth K (2018) A framework to identify factors affecting the performance of third-party B2B e-marketplaces: a seller's perspective. Electron Mark 28:129–147
66. Yamin M, Mattar R (2016) e-Government in Saudi Arabia—an empirical study. Int J Inf Technol 8(1):944–949

cogent engineering

Cogent Engineering

Taylor & Francis
Taylor & Francis Group



ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/oaen20

Geotechnical parameter assessment of sediment deposit: A case study in Pasakha, Bhutan

Karma Tempa , Nimesh Chettri , Raju Sarkar , Sunil Saha , Lily Gurung , Tshering Dendup & Bhawani Shankar Nirola |

To cite this article: Karma Tempa , Nimesh Chettri , Raju Sarkar , Sunil Saha , Lily Gurung , Tshering Dendup & Bhawani Shankar Nirola | (2021) Geotechnical parameter assessment of sediment deposit: A case study in Pasakha, Bhutan, Cogent Engineering, 8:1, 1869366, DOI: 10.1080/23311916.2020.1869366

To link to this article: https://doi.org/10.1080/23311916.2020.1869366

© 2020 The Author(s). This open access article is distributed under a Creative Commons Attribution (CC-BY) 4.0 license.

Published online: 04 Jan 2021.

Submit your article to this journal

Article views: 121

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=oaen20

cogent engineering



Geotechnical parameter assessment of sediment deposit: A case study in Pasakha, Bhutan

Karma Tempa, Nimesh Chettri, Raju Sarkar, Sunil Saha, Lily Gurung, Tshering Dendup and Bhawani Shankar Nirola

Cogent Engineering (2021), 8: 1869366

cogent oa

CrossMark

Tempa et al., *Cogent Engineering* (2021), 8: 1869366
https://doi.org/10.1080/23311916.2020.1869366

cogent engineering

CIVIL & ENVIRONMENTAL ENGINEERING | RESEARCH ARTICLE

Geotechnical parameter assessment of sediment deposit: A case study in Pasakha, Bhutan

Karma Tempa[1*], Nimesh Chettri[2], Raju Sarkar[2], Sunil Saha[3], Lily Gurung[2], Tshering Dendup[2] and Bhawani Shankar Nirola[2]

Received: 07 September 2020
Accepted: 22 December 2020

*Corresponding author: Karma Tempa, Civil Engineering and Architecture Department, College of Science and Technology, Royal University of Bhutan, Bhutan
E-mail: karmatempa.cst@rub.edu.bt

Reviewing editor:
Ramon Fernando Colmenares, Faculty of Engineering, Universidad Cooperativa de Colombia, Colombia

Additional information is available at the end of the article

Abstract: The research area falls at the downstream level of Bhalujhora landslide area in Alay which is located in Pasakha, Bhutan. Local Government authority proposed for construction of custom station between the commercial town Phuentsholing and Pasakha Industrial estate. However, downstream region which connects the core town to industrial estate are severely affected by the flash floods every year and the runoff causes huge sediment deposits leading risk for community and substantial damages to the infrastructures. The site is seasonally eroded and subsequently deposited by the sediments from upstream. Although, the formation level for infrastructure are proposed on the deposit layer, the assessment of soil parameters is necessary for design and planning to ensure the overall safety and long-term economic benefits. Through this study, the subsurface soil parameters are assessed by in-situ open pit investigation, geophysical survey and Standard Penetration Test (SPT), while the engineering properties of the sediment deposits were determined from laboratory tests. The field investigation was important to decide the suitability of foundation type and depth. Further, soil parameters were determined to classify the soil type, understand permeability characteristics,



Karma Tempa

ABOUT THE AUTHOR

Karma Tempa is Lecturer and Head of Department in Civil Engineering and Architecture Department at College of Science and Technology, Royal University of Bhutan, Phuentsholing, Bhutan. He is also Program Leader for B.E. in Engineering Geology program. He has MSc in Civil Engineering from Università degli Studi del Sannio, Benevento, Italy and B.E. in Civil Engineering from College of Science and Technology. He has worked as Construction Manager for Royal Government of Bhutan for public infrastructure development projects. He has continuously supported Construction Development Board (CDB) in training contractors to enhance quality in Bhutanese construction industry. He has worked as principal investigator for World Bank project, South Asia Subregional Economic Cooperation-SASEC project and R&D-based project with Druk Holding and Investment (DHI). His research interest includes landslide hazard vulnerability and susceptibility assessment, geotechnical site characterization, local seismic response studies, geohazards risk management and pavement geotechnology.

PUBLIC INTEREST STATEMENT

Geotechnical investigation is conducted to assess the subsoil characteristics for infrastructure development projects in the study area. The investigated physical and mechanical parameters are mainly required for design of foundations, retaining structures and site development works. This study assures safety and economic benefit in long run in terms of functional and structural stability which may be caused due to deformation or settlement of subsoil and risk associated to natural hazards. Also, this geotechnical investigation will ensure precautions to minimize damages, cost savings and reduce frequency of future repair and maintenance. This geotechnical investigation will form a part of Detail Project Report (DPR) for strategic planning, technical feasibility and suitability of the area which is a pre-requisite in construction industry. The experimental results and quantity estimate of sediment deposit will allow the relevant stakeholders to propose alternate use of the sediment in road works.

cogent oa

© 2020 The Author(s). This open access article is distributed under a Creative Commons Attribution (CC-BY) 4.0 license.

and predict shear strength parameters and compaction behaviour. The result shows that the sediments consist of mixture of natural aggregates and sand mixed with silt and clay. An estimate of sediment accumulation was carried out to determine the volume of the sediment deposited. The outcome will allow the relevant stakeholders to decide the management of the risk from this future disaster and propose alternative use of the sediment materials in road works.

Keywords: landslide; trial pit; geophysical survey; laboratory test; sediments; industrial estate; Bhutan

1. Introduction

Soil settings of an area have a crucial role in construction of the building. Evaluation of the engineering properties of the subsoil at the project site is important for the generation of appropriate input data for the design and construction of foundations for the proposed structures, according to various researchers (Laskar & Pal, 2012; Oke & Amadi, 2008). The proper design and implementation of civil structures avoid an adverse effect on the environment or structural damage or post—project problems (Ngah & Nwankwoala, 2013; Oghenero et al., 2014; Youdeowei & Nwankwoala, 2013). In this regard, the data about the surface and sub-surface features is important for structural design and construction planning. The objective of this study is to determine the long-term structural safety of the infrastructures, since the proposed site is earmarked for development of land custom station which will cater to the major export and import transaction between the third country and the largest industrial hub of Bhutan—Pasakha. However, the site is characterized by layers of sediment deposit due to monsoon flash floods and runoff from upstream landslide (Figure 2) which requires appropriate engineering intervention to assess the sub-soil conditions and to further analyze the soil parameters for the use in planning and design. For geotechnical engineers such deposits are extremely challenging to deal with. Construction of the custom office on the deposited site requires information of the geotechnical conditions. When the bases of any building are founded on compressible earth, this results in settling. Awareness of the pace at which soil contraction occurs is crucial from proposed design. Soil properties such as plasticity, compressibility or soil hardness often influence building design. During the monsoon season, the area along the Pasakha highway in Phuentsholing outnumbers the frequency of landslide compared to any other regions in the country, therefore alarming the local municipal authority. Soil erosion during monsoon and fragile geology of the region are the main causes of landslides. In 1999, a slope failure occurred in Bhalujhora, Pasakha which consequently caused significant damage to agricultural lands and disrupting transportations network within the localities (Chettri, Sarkar et al., 2019). One of the significant factors for the rise and fall of the frequency of natural disasters is the existence of an interaction between the physical environment and socio-economic conditions (Mata-Lima et al., 2013). Transportation of sediment is generally a combination of fluid movement where the sediments are under the action of gravity or they are entrained. Catastrophic failure of buildings located near the coastal area is one of the effects of this problem (Zawawi et al., 2019). Ali and Fakhraldin (Ali & Fakhraldin, 2016) analyzed the behaviour of sediments under the action of structural loading in order to save costs of design and construction, after conducting a parametric study on sediments to prognose the geotechnical properties, subsurface materials' nature and stratigraphy. For geotechnical engineers and geologists, the site investigation is the most important aspect of the work; without an accurate depiction of the sub surface conditions and the geology of the site, subsequent analysis is an assumed work (Tempa et al., 2016). In this study, soil parameters were extracted from geotechnical site investigation (GSI) using trial pits sampling and geophysical methods and the laboratory tests of the collected soil samples from the study location. The quantified and detailed information of the parametric analyses and statistical inference are presented in this paper. Statistical analysis of the rainfall

pattern of the study area was conducted using 24 years of precipitation data (1996 to 2019) and quantity of sediment deposit were estimated.

2. Study area

The total population of Phuentsholing city in 2017 is 27,658 considering 3% growth rate and 30,518 considering 5% growth rate from 2012. Based on the land use pattern, the population of Pasakha in 2017 is 10,124 with the growth rate of 3% to 5% per year, in average of 4% annually (Phuentsholing Thromde, 2002).

The study area stretches along with the Buxa group of Lesser Himalayan towards the Greater Himalayan rocks. The geology of the study area is presented in Figure 3. The area considered falls in the structurally fragile crystalline weak sheet of bed rock mostly comprising paragneiss, quartzite, schists of Precambrian, orthogneiss and early Paleozoic ages and Miocene leucogranitic intrusions.

The bordering geology to the study area consists of different aged Tethyan sediments (Paleozoic to Eocene) predominated by limestone, shales and sandstones (Dini et al., 2019). These ideally mimic the fragile phyllite and quartzite that easily disintegrates during rainy season. Predominantly, Carbonaceous phyllite is the dominant rock type in the area followed by variegated phyllite and a thin band of greenish grey quartzite which gives rise to reddish brown to greyish brown residual soil (Department of Geology and Mines, 2010).

3. Methodology

In the present study for assessing the surface and sub-surface soil geophysical setting to establish the custom station Open pit investigation, Standard Penetration Test (SPT), geophysical survey, laboratory test of soil properties (Table 1) and measurement of sediment volume were carried out (Figure 4). The location of four trial pits for geotechnical investigation of properties of the sediments is presented in Figure 1(b). The detailed examination of the conditions of underlying soil and rocks for shallow depth can be performed through trial pits and trenches at relatively low cost.

Table 1. Summary of in situ tests, sampling methods used and geophysical tests performed at the research site

	Testing methods in the field
Sampling	Open trial pits and trenches
In-situ tests	Standard Penetration Test (SPT)
Geophysical	Electrical Resistivity Tomography (ERT, Electrical Imaging)

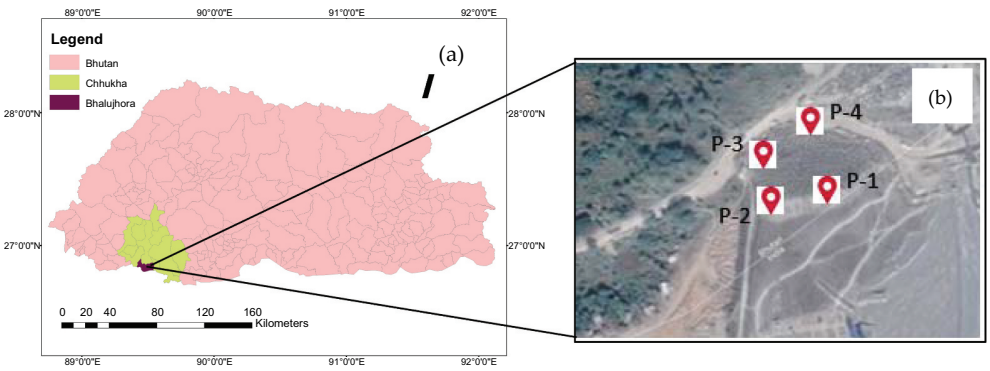


Figure 1. Location of the study area: (a) Phuentsholing, Chhukha, Bhutan; (b) Location of trial pits for study area.



Figure 2. Pictorial representation of study area: (a) Upstream hill, (b) Near bridge sediments, (c) Downstream sediments, (d) Proposed area for developmental activities, (e) Runoff and sediment deposit and (f) Existing road infrastructure and settlements.

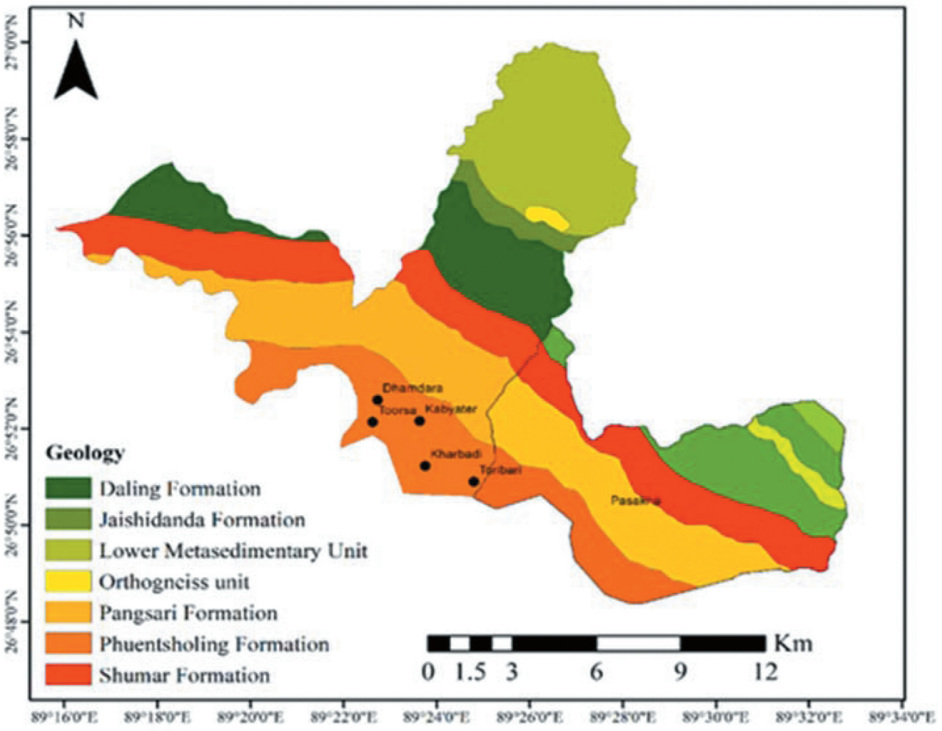


Figure 3. Geological setting of study region (Tempa et al., 2020).

Figure 4. Methodology adopted.

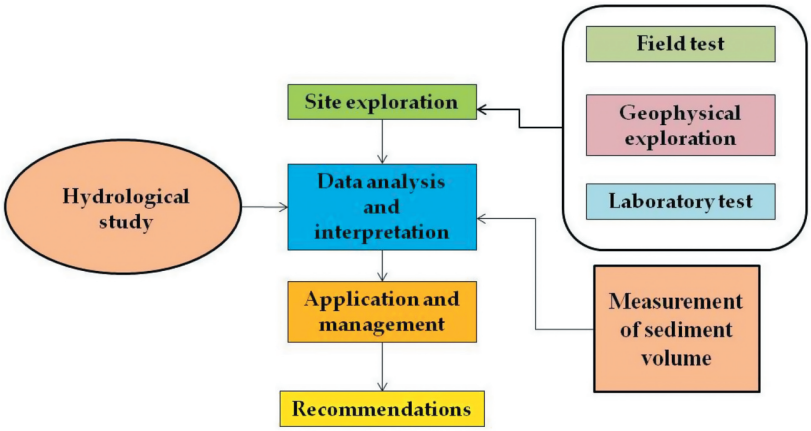


Table 2. Location of the test pits

Pit No.	Test locations	Remarks	Coordinates	Elevation (m)
P-1	Parking area	Open pit	N 26 48 7.2 E 89 28 33.2	283
P-2	Entrance area	Open pit	N 26 48 7 E 89 28 32.1	272
P-3	Minor structure area	SPT	N 26 48 7.6 E 29 28 32.5	324
P-4	Major structure area	SPT	N 26 28 7.7 E 89 28 32.9	344

3.1. Open pit investigation

Exploration pits are considered as integral part of geotechnical investigation for exploring and sampling the significant variations in soil conditions that takes place either in vertical and horizontal direction. The detailed examination of the soil and rock conditions at shallow depths at relatively low cost can be performed by examining through open trial pits (Table 1) and trenches. Moreover, it can also be extended to materials like debris, cobbles and boulders (non-soil) which cannot be sampled with conventional method. Soil profile is the natural strata below the ground surface which represents the alteration in the soil material as a result of numerous weathering processes (Chettri et al., 2019).

The open pits were excavated with mechanical equipment (backhoe). Investigation through open pit was carried out to study the nature of soil deposit along the depth below the ground surface which indicates the shallow soil profile (Table 2). This has allowed investigator to assess the suitability of the foundation depth as well as study the indication of presence of any water table at the investigated depth. To ensure that the uniform horizon of the depsosit was lifted each time during excavation, the bottom of the pit was kept relatively levelled. The excavated material for each level was stacked separately on the surface, in an orderly manner to identify the depth, against the adjoining pit which were used for sampling and conducting laboratory test. The sides of the pits were vertically and horizontally maintained so as to achieve a clean face of soil profile or rock face, if any.

3.2. Standard penetration test (SPT)

The SPT test is one of the reliable in-situ tests used to measure penetration resistance which can be used for geotechnical design purposes (IS: 2131–1981). SPT was carried out at 2.0 m depth at two test pits where major structures are proposed and the no. of blows/penetration resistance



Figure 5. Field test by SPT in test pits P-2 and P-3; (a) Installation of SPT test; (b) Conduction of SPT test at pit.

value (*N*-value) was recorded (Figure 5(a,b)). The field *N*-values are further applied for overburden pressure correction to get the corrected value N_{cor}.

Since, there was no indication of ground water table, depth of water table was considered to be at a greater depth (Water table depth > *B*), hence effect due to water table on bearing capacity was neglected. SPT test is usually carried out to determine the resistance to penetration of the bearing soil. Penetration resistance is further used to compute the bearing capacity of soil at an identified location. The bearing capacity is determined based on the lower SPT test value at pit P-4 considering the worst-case scenario. Based on the corrected *N*-value, the angle of internal friction is estimated using Meyerhof, 1943 (Ranjan & Rao, 2000) (Table 2). Bearing capacity factors are based on ϕ-value proposed by Peck, Hanson and Thornburn, 1974 (Ranjan & Rao, 2000).

To account the shear criteria for non-cohesive soil, the Terzaghi's ultimate bearing capacity is used for strip foundation (Equation 1).

$$q_{ult} = cN_c + \overline{q}N_q + 0.4\gamma BN_\gamma \tag{1}$$

Where $\overline{q} = \gamma D_f$ is overburden pressure (kN/m^2), N_q and N_γ are bearing capacity factors (after Peck, Hanson and Thornburn, 1974), γ = bulk unit weight of soil (kN/m^3), *B* = width of the footing (m) and corresponding ϕ = 30° (worst condition of pit P-3 and P-4). The bearing capacity factors are estimated to be N_q = 8.5 and N_γ = 15. The safe bearing capacity were obtained by applying factor of safety 3.0 as per IS code.

Similarly, the settlement criteria were checked as per Peck, Hanson and Thornburn (Oke & Amadi, 2008) with the following Equation 2.

$$q_{a-net} = 0.44C_w \cdot N \cdot S_a \tag{2}$$

Where $q_{a\text{-}net}$ = allowable bearing pressure (kN/m^2), C_w is correction factor for water table position and S_a as recommended by IS code is the permissible settlement (25 mm for RCC structures) (Ranjan & Rao, 2000).

3.3. Geophysical survey

The study location was explored through geophysical survey to determine geological stratigraphy using Electrical Resistivity Tomography (ERT, Electrical Imaging) technique (Figure 6(a,b)). There are several geophysical investigation methods available based on different equipment and working principles to assess various physical properties (Tempa et al., 2020). For this study, the surface electrical resistivity method was used. Surface electrical resistivity survey works on the principle of potential gradient. The electrical resistivity of soil or rocks and their distribution decides on how the

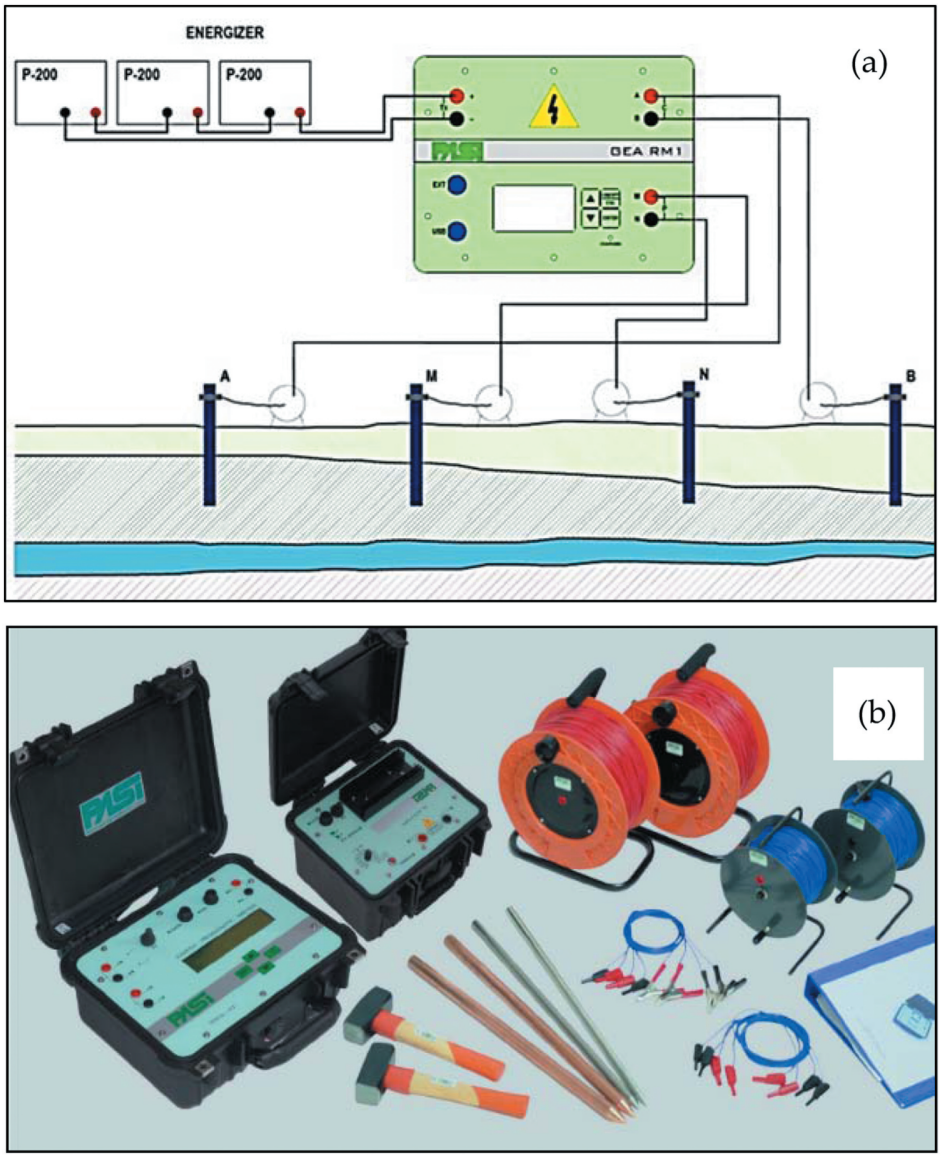
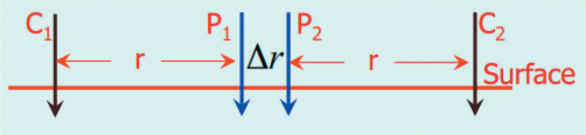


Figure 6. (a) Schematic diagram of the electrical resistivity technique; (b) Electrical resistivity meter RM1 used for site exploration (https://www.pasisrl.it/**).**

electrical potential in the ground around a current-carrying electrode is distributed. Four electrodes were pegged in the ground, where two were supplied with direct current (DC) while the difference of potential between the other two was measure. As depicted in Figure 6(a), the potential electrodes were arranged in line between the current electrodes. Other than direct current, commutated direct current (a square-wave-alternating current) or alternating current of low frequency (typically about 20 Hz) can also be used.

Schlumberger Array-based system is followed for the electrode arrangement sequence which is pictorially depicted in Figure 7. To investigate resistivity variation with depth, the electrode sizes can be changed. As the electrodes' spacing is increased, the depth (material volume) largely increases which affects the apparent resistivity. Thus, vertical variations in resistivity can be inferred from a plot of apparent resistivity against electrodes' spacing. Schlumberger Array system is most commonly used.

Figure 7. Schlumberger array.



For this arrangement, as *Δr* approaches zero, potential gradient at the midpoint of the array can be quantified by the value of V/*Δr*. The ratio of "*r*" to *Δr* is limited, in practice, by the sensitivity of the instrument and it usually ranges from 3 to 30. Therefore, knowing the potential (v), current (I) and resistivity of medium, the geometric factor can be computed using Equation 3 and a finite electrode spacing (Keller & Frischknecht, 1982).

$$V = \frac{\rho I}{2\pi}\frac{1}{K} \tag{3}$$

Where, *V* is the potential difference between two potential electrodes (P_1 and P_2), *ρ* is the resistivity of the medium (i.e soil or rock strata or water table in the current case), *I* is the current flowing from electrode C_1 to C_2) and 1/*K* is the array geometric factor based on various electrode spacing among $C_1P_1P_2C_2$.

In the scenario, whereby the measurements are made over a real heterogeneous earth or the soil mass with varying depth, the symbol *ρ* (Equation 3) is replaced by ρ_a for apparent resistivity. These enable the geotechnical engineer to have proper demarcation of anomalies subsurface strata. Therefore, equation in terms of apparent resistivity (ρ_a) is given by following Equation 4.

The apparent resistivity (ρ_a) is given by Equation 4:

$$\rho_a = \pi\left[\frac{r^2}{\Delta r} - \frac{\Delta r}{4}\right]\frac{V}{I} = \pi\Delta r\left[\left(\frac{r}{\Delta r}\right)^2 - \frac{1}{4}\right]\frac{V}{I} \tag{4}$$

Where, r = spacing between current electrode to potential electrode (i.e C_1 to P_1 or C_2 to P_2), Δr is the spacing between potential electrode, *V* is the potential difference between two potential electrode and *I* is the current between current electrode (C_1 and C_2).

During the operations, the potential electrodes are fixed at a position and the current electrodes are positioned varying the distance "*r*" as per Schlumberger array. During the measurement, the spacing *Δr* is changed decreasing sensitivity is observed. The spacing *Δr* must always be smaller than 0.4s otherwise the assumption of potential gradient becomes invalid. The "*r*" must be held constant while the spacing between electrodes (*Δr*) needs to be changed, so as to find out the presence of local non-homogeneities or lateral changes in the vicinity of the inner electrodes. The electrical resistivity test was conducted along the stretches of sediments as highlighted in Figure 8. Pre-feasibility studies were also conducted before the real field test.

3.4. Laboratory tests

For analysing the suitability of the study area for the construction of custom office laboratory test of soil properties was done. In Table 3, details of tests and testing method in laboratory are mentioned.

3.5. Determination of sediments volume

Nowadays, sediment erosion is a global concern for researchers around the globe. Mohsen & Abayati (Mohsen & Abayati, 2009) suggested multifarious techniques to executive the process involve in the prediction of sediments transport and accumulation. 1D and 2D models are





Figure 8. Electrical resistivity test in the study site of Pasakha.

Table 3. Summary of laboratory tests performed on samples

Parameters	Testing methods in the laboratory	Reference
Specific Gravity	Pycnometer method	IS: 2720 (Part 3- section 2)-1980
Grain size distribution	Sieve Analysis considering Indian Standard Soil Classification System	IS: 2720 (Part 4)—1985
Consistency Limits	Atterberg's limit	IS: 2720 (Part 5)-1995
Shear strength parameters	Direct Shear test	IS: 2720 (Part 13)—1997
Permeability	Variable head method	IS: 5529 (Part-1)-1985
Compactness of Soil, Maximum Dry Density (MDD) and Optimum Moisture Content (OMC)	Modified Proctor Test	IS: 2720 (Part 8)-l 983

commonly used for predicting the sediment scour and deposition. Those models are adequate for many engineering applications; however, for some problem, only a three-dimensional model should be used. Several baseline studies have been carried out for determining the sedimentation rate out of which universal soil loss equation (USLE) is most common due to its simplicity. The main problem with USLE was that it was based on empirical erosion data collected from relatively small area giving results for limited situations (Kitahara et al., 2000). Revised version known as RUSLE Model was developed which could also incorporate GIS into the process of finding the sediments. It is commonly used and accepted model to predict the average soil erosion rate of area of interest (Gaubi et al., 2017). The RUSLE is expressed as shown in Equation 5.

$$A = R \cdot K \cdot L \cdot S \cdot C \cdot P \tag{5}$$

where,

A = Average annual soil loss (metric tons $Km^{-1}year^{-1}$)

R = Rainfall/runoff erosivity (MJ mm h^{-1}Km-1$year^{-1}$)

K = Soil erodibility (metric tons $Km^{-1}MJ^{-1}mm^{-1}$)

LS = Slope Length and Steepness Factor (Dimensionless)

C = Cover & management factor (Dimensionless: 0–1)

P = Support practice factor (Dimensionless: 0–1)

Source: Environmental GIS: Lab 10

Rainfall (R) plays a significant role in the process of runoff and it represents the relation between precipitation occurring over an area and its runoff potential. It is the product of the annual sum of the rainfall energy and the maximum 30-min rainfall intensity in each rainfall event of greater or equal to 13 mm per hour (Kitahara et al., 2000). Due to unavailability of 30-min rainfall data for the study area, alternative formulas developed by Jain et al., (2001) were used for computing R-factor for this study (Equation 6). K factor represents the soil susceptibility to erosion (Cassol et al., 2018) and the amount and rate of runoff (Ashiagbor et al., 2012). It depends on the texture of soil, presence of organic matter, structure soil and permeability. The K factor is calculated using Equation 8.

The Slope Length and Steepness Factors (LS) represents erodibility due to combinations of slope length and steepness relative to a standard unit plot. It expresses the effect of topography, specifically hill slope length and steepness on soil erosion. The slope length factor (L) is defined as the distance from the source of runoff to the point where either deposition begins or runoff enters a well-defined channel that may be part of a drainage network. On contrary, the steepness factor (S) reflects the influence of slope steepness on erosion. An increase in hill slope length and steepness results in an increase in the LS factor (Wordofa, 2011). These two factors are sometimes combined as single index (LS) known as topographic factor. Digital Elevation Model (DEM) is used to calculate the LS factor by most researchers, however, for the current study, field measurement (topographic survey) is done to obtain the parameters required in the Wischmeier's formula and validation was done in Arc GIS. The modified equation for computing the topographic factor (LS factor) in GIS environment is employed by the formula recommended by (Wischmeier and Smith, 1978) in Equation 6.

$$LS = (\ell/72.6)\ m\ \left(65.41\ sin^2\beta + 4.56\ sin\ \beta + 0.065\right) \tag{6}$$

Where l is the cumulative slope length in feet (field measured); β is the downhill slope angle (field measured); m is a slope contingent variable; 0.5 if the slope angle is greater than 2.86 (Wordofa, 2011). The Arc GIS program that calculates the LS values requires a DEM coverage (90-meter resolution obtained from National Centre of Hydrology and Meteorology, NCHM) and two input values: the cell resolution units (feet or meters) and the cutoff slope angle required for identifying cells with net deposition rather than erosion. It is to note the limitation of the DEM resolution i.e. DEM resolution and accuracy increase, the landscape will be more accurately described and erosion estimates will approach actual values. Therefore, authors opted to scope out the same study in future with greater accuracy with the availability of input data. Cell size is the size of the cells being used in the grid-based representation of the landscape. Finally, the LS factor was derived using the above formula in ArcGIS spatial analysis raster calculator function.

The Cover and Management Factor, C is used in reflecting cropping and management practices' effect on erosion rates and is used to make comparison of relative impacts of management options on mitigation plans. The C factor in landslide prone areas (as in the case of the current study area) is 0.0350 (Gaubi et al., 2017). The support practice factor, P is the ratio of soil loss with a specific support practice to the corresponding loss with upslope and downslope tillage (Dutta, 2016). A low P-factor (approaching 0) indicates effective conservation practices, while a high P factor (approaching 1.0) indicates ineffective conservation practices. As per the interviews with the locals it was found out that the entire-affected areas were once used for the agriculture. The area consists of an average slope percentage between 30°- 40° hence, the interpolated value of P = 0.25 is assigned to it (Wischmeier & Smith, 1978). For water bodies it is assigned P-value of 0 and other land use type P value is assigned 1. As RUSLE Model gives the gross soil erosion, and hence brief analysis is done using SDR. Sediment Delivery Ratio (SDR) is a fraction of gross erosion that is transported from a given area in a given time interval. It is also referred to as a transmission coefficient; fraction of gross soil erosion by water that is delivered to a particular point in the drainage system (Ouyang et al., 2005). In the current study the total soil loss from hill side was

calculated based on field data collected. However, with the decrease in the velocity of runoff along the slope some sediments are deposited along the sediments channel and does not reach the slope end. Therefore, the actual sediments deposit was calculated analytically and compared with the empirical formula given by Vanoni (Equation 7).

$$R = 79 + 0.363R_N \tag{7}$$

Where, R_N = mean annual rainfall (mm). In the current study highest annual mean rainfall is taken as 6699 mm for the year 1998. Therefore, K factor can be calculated using the soil erodibility nomograph solution proposed by Wischmeier and Smith (1978) where the detail soil parameters are required as shown in Equation 8:

$$100k = 2.73x10^{-6}xM_{1.14}(12 - a) + 3.25x10^{-2}(b - 2) + 2.5x10^{-2}(c - 3) \tag{8}$$

where,

M = particle size diameter

a = % organic matter

b = the soil-structure code used in soil classification, and

c = the profile-permeability class

These parameters are obtained by doing laboratory test and field test described in the paper. The product of all the factors (Equation 5) will give the estimated total soil erosion in the catchment area hence sediment loss at each grid/cell can be deducted (Department of Engineering Services, 2020).

4. Results

The occurrence of the significant variation of soil conditions (horizontally and vertically), buried features and existence of huge soil and/or non-soil materials like boulders and debris, that cannot be sampled with conventional methods can be substantially determined by the trial pits (Mayne et al., 2002). Two open pit investigation was carried out at trial pit P-1 and P-2 and SPT test at test pit P-3 and P-4. The open pit investigation was carried out to a depth of 3.0 m below the formation level to determine the stratigraphy of the soil to ascertain the condition of the subsoil, material composition and relative density. Pit loggings were recorded and the soil samples were collected for laboratory analysis. Thus, the integral part of assessment of true character of a soil is through its soil profile, wherein there is a display of a succession of layers. Although various properties like color, texture, structure, biological composition, porosity, organic matter content, consistency and constitution of chemical may vary in those various layers. Unlike borehole and drill-hole investigation, the structure and stratification of the underlying ground can be seen and logged in 3D (Figure 9) with trial pitting method.

4.1. Textural characteristics of soil profile

Open pit investigation through open pit was carried out for analyzing the nature of soil deposit at different depth below the surface which indicates the shallow soil profile. For analyzing the suitability of the establishment of building as well as foundation depth on the deposited landform particularly, in the mountainous region it very essential to investigate subsurface soil condition and water table.

From open pit investigation, for test pit 1, up to 1.0 m depth the soil consists of mostly gravel with substantial amount of sand. Between 1 and 1.8 m, the soil consists of substantial amount of

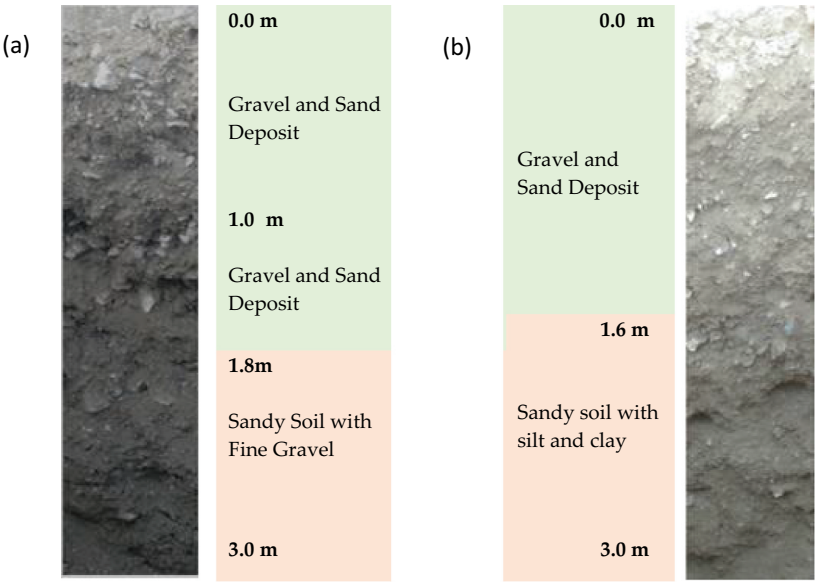


Figure 9. Soil profile/stratigraphy at trial pit P-1 and P-2 respectively.

gravel and less amount of sand while from 1.8 to 3.0 m, the soil is sandy with some amount of gravel fractions. Similarly, for test pit P-2, up to 1.6 m the soil is mostly gravel with good amount of sand and between 1.6 to 3.0 m, the soil is mostly sand with some fractions of gravel. The actual distribution of the particles in each pits are discussed in section 4.4.2.

4.2. Penetration resistance/bearing capacity

SPT test was carried out for measuring the penetration resistance of the soil to establish a construction in the study site. In this test to estimate the ultimate bearing capacity, the basic equation proposed by Terzaghi (Ranjan & Rao, 2000) is used and settlement criteria are verified by using correlation proposed by Peck, Hanson and Thornburn (Ranjan & Rao, 2000) for permissible settlement of 25 mm for RCC structures as recommended by IS: 1904–1986 (Ranjan & Rao, 2000). The safe bearing capacity are further calculated by applying factor of safety (FoS) of 3.0 as suggested by IS code. The graphical representation of the results is presented in Figure 10(a,b).

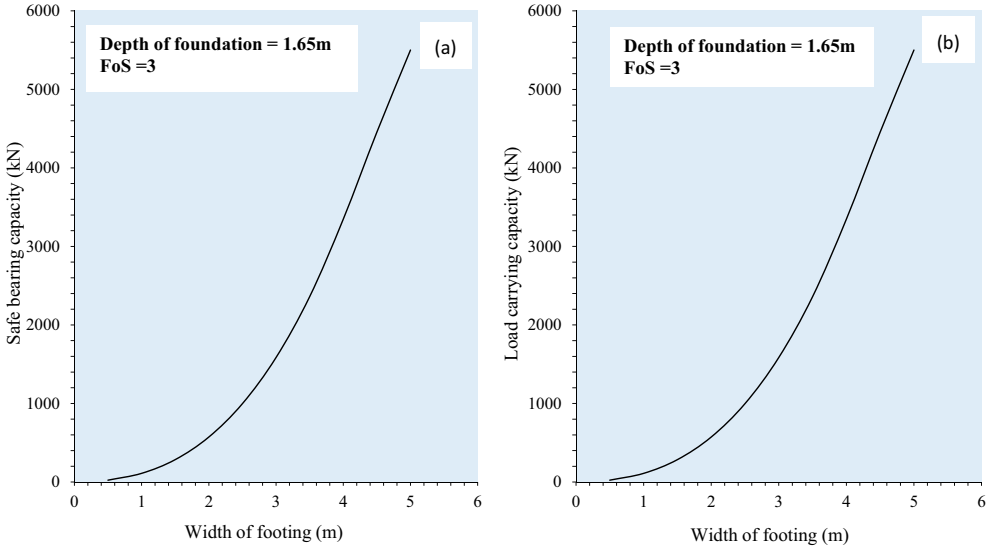


Figure 10. (a) Safe bearing capacity; (b) Load carrying capacity.

For any infrastructure development, assessment of foundation subsoil becomes important in determining the safe bearing capacity and foundation depth. The safe bearing capacity of the subsoil decides the proportioning of the size and type of the footing to be proposed. Figure 10(a) shows the relationship between the footing width and the safe bearing capacity of the subsoil for the study area. The results indicate the increase in the bearing capacity as the width of the footing increases. In structural design for the footing, the ultimate design load is utilized, however it should not exceed the load-carrying capacity (Figure 10(b)) corresponding to safe bearing capacity of the subsoil for required width of the footing. This has a direct cost and safety aspects of the foundation of the proposed project. The overburden correction for *N*-vales at 1.65 m embedment depth and corresponding angle of friction are presented in Table 4.

4.3. Depth-wise resistance of soil

Figure 11 shows the result of the resistivity survey. The profile is located along the Phuentsholing-Pasakha highway at Bhalujhora adjacent to new bridge constructed by Construction Development Corporation Limited in 2019 which is currently covered by sediments. The depth of investigation is about 43.1 m with the profile length of 200 m. This is achieved by maintaining spacing of electrodes as 10 m. From Figure 11, the results indicate the deposition of huge quantity of sediments (around 31.3 m) of alluvium typically of quartzite which are prone to erosion. Pockets of phyllite deposition are present along with the horizontal profile (80–120) m and in 160 m. A layer of phyllite was also detected beneath the quartzite deposition till the depth of 43.1 m. The phyllite layer is dissolvable, when water seeps during monsoon creating a huge plunge over the surface which is further filled by upcoming sediments.

4.4. Laboratory tests

Various tests were conducted in the laboratory to determine the physical properties of the sediment. All tests comply with the Indian standard IS codes.

Table 4. Overburden correction to "*N*-values" filed value and corresponding angle of internal friction

Pit No.	Field SPT Value "*N*"	Bulk Unit Weight (kN/m^3)	Depth of foundation, D_f (m)	Over burden Pressure, $\sigma' = \gamma.D_f$ (kN/m^2)	Correction for Over burden Pressure, kN/m^2	Corrected SPT Value "N_{cor}"	ϕ Meyerhof, 1956
P-3	15	16.5	1.65	30.53	1.4	21	32°
P-4	14	16.0	1.65	30.53	1.4	20	30°

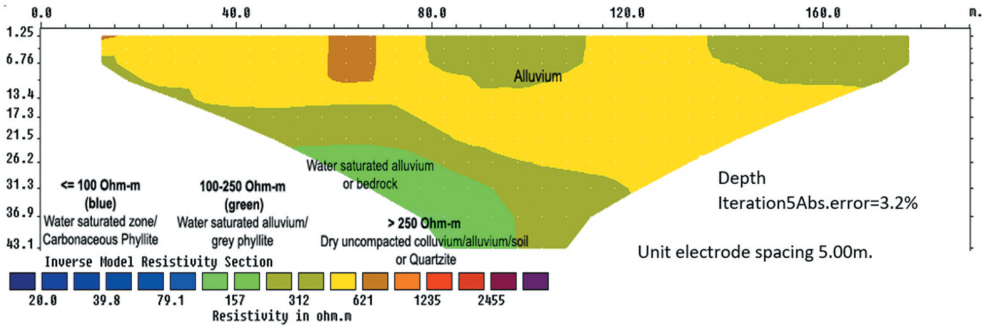


Figure 11. Resistivity profile of the study area.

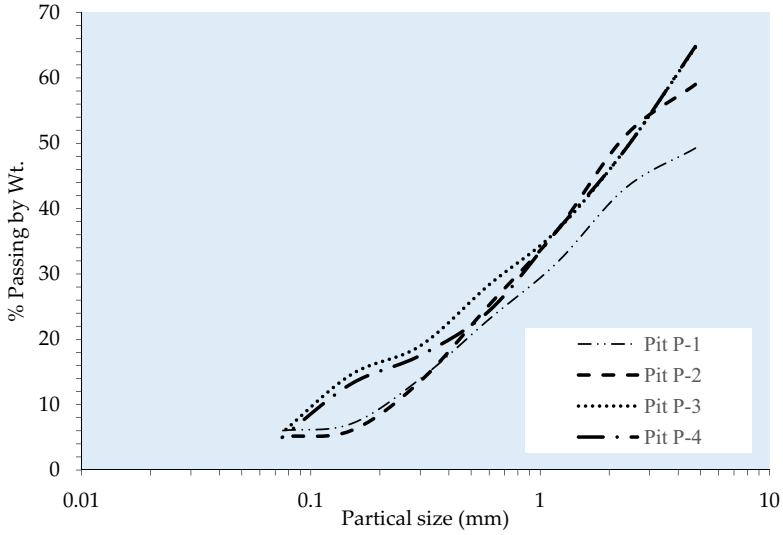


Figure 12. Grain size distribution graph of four pits.

4.4.1. Specific gravity

Specific gravity is the ratio of the weight of a given volume of material to the weight of an equal volume of water (at 20°C). The specific gravity indicates how much heavier (or lighter) the material is than water according to IS: 2720 (Part 3- section 2)-1980. Specific gravity is an integral gravimetric property useful for determining various soil properties such as the void ratio, unit weight, degree of saturation and most importantly the porosity of the soil. It can also be used in wet analysis for determining particle size. Thus, specific gravity of soil should be precisely determined. The Pycnometer method was adopted for determining specific gravity of soil for fraction passing 4.75 mm I. S. sieve and results are presented in Table 5.

4.4.2. Particle size analysis

The soil classification is done Indian Standard Soil Classification System (ISSCS) which essentially gives particle size distribution of granular material materials by passing through different sizes arranged sieves as per the standard. Sieve analysis is used in engineering classification of soils, particularly coarse-grained soils. It is one of the suitability criteria to be accounted for the soils to be used for road, dam, airfield and other embankment construction. The soil water content/ movement can be predicted from grain size analysis. Also, the grain size distribution is suitable parameter to predict phenomenon of frost action in soil in colder climate. From the grain size analysis (Figure 12), soils are broadly classified as coarse and fine-grained soil. The sieve analysis was carried out according to IS: 2720 (Part 4)—1985.

4.4.3. Consistency limits

In multifarious states that the soil exists, the engineering properties of the soil vary based on behaviour of soil and its consistency. Thus, by properly examining the changes in soil's behaviour,

Table 5. Physical properties of soil deposit

Pit No.	G_s	Consistency	c (kN/m^2)	ϕ	ρ_{dmax} (g/cc)	OMC (%)	k (m/sec)
P-1	2.5	Non-plastic	5.5	36°	1.74	18.61	$3.40x10^{-4}$
P-2	2.6	Non-plastic	8.5	35°	1.55	21.59	$2.71x10^{-4}$
P-3	2.6	Non-plastic	9.0	34°	1.55	24.51	$2.52x10^{-4}$
P-4	2.4	Non-plastic	6.0	37.5°	1.76	20.39	$2.54x10^{-4}$

the boundary between each state can be defined. Therefore, Atterberg limits test based on IS guidelines is performed to distinguished properties of clay and silt. Liquid and Plastic limit test was conducted as per IS: 2720 (Part 5)-1995 with the soil passing through 0.075 mm IS sieve. The soil was found non-plastic in nature in all the test pits indicating larger portion of silt content. The soil samples are from depth 1.0 m to 3.0 m. Sieve analysis have also shown that the fraction passing 0.075 mm IS sieve is substantially less (4 to 6.5%) of the soil sample.

4.4.4. Direct shear

The objective of the direct shear test conducted as per IS: 2720 (Part 13)—1997 was to determine the mechanical properties such as cohesion (c) and angle of internal friction (ϕ) (Figure 13). Shear strength parameters can be used to determine soil shear strength as well as to estimate the bearing capacity of the soil.

4.4.5. Compaction test

Modified proctor test refers to heavy compaction test, an improved version of light compaction test. The proposed site is expected to have heavy traffic load due to which settlement could be one issue. The daily traffic is more than 500 to 1500 CVPD (Department of Roads, 2005). To achieve the desired compaction, modified proctor test was conducted to determine Maximum Dry Density (MDD) and Optimum Moisture Content (OMC) (Ouyang et al., 2005). Figure 14 shows the relationship between the dry density and moisture content for each sample. To determine the degree of compaction in the field, field test could be conducted to achieve 98% of maximum dry density and + or—2% OMC for embankment, sub-grad, sub-base and base material (Department of Engineering Services, 2020). Based on the type of soil either core cutter or sand replacement methods can be used.

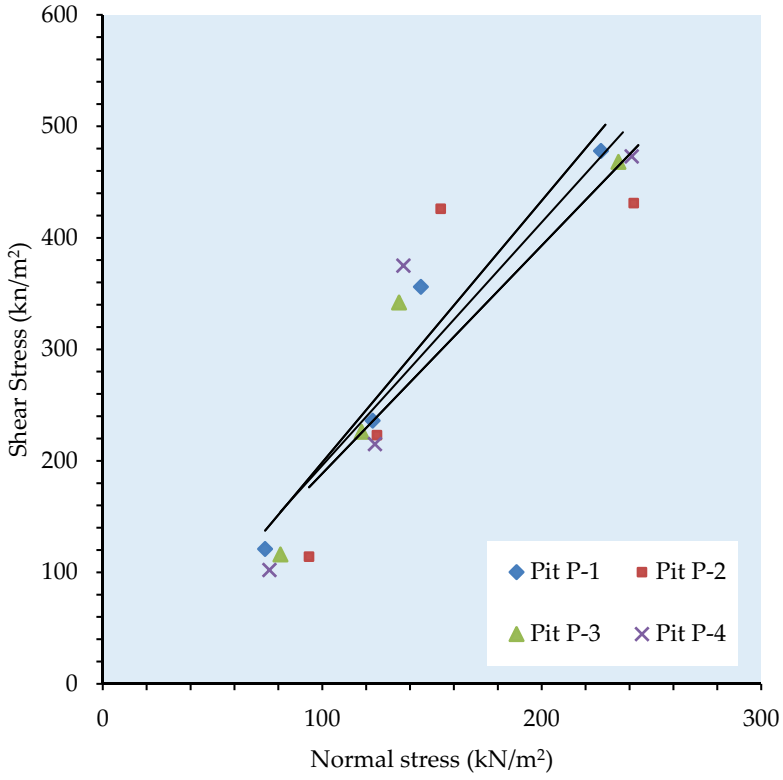
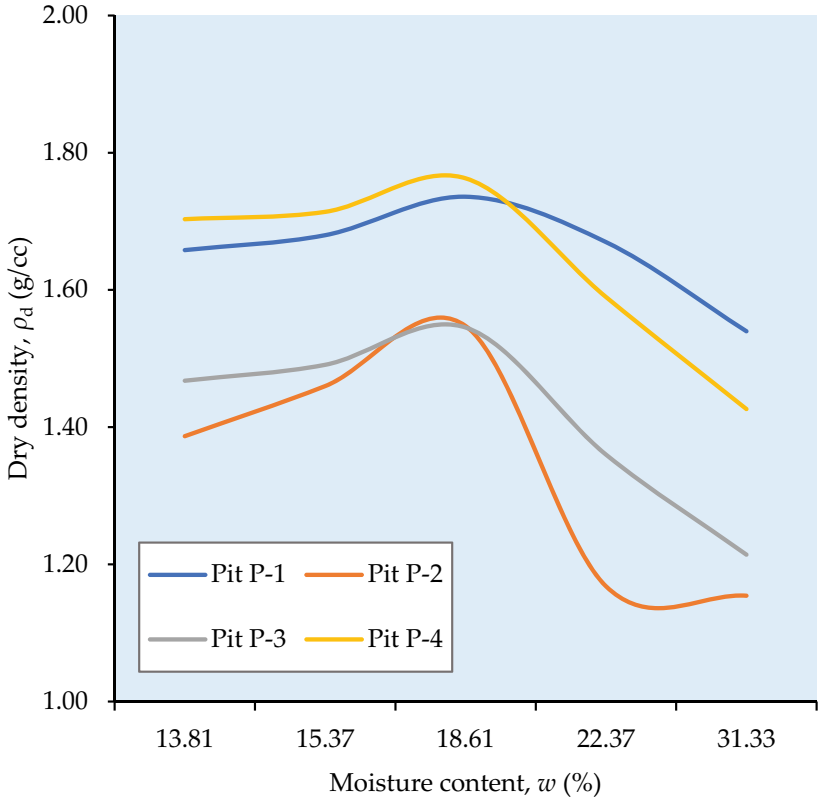


Figure 13. Relationship between normal stress and shear stress.

Figure 14. Relationship between dry density and moisture content.



4.4.6. Permeability

The property of a soil which allows water to infiltrate through it, is called permeability. It is very important factor for the structures which are in contact with water. The permeability of a soil is paramount property and is useful in understanding various engineering issues, e.g., seepage analysis through soils, drainage, rate of settlement of compressible layers, etc. Prakash and Jain (2002) (Prakash & Jain, 2002) clarified that water flowing through the soil exerts a significant seepage forces which directly affect the safety of hydraulic structures. Any material having continuous voids is permeable. The permeability of different soils varies widely. Gravels and sands are more permeable as compared to silts, whereas clays are least permeable. The permeability of soil is apprised by permeability coefficient 'k'. It is measured as the velocity of flow under unit hydraulic gradient. The permeability of sediments is quantified based on IS: 5529 (Part-1)-1985 and projected in Table 5 with other soil parameters.

4.5. Precipitation pattern in study area

Climate in Bhutan vagaries as a function of seasonal monsoon winds and the geographical conditions. The climatic zones widely vary from hot and humid tropical in the South to alpine in the North dividing Bhutan into the six major climatic zones (alpine, cool temperature, subtropical, tropical, tundra and warm temperate). Monsoon season in Bhutan falls within May to September where 90% of annual precipitation has been recorded by local meteorological agencies. The mean annual rainfall widely varies from as low as 250 mm in the North to a maximum of 6000 mm in South with an average rainfall of 2000 mm in most part of the country. Records show that uneven daily precipitation exist during monsoon whereby around 10% of the total annual rainfall occurs in one day whereas 50% of the total annual rainfall was observed in 10 days during summer. Such vagaries in the precipitation initiate extensive runoff of the top soil and sediments deposition in settlement area.

The studied area falls in the sub-tropical climate zone with warm to hot summer and pleasant to cold winter. From the past records the mean annual rain fall varies. There is a drastic fluctuation in annual rainfall. The maximum rain fall was recorded 6699 mm during the year 1998 and the minimum was in 2006 where the total rain fall recorded was 1996 mm (Department of Geology and Mines, 2010).

Double ring infiltrometer test was conducted as per ASTM D5093 in 26 pocketed locations within the study area as shown in Figure 17. The highest Infiltration rate was found to be 23.29 mm/hr. on average. It depicts low runoff potential and high infiltration rates (*Wischmeier & Smith, 1978*). The soil-based on texture is categorized as well-drained sand and gravel.

4.6. Volume of sediments

The precipitation of the study area is as high as 6000 mm (Figures 15, 16) in monsoon season which has led to high erosion at the upstream of the study area causing huge sediment deposit at the downstream. For this reason, measuring the sediment volume is essential for the study site. RUSLE was used for measuring the sediment volume of the study area. It was confirmed that no organic materials are presents in the sediment and predominant soil

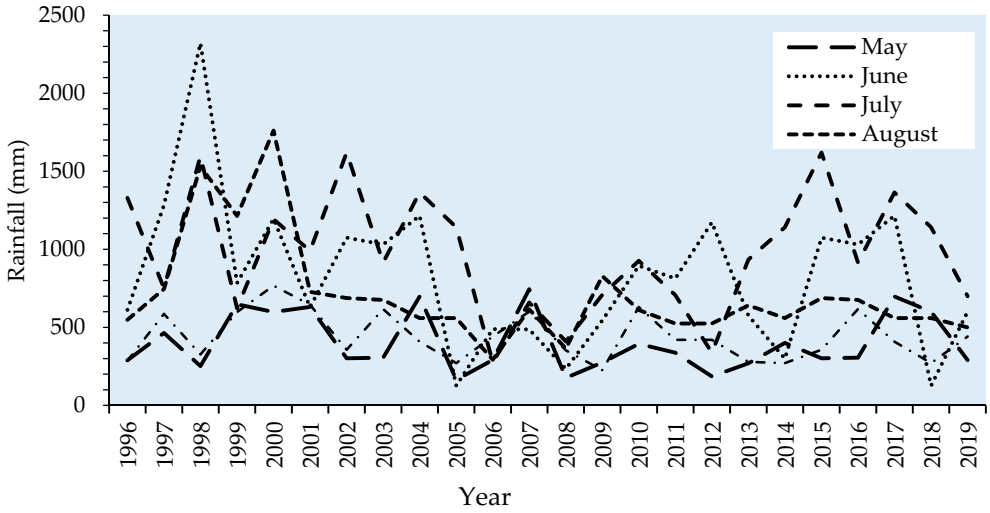


Figure 15. Variation of rainfall during peak monsoon in study area from May to September.

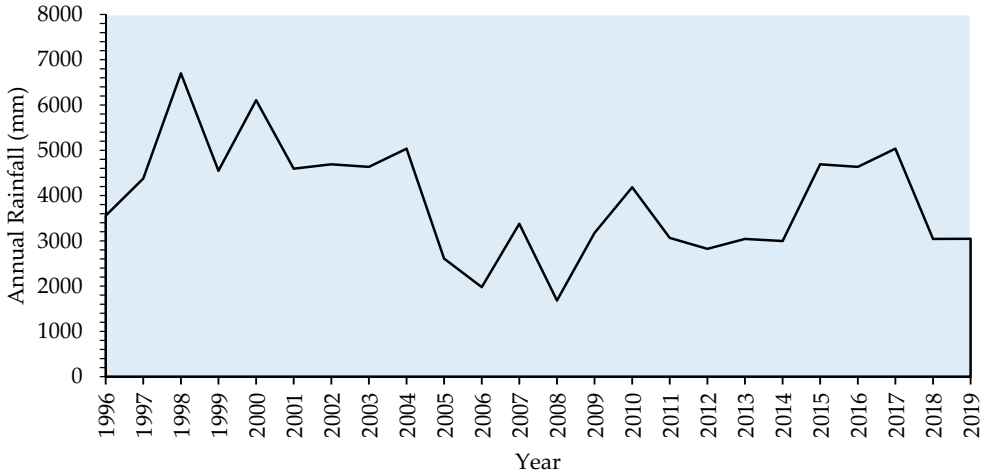


Figure 16. Variation of annual rainfall (1996–2019).



Figure 17. Double ring infiltrometer test.

deposit is grey phyllite sand whose particle size varies to 4.25 mm. The test for permeability was conducted and results obtained are used in the calculation. It was found out that total soil loss from the hill slope is 593,267,807.9 tons/Km^2/year. Sediment delivery ratio was calculated using Vanoni SDR equation i.e. SDR = 0.42 *A − 0.125. The actual total sediment deposited was roughly estimated in the basin area (3 Km x 2 Km) as 144,600,000 tons/Km^2/year. However, using SDR ratio the sediments yield was found out to be 199,173,322.1 tons/Km^2/year.

5. Conclusions

In the present study, the subsurface soil parameters were assessed by in-situ open pit investigation, geophysical survey and Standard Penetration Test (SPT), while the engineering properties of the sediment deposits were determined from laboratory tests. Following conclusions are drawn from the present study:

(1) Based on the geotechnical investigation, the sediment deposit at the study area is coarse-grained soil, in which 35–50%—gravel, 50–60%—sand and 4–6.5% fines in all the four test pits.

(2) As per Indian Standard Soil Classification System (ISSCS), the soil is classified as SG-SM (Sandy Gravel-Sandy silt), non-plastic soil.

(3) The depth of foundation of 1.65 m from prepared or reclaimed ground level is proposed based on the present investigation

(4) There is no influence of water table in the current study.

(5) Proper drainage system around the planed area is highly proposed.

(6) The study indicated huge volume of sediments and geophysical survey validated the deposit up to depth of 43.1 m.

(7) An alternative use of the sediment deposit is recommended as cost recovery measures such as in road works: sub-base and base course prior to relevant quality test.

Acknowledgements
The authors are thankful to College of Science and Technology, Royal University of Bhutan and Department of Geology and Mines for providing technical reports.

Funding
The authors have no direct funding for this research.

Author details
Karma Tempa[1]
E-mail: karmatempa.cst@rub.edu.bt
ORCID ID: http://orcid.org/0000-0001-9558-8692
Nimesh Chettri[2]
E-mail: nimeshchhetri.cst@rub.edu.bt
ORCID ID: http://orcid.org/0000-0002-4027-7763
Raju Sarkar[2]
E-mail: rajusarkar@dce.ac.in
Sunil Saha[3]
E-mail: sunilgeo.88@gmail.com
ORCID ID: http://orcid.org/0000-0003-2739-3716
Lily Gurung[2]
E-mail: lilygurung.cst@cst.edu.bt
Tshering Dendup[2]
E-mail: tsheringdendup.cst@rub.edu.bt
Bhawani Shankar Nirola[2]
E-mail: bhawani.cst@rub.edu.bt
[1] Civil Engineering and Architecture Department, College of Science and Technology, Royal University of Bhutan, Phuentsholing, Bhutan.
[2] Department of Civil Engineering, Delhi Technological University, Delhi, India.
[3] Department of Geography, University of Gour Banga, Malda, India.

Citation information
Cite this article as: Geotechnical parameter assessment of sediment deposit: A case study in Pasakha, Bhutan, Karma Tempa, Nimesh Chettri, Raju Sarkar, Sunil Saha, Lily Gurung, Tshering Dendup & Bhawani Shankar Nirola, *Cogent Engineering* (2021), 8: 1869366.

References

Ali, T. S., & Fakhraldin, M. K. (2016). Soil parameters analysis of Al-Najaf City in Iraq: Case study. *Journal of Geotechnical Engineering*, *3*(1), 1987–2394. http://engineeringjournals.stmjournals.in/index.php/JoGE/article/view/3749

Ashiagbor, G., Forkuo, E. K., Laari, P., & Aabeyir, R. (2012). Modeling Soil erosion using rusle and Gis tools. *International Journal of Remote Sensing & Geoscience (IJRSG)*, *2*(2), 7–17. http://www.ijrsg.com/Issue.aspx?IssueId=24

Cassol, E. A., da Silva, T. S., Eltz, F. L. F., & Levien, R. (2018). Soil erodibility under natural rainfall conditions as the K factor of the universal soil loss equation and application of the nomograph for a subtropical Ultisol. *Revista Brasileira de Ciencia do Solo*, *42*, 1–12. https://doi.org/10.1590/18069657rbcs20170262

Chettri, N., Chhetri, I. B., & Yangzom, K. (2019). Profile of subgrade subsidence in Sorchen. *International Journal on Integrated Education*, *2*(6), 47–54. https://doi.org/10.31149/ijie.v2i6.197

Chettri, N., Sarkar, R., Adhikari, K., & Chettri, I. B. (2019). Forecasting landslide in Chukha from index properties of soil-research to policy making. *World Scientific News an International Scientific Journal*, *135*, 227–248. http://psjd.icm.edu.pl/psjd/element/bwmeta1.element.psjd-e322eeac-ef79-42d1-9c71-85f0a50c0682

Department of Engineering Services. (2020). *Specifications for building and road works*. Ministry of Works and Human Settlement.

Department of Geology and Mines. (2010). *Technical Report on proposed landslide mitigations in Phuentsholing Thromde* (pp. 1–60). Ministry of Economic Affairs.

Department of Roads. (2005). *Pavement design manual*. Ministry of Works and Human Settlement.

Dini, B., Manconi, A., & Loew, S. (2019). Investigation of slope instabilities in NW Bhutan as derived from systematic DInSAR analyses. *Engineering Geology*, *259*, 105111. https://doi.org/10.1016/j.enggeo.2019.04.008

Dutta, S. (2016). Soil erosion, sediment yield and sedimentation of reservoir: A review. *Modeling Earth Systems and Environment*, *2*(3), 1–18. https://doi.org/10.1007/s40808-016-0182-y

Gaubi, I., Chaabani, A., Ben Mammou, A., & Hamza, M. H. A. (2017). A GIS-based soil erosion prediction using the revised universal soil loss equation (RUSLE) (Lebna watershed, Cap Bon, Tunisia). *Natural Hazards*, *86*(1), 219–239. https://doi.org/10.1007/s11069-016-2684-3

Keller, G. V., & Frischknecht, F. C. (1982). *Electrical methods in geophysical prospecting* (5th ed.). Pergamon Press.

Kitahara, H., Okura, Y., Sammori, T., & Kawanami, A. (2000). Application of universal soil loss equation (USLE) to Mountainous Forests in Japan. *Journal of Forest Research*, *5*(4), 231–236. https://doi.org/10.1007/BF02767115

Laskar, A., & Pal, S. K. (2012). Geotechnical characteristics of two different soils and their mixture and relationships between parameters. *EJGE*, *17*, 2821–2832. http://www.ejge.com/2012/Ppr12.261alr.pdf

Mata-Lima, H., Alvino-Borba, A., Pinheiro, A., Mata-Lima, A., & Almeida, J. A. (2013). Impacts of natural disasters on environmental and socio-economic systems: What makes the difference? *Ambiente e Sociedade*, *16*(3), 45–64. https://doi.org/10.1590/S1414-753X2013000300004

Mayne, P. W., Christopher, B. R., & DeJong, J. (2002). *Subsurface investigations — Geotechnical site characterization reference manual*. U.S. Department of Transportation Federal Highway Administration.

Mohsen, M., & Abayati, A. (2009). Review study on sediment models. *Journal of Engineering and Applied Sciences*, *4*(2), 152–160. https://medwelljournals.com/abstract/?doi=jeasci.2009.152.160

Ngah, S. A., & Nwankwoala, H. O. (2013). Evaluation of geotechnical properties of the sub-soil for shallow foundation design in Onne, Rivers State, Nigeria. *The IJES*, *2*(11), 08–16. http://www.theijes.com/papers/v2-i11/Part.3/C02110308016.pdf

Oghenero, A. E., Akpokodje, E. G., & Tse, A. C. (2014). Geotechnical properties of subsurface soils in Warri, Western Niger Delta, Nigeria. *Journal of Earth Sciences and Geotechnical Engineering*, *4*(1), 89–102. http://www.scienpress.com/Upload/GEO/Vol%204_1_8.pdf

Oke, S. A., & Amadi, A. N. (2008). An assessment of the geotechnical properties of the sub-soil of parts of Federal University of Technology, Minna, Gidan Kwano Campus, for foundation design and construction. *Journal of Science Education and Technology*, *1*(2), 87–102. https://www.researchgate.net/publication/317851129_An_assessment_of_the_geotechnical_properties_of_the_subsoil_of_parts_of_Federal_University_of_Technology_Minna_Gidan-Kwano_Campus_for_foundation_design_and_construction

Ouyang, D., Bartholic, J., & Selegean, J. (2005). Assessing sediment loading from agricultural croplands in the Great Lakes Basin. *The Journal of American Science*, *1*(2), 14–21. doi: 10.7537/marsjas010205.03

Phuentsholing Thromde. (2002). *Urban development plan: 2002-2017*.
Prakash, S., & Jain, P. K. (2002). *Engineering soil testing*. Nem Chand & Bros.
Ranjan, G., & Rao, A. S. R. (2000). *Basic and applied soil mechanics* (2nd ed.). New Age International (P) Ltd.
Tempa, K., Chhetri, N., Bdr, I., & Yangzom, D. (2016). Subsurface exploration for road sinking at Sorchen and counter measures. *International Civil Engineering Congress: Karachi, Pakistan*, *8*, 295–303. http://iepkarachi.org.pk/8th_International.html?fbclid=IwAR29Y0DnCWvF4X99-7QBX1QpQgLVlmpqnAcjAWt_d9N431jAAHYEf-whuFM
Tempa, K., Sarkar, R., Dikshit, A., Pradhan, B., Simonelli, A. L., Acharya, S., & Alamri, A. M. (2020). Parametric study of local site response for bedrock ground motion to earthquake in Phuentsholing, Bhutan. *Sustainability*, *12* (13), 5273. https://doi.org/10.3390/su12135273
Wischmeier, W. H., & Smith, D. D. (1978). *Predicting rainfall erosion losses: A guide to conservation planning*. Science and Education Administration, United States Department of Agriculture in cooperation with Purdue Agricultural Experiment Station, U. S.
Wordofa, G. (2011, June). Soil erosion modeling using Gis and rusle on the Eurajoki Watershed Finland [Bachelor's degree, Tampere University of Applied Sciences].
Youdeowei, P. O., & Nwankwoala, H. O. (2013). Suitability of soils as bearing media at a freshwater swamp terrain in the Niger Delta. *Journal of Geology and Mining Research*, *5*(3), 58–64. https://doi.org/10.5897/JGMR11.046
Zawawi, M. H., Azman, A., Radzi, M. R. M., Abas, A., & Aziz, N. A. (2019). Coastal sediment transport study along shoreline by using smooth particle hydrodynamics. *IOP Conference Series: Materials Science and Engineering*.

cogent oa

© 2020 The Author(s). This open access article is distributed under a Creative Commons Attribution (CC-BY) 4.0 license.

You are free to:
Share — copy and redistribute the material in any medium or format.
Adapt — remix, transform, and build upon the material for any purpose, even commercially.
The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:
Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
No additional restrictions

You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

Cogent Engineering **(ISSN: 2331-1916) is published by Cogent OA, part of Taylor & Francis Group.**

Publishing with Cogent OA ensures:

- Immediate, universal access to your article on publication
- High visibility and discoverability via the Cogent OA website as well as Taylor & Francis Online
- Download and citation statistics for your article
- Rapid online publication
- Input from, and dialog with, expert editors and editorial boards
- Retention of full copyright of your article
- Guaranteed legacy preservation of your article
- Discounts and waivers for authors in developing regions

Submit your manuscript to a Cogent OA journal at www.CogentOA.com

Materials Today: Proceedings xxx (xxxx) xxx

Contents lists available at ScienceDirect

Materials Today: Proceedings

journal homepage: www.elsevier.com/locate/matpr

*h*GWO-PS optimized dissimilar fractional order 2DOF-PID (FO-2DOF-PID) controllers for performance analysis of interconnected thermal power system

Shailaja Yogesh Kanawade [a,*], Vikas Soni [b]

[a] ECE-Department, Career Point University, Kota, Rajasthan, India
[b] ECE-Department, Modi Institute of Technology, Kota, Rajasthan, India

ARTICLE INFO

Article history:
Received 23 October 2020
Accepted 4 November 2020
Available online xxxx

Keywords:
Automatic Generation Control (AGC)
GWO
ITAE
Interconnected thermal power systems
PS algorithm
FO-2DOF-PID

ABSTRACT

Here, a hybridization of Grey Wolf Optimization (GWO) and Pattern Search algorithms has been implemented into account and the resultant mixed/hybrid algorithm: *h*GWO-PS has been used in AGC of interconnected thermal power system having two dissimilar fractional order 2DOF-PID (FO-2DOF-PID) controllers. The ITAE (objective function) has been minimized using *h*GWO-PS for evaluating the final optimized parameters of FO-2DOF-PID. The comparison of transient analysis of the system under investigation obtained by the implemented approach: *h*GWO-PS/FO-2DOF-PID with some well known techniques is also given. The percentage improvement/enhancement has also been carried by the implemented approach for the same with respect to the other approaches. The sensitivity/robustness analysis of the same with *h*GWO-PS/FO-2DOF-PID has also been carried out. The simulation results show that the investigated power system with proposed approach delivers outstanding results in terms of aforesaid parameters, system dynamic responses and better robustness when compared with other well known approaches in literature.
© 2021 Elsevier Ltd. All rights reserved.
Selection and peer-review under responsibility of the scientific committee of the Emerging Trends in Materials Science, Technology and Engineering.

1. Introduction

Intelligent technique based interconnected power systems (IPS) consist of enhanced power quality [1–3] in which AGC plays a key role since it makes a balance of generation and system load, controls the power interchanges and frequency and monitors tie-line flows, etc [4–6]. Different types of controller's structures have been used for AGC of IPS [7–11] for better control of load frequency.

Various Evolutionary Techniques have already been implemented to tune/optimize the different controller's structures in AGC of IPS like: BFOA based PI controller [12,13], ICA based PID [14], DE optimized PI and other classical frequency controllers [15,16], TLBO optimized I and PID for MUMS-PS [17], Emotional Learned controllers [18], FA + Online wavelet filter for unequal areas of interconnected power plants [19] and ABC for equal power plants [20], GSA tuned FOPID [21], QOHSA optimized PID [22] and TLBO optimized PI/PID [23], etc.

In [24], various popular and modern approaches for interconnected thermal power system have been employed for improving the performance of the same like: DSA/ PI, PID and PIDF, CLPSO/ PI and PID, EPSDE / PI and PID and SHADE / PI and PID.

For further improvement in performance of the aforesaid IPS, hybrid/mixed soft computing techniques are being applied in the literature such as; *h*BFOA-PSO/PI [25], hybrid PSO-PS/fuzzy PI [26] and *h*GWO-PS/2DOF-PID [27–29], etc. In continuation, using different controller named fractional order 2DOF-PID (FO-2DOF-PID) with a hybrid technique; *h*GWO-PS, a different approach called as *h*GWO-PS/FO-2DOF-PID has been proposed for the AGC of interconnected thermal power system. The *h*GWO-PS is hybridization of GWO [30] and PS [31]. The present work includes the comparison of proposed approach with some popular and modern approaches for the same and robustness analysis in order to show the superiority over them.

2. AGC of standard interconnected thermal power system

Here, the widely used AGC of interconnected thermal power system is taken for consideration as shown in Fig. 1. There are

* Corresponding author.

https://doi.org/10.1016/j.matpr.2020.11.078
2214-7853/© 2021 Elsevier Ltd. All rights reserved.
Selection and peer-review under responsibility of the scientific committee of the Emerging Trends in Materials Science, Technology and Engineering.

Please cite this article as: S. Yogesh Kanawade and V. Soni, *h*GWO-PS optimized dissimilar fractional order 2DOF-PID (FO-2DOF-PID) controllers for performance analysis of interconnected thermal power system, Materials Today: Proceedings, https://doi.org/10.1016/j.matpr.2020.11.078

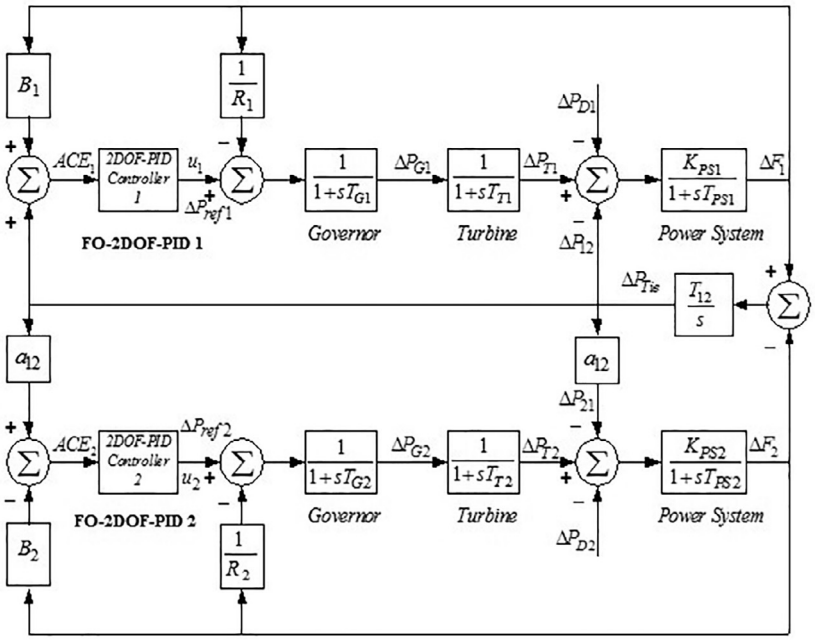


Fig. 1. AGC of two area interconnected non reheat thermal-thermal power system.

two area of 2000 MW at 1000 MW nominal load which are interconnected through tie-line. The nomenclature, mathematical modelling of system and detailed description can be found in [5,7,12,22–24,27–29] and the nominal parameters for implementation has been given in Table 1.

3. Proposed approach

In the present work, the followings have been included since these all are responsible for the system performance:

- FO-2DOF-PID controller's structure.
- ITAE (objective function).
- Controller's design problem.
- *h*GWO-PS algorithm in Brief.

3.1. FO-2DOF-PID controller's structure

The fractional order PID controller ($PI^{\lambda}D^{\mu}$) is the simplest form of FOC, whereλ and μ represent the operators of fractional order (real numbers). The conventional FOPID controller is defined as:

$$G_C(s) = K_P + \frac{K_I}{s^{\lambda}} + K_D s^{\mu} \tag{1}$$

In the present work, two dissimilar fractional order 2DOF-PID (FO-2DOF-PID) controllers shown in Fig. 2 has been considered in the aforesaid investigated AGC of power system. In FO-2DOF-PID, a filter is cascaded to the fractional derivative of FOPID for minimizing the noise of high frequency [11].

In Fig. 2, C(s) and F(s) are as follows:

$$C(s) = K_P + \frac{K_I}{s} + \frac{K_D s^{\mu}}{1 + \frac{K_D s^{\mu}}{K_P N_F}} \tag{2}$$

$$F(s) = K_P P_W + \frac{K_I}{s^{\lambda}} + \frac{K_D s^{\mu}}{1 + \frac{K_D s^{\mu}}{K_P N_F}} D_W \tag{3}$$

Hence, the U(s) can be expressed as:

$$U(s) = K_P\{P_W R(s) - Y(s)\} + \frac{K_I}{s^{\lambda}}\{R(s) - Y(s)\} + \frac{K_D s^{\mu}}{1 + \frac{K_D s^{\mu}}{K_P N_F}}\{D_W R(s) - Y(s)\} \tag{4}$$

3.2. Objective function

Here, the ITAE has been chosen as an objective function and it is given below:

Table 1
Implementation Parameters of Interconnected Thermal Power System [5,7,12,22–24,27–29].

P_R = 2000.0 Mega Watt
P_L = 1000.0 Mega Watt (nominal loading);
f = 60.0 Hertz;
B_1, B_2=0.4250p.u. Mega Watt/Hertz;
R_1=R_2 = 2.40 Hertz/p.u.;
T_{G1}= T_{G2}=0.08 s;
T_{T1}=T_{T2} = 0.30 s;
K_{PS1}=K_{PS2} = 120 Hertz/p.u. Mega Watt;
T_{PS1}=T_{PS2} = 20 s;
T_{12} = 0.545p.u.;

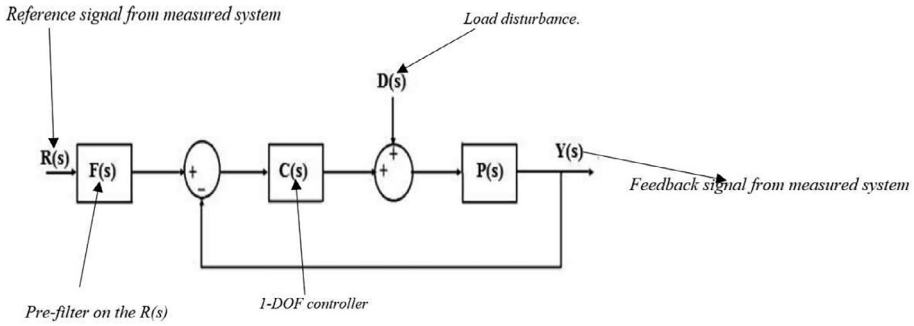


Fig. 2. FO-2DOF-PID controller's structure.

$$J = ITAE = \int_0^{t_{Sim}} w_1.(\lfloor \Delta f_1 \rfloor + \lfloor \Delta f_2 \rfloor + \lfloor \Delta P_{Tie} \rfloor).t.dt \quad (5)$$

where, Δf_1, Δf_2 and ΔP_{Tie} are frequency and tie-line power deviations and t_{Sim} is simulation time.

3.3. Controller's design problem

Since, ITAE has been minimized, the controller design problem is framed as:

Minimize J

Subject to

$$K_{Pmin} \le K_P \le K_{Pmax}, K_{Imin} \le K_I \le K_{Imax}, K_{Dmin} \le K_D \le K_{Dmax} \quad (6)$$

$$P_{Wmin} \le P_W \le P_{Wmax}, D_{Wmin} \le D_W \le D_{Wmax}, N_{Fmin} \le N_F \le N_{Fmax} \quad (7)$$

$$\lambda_{min} \le \lambda \le \lambda_{Wmax}, \mu_{min} \le \mu \le \mu_{Wmax} \quad (8)$$

The abbreviations and controller parameters' ranges are mentioned in Table 2.

3.4. hGWO-PS algorithm in brief

Here, the effort is towards the implementation of a hybrid/mixed algorithm: *h*GWO-PS in optimal design of two dissimilar FO-2DOF-PID controllers existing investigated AGC of interconnected power system. The detailed description, mathematical modeling, flow chart and pseudo codes of GWO and PS algorithms are found separately in [30,31,27–29]. The hybrid *h*GWO-PS algorithm can be understood by the flow chart shown in Fig. 3. The best set of values corresponding to the minimum value of objective function obtained from optimal GWO has been used as the starting points of PS algorithm.

The parameters of FO-2DOF-PID controllers obtained by *h*GWO-PS along with ITAE, settling time (ST) and overshoots of Δf_1, Δf_2, and ΔP_{Tie} have been given in Table 3. Area control error; *ACE*, frequency and tie line power deviations at area 1 and 2 have been shown in Figs. 4–7. From Figs. 4–7, it is clear that smooth dynamic responses with low value of overshoots/undershoots, settling times (ST) and ITAE are obtained by proposed approach. From Figs. 4–7, it is also clear that oscillations in dynamic responses are maximally attenuated by proposed approach which in turn gives the smooth response.

Table 2
Ranges for Parameters of FO-2DOF-PID.

Parameters	*Range*
K_{Pmin}, K_{Pmax}	[0, 2]
K_{Imin}, K_{Imax}	[0, 2]
K_{Dmin}, K_{Dmax}	[0, 2]
P_{Wmin}, P_{Wmax}	[0, 4]
D_{Wmin}, D_{Wmax}	[0, 4]
N_{Fmin}, N_{Fmax}	[10, 200]
$\lambda_{min}, \lambda_{Wmax}$	[0, 1]
μ_{min}, μ_{Wmax}	[0, 1]

4. Results and discussions

Simulation results obtained by proposed approach; *h*GWO-PS/FO-2DOF-PID for the system under investigation and relative comparison with other well known approaches have been discussed below:

4.1. Dynamic responses with and without hGWO-PS optimized FO-2DOF-PID

A relative comparison between the dynamic responses of the same with *h*GWO-PS optimized FO-2DOF-PID controllers and without any controllers have been shown in Fig. 8 (a) to (d) which demonstrates that non oscillatory, stable and fast return back to zero responses with less settling times/overshoots are achieved with the hGWO-PS optimized FO-2DOF-PID controllers while the reverse results are obtained without controllers.

4.2. Comparison of obtained controllers' parameters and ITAE values

The obtained controllers' parameters by some well known approaches have been given in Table 4. The ITAE values obtained by proposed approach: *h*GWO-PS/FO-2DOF-PID and all other well known existing techniques have been shown in Table 5 which shows that proposed approach gives less ITAE value (0.1082) for the same system under investigation as compared to some well known approaches.

The value of ITAE obtained by proposed and some existing approaches has also been represented by bar chart in descending order (shown in Fig. 9) for better illustration and comparison of ITAE values at a glance.

4.3. Comparison of dynamic responses for various step load perturbations (SLP) & locations

Two comparative studies called *Comparison 1* and *Comparison 2* for better illustration of proposed approach have been included in this Section.

Comparison 1: In this section, the dynamic responses with proposed approach: *h*GWO-PS/FO-2DOF-PID have been compared with some popular approaches [15,21,23,27–29,32] in order to

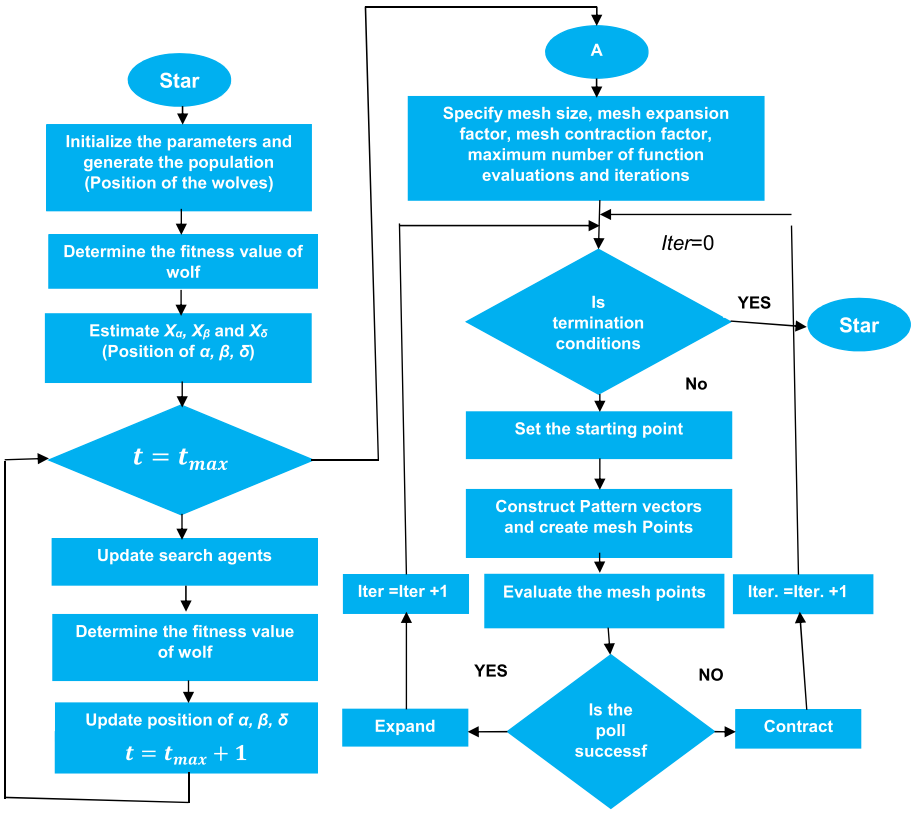


Fig. 3. Flow chart of *h*GWO-PS algorithm.

Table 3
Optimal parameters, settling times/overshoots/undershoots obtained by proposed approach; *h*GWO-PS/FO-2DOF-PID.

Parameters of FO-2DOF-PID Controllers and ITAE		ST (2% Tolerance Band)		Overshoots/Undershoots ('+' represents overshoots and '–' represents undershoots)	
K_{P1} = 0.0082	K_{P2}=0.0018	Δf_1	1.25	Δf_1	0/-0.012
K_{I1} = 0.1967	K_{I2} =0.3019				
K_{D1} = 0.1463	K_{D2} = 0.00085	Δf_2	2.2	Δf_2	0/-0.003
P_{W1} = 0.0668	P_{W2} = 0.2041				
D_{W1} = 0.0018	D_{W2} = 0.1071	ΔP_{Tie}	1.256	ΔP_{Tie}	+0.0096/0
N_{F1} = 155	N_{F2} = 162				
λ_1 =0.9213	λ_2 = 0.0079				
μ_1 =0.2468	μ_2 = 0.0271				
ITAE = 0.1082					

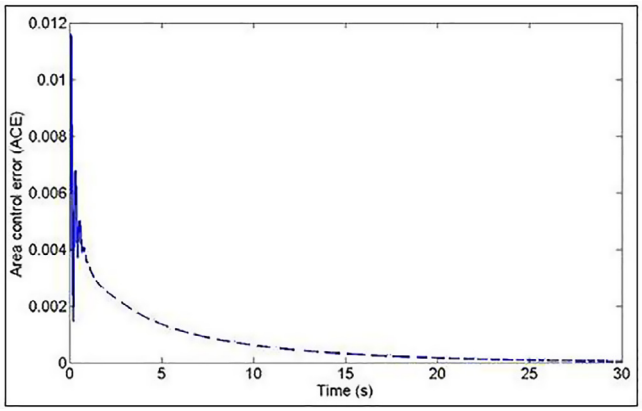


Fig. 4. Area control error (*ACE*).

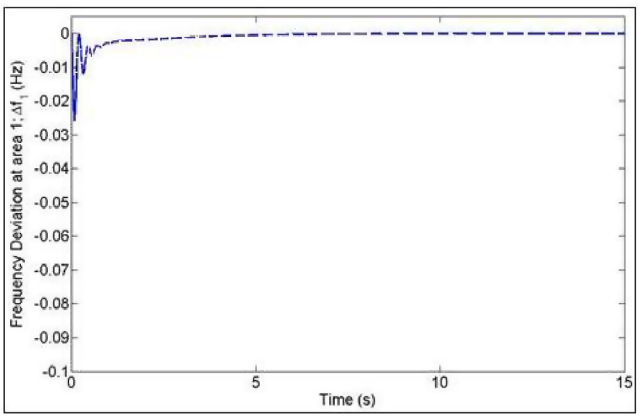


Fig. 5. Response of Δf_1 (Hz).

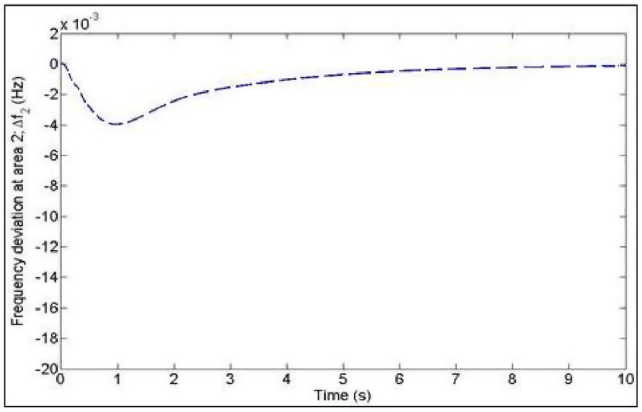


Fig. 6. Response of Δf_2 (Hz).

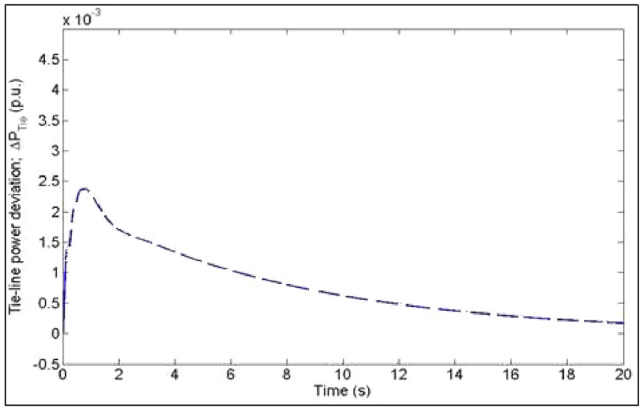


Fig. 7. Response of ΔP_{Tie} (p.u.)

prove the proposed approach superior. The comparative study of the same is shown in Figs. 10–13 for 0.1p.u. SLP in area 1 at t = 0 s and Figs. 14–17 for 0.1p.u. SLP in area 1 and 0.2p.u. SLP in area 2 at t = 0 s. The *Comparison 1* concludes that proposed approach gives far better performance than other mentioned approaches in terms of overshoots, settling times, ITAE and smoothness of the response.

Comparison 2: Further, the superiority of the proposed approach has also been proved by comparing the dynamic responses obtained by the proposed approach and most recently published modern heuristic approaches [24,33] for the same system under investigation. Comparisons are shown in Figs. 18–21 for 0.1p.u. step load increase in area 1 at t = 0 s.

4.4. Comparative analysis of settling times of $\Delta f_1, \Delta f_2$ **and** ΔP_{Tie}

The comparison of settling times of dynamic responses obtained by proposed approach and other approaches have been shown in Figs. 22–24 and tabulated in Table 6. The simulation results show that less settling times along with quick return back to zero responses have been obtained by the proposed approach as compared to all other approaches mentioned in Table 6.

4.5. Comparative analysis of percentage improvement in ITAE and settling times of $\Delta f_1, \Delta f_2$ **and** ΔP_{tie}

The % improvement carried by proposed approach in ITAE and settling times (ST) (for 0.1p.u. SLP in area 1 at t = 0 s) with respect to other approaches have been given in Table 7. The calculation of the same for proposed approach; *h*GWO-PS/FO-2DOF-PID with respect to the other approaches mentioned herein.

For example, % improvement in ITAE and ST carried by proposed approach; *h*GWO-PS/FO-2DOF-PID with respect to SHADE/PID [24] is computed as under:

$$\%Improvement\ in\ ITAE = \frac{ITAEwithOtherApproach-}{ITAEwithotherApproach} \tag{4}$$

$$\%Improvement\ in\ ITAE = \frac{0.1437 - 0.1082}{0.1437} \times 100 = 24.70$$

$$\%Improvement\ in\ STof\ \Delta f_1 = \frac{STwithotherApproach-}{STwithotherApproach} \tag{5}$$

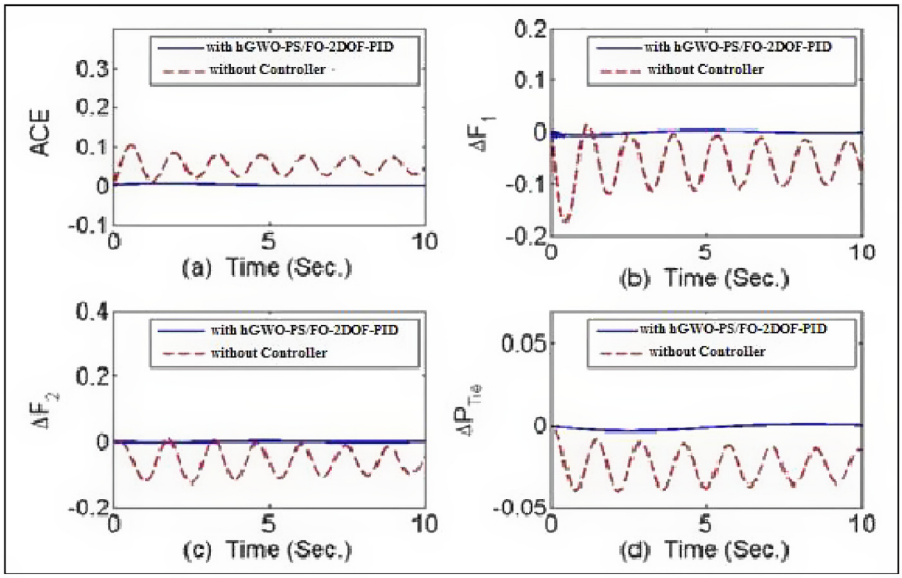


Fig. 8. System Responses with and without FO-2DOF-PID Controllers: *(a) ACE; (b) Response of* Δf_1 *(c) Response of* Δf_2 *and (d) Response of* ΔP_{Tie}.

Table 4
Optimal parameters for different controllers optimized by numerous approaches.

Approaches	K_P	K_I	K_D	P_W	D_W	N_N	K_1	K_2
Z-N/PI [12]	0.7005	0.3802	–	–	–	–	–	–
BFOA/PI [12]	−0.3317	0.4741	–	–	–	–	–	–
GA/PI [12]	−0.2346	0.2662	–	–	–	–	–	–
*h*BFOA- PSO [25]	−0.3317	0.4741	–	–	–	–	–	–
PSO/PI [25]	−0.3597	0.4756	–	–	–	–	–	–
*h*PSO-PS/fuzzy PI [26]	0.9852	0.5595	–	–	–	–	0.9336	0.7203
NSGA-II/PI [8]	−0.428	0.2967	–	–	–	–	–	–
NSGA-II/PIDF [8]	0.4262	0.8063	0.6146	–	–	384.58	–	–
GSA/PI [21]	−0.188	0.6179	–	–	–	–	–	–
GSA/PIDF [21]	1.1884	1.9589	0.3456	–	–	54.3260	–	–
FA/PI [32]	−0.3267	0.4296	–	–	–	–	–	–
FA/PID [32]	1.0556	1.0373	0.9626	–	–	–	–	–
DE/PI [15]	−0.4741	0.3047	–	–	–	–	–	–
PSO/fuzzy PI [26]	0.8176	0.7948	–	–	–	–	0.5085	0.5108
PS/fuzzy PI [26]	0.4509	0.5470	–	–	–	–	0.7317	0.6477
PSO/PI [26]	−0.3597	0.4756	–	–	–	–	–	–
QOHSA/PID [22]	4.0000	3.9156	1.6928	–	–	–	–	–
	1.6662	0.4925	0.0011					
TLBO/PI [23]	−0.2974	0.4378	–	–	–	–	–	–
TLBO/PID [23]	1.1726	1.9370	1.1635	–	–	–	–	–
DSA/PI [24]	−0.4752	0.9945	NA	–	–	–	–	–
	0.5517	−0.0006						
DSA/PID [24]	1.8716	1.8980	0.6773	–	–	–	–	–
	1.0013	1.7907	0.1326					
DSA/PIDF [24]	1.6205	1.9805	0.4638	–	–	231.67	–	–
	1.2095	1.0345	1.1079			268.24		
CLPSO/PI [24]	0.1467	0.9870	NA	–	–	–	–	–
	−0.7835	0.0895						
CLPSO/PID [24]	1.9545	2.00	0.7940	–	–	–	–	–
	1.2690	1.7357	0.8981					
EPSDE/PI [24]	−0.0988	1.000	NA	–	–	–	–	–
	−0.9386	−0.0016						
EPSDE/PID [24]	1.9109	1.8765	0.7414	–	–	–	–	–
	2.0000	1.7178	1.8892					
SHADE/PI [24]	−0.0515	0.9092	NA	–	–	–	–	–
	−0.5156	−0.0059						
SHADE/PID [24]	1.4011	1.9785	0.4619	–	–	–	–	–
	0.6893	1.0082	0.8041					
*h*GWO-PS/2DOF-PID [28]	1.9996	1.4433	1.3706	9.5076	0.0851	433.1273	–	–
***h*GWO-PS/FO-2DOF-PID (Proposed) with λ_1 =0.9213, λ_2=0.0079,**	***0.0082***	***0.1967***	***0.1463***	**0.0668**	**0.0018**	***155***	–	–
μ_1 =0.2468, μ_2=0.0271	**0.0018**	**0.3019**	***0.00085***	**0.2041**	**0.1071**	**162**		

$$\%Improvement\ in\ Settling\ Time(ST)of\Delta f_1 = \frac{4.67 - 1.25}{4.67} \times 100 = 73.23\%$$

$$\%Improvement\ in\ Settling\ Time(ST)of\Delta f_2 = \frac{4.14 - 2.2}{4.14} \times 100 = 46.85\%$$

$$\%Improvement\ in\ Settling\ Time(ST)of\Delta P_{Tie} = \frac{4.35 - 1.256}{4.35} \times 100 = 71.13\%$$

These values have been given in Table 7 in which the % improvement in ITAE and ST using proposed approach with reference to the other approaches have been evaluated with the same procedure given above.

For better illustration, bar charts have also been drawn to summarize the same in Figs. 25–28. The percentage reduction in above said parameters is equivalent to percentage improvement of these parameters.

4.6. Robustness/sensitivity analysis with parametric variation

The robustness/sensitivity analysis of the same with the proposed approach has been carried out for the wide range from −50% to +50% by varying the load and parameters of system. The sensitivity analysis for the system has been performed by following two ways:

- **Sensitivity Analysis under Varying operating Load conditions:** *In this, the sensitivity analysis is carried out by varying the SLP for wide range of [−50%, +50%] keeping Constant T_G, T_T and T_{12}.*
- **Sensitivity Analysis under Variation in System Parameters:** *In this procedure, the sensitivity analysis is carried out by varying T_G, T_T, and T_{12} for wide range of [−50%, +50%] one at a time keeping constant operating load conditions. So, this sensitivity analysis can further be performed by three ways:*
  i. *The sensitivity analysis carried out by varying T_G for wide range of [−50%, +50%] keeping constant operating load condition, T_T and T_{12}.*
  ii. *The sensitivity analysis carried out by varying T_T for wide range of [−50%, +50%] keeping operating load condition, T_G and T_{12} constant.*

Table 5
Comparison of ITAE values obtained by proposed & numerous approaches.

Abbreviation for Bar Chart Representation	*Approaches*	*Controller*	*ITAE*
A	Z-N/PI [12]	PI	3.7568
B	BFOA/PI [12]	PI	1.7975
C	GA/PI [12]	PI	2.7475
D	*h*BFOA- PSO [25]	PI	1.1865
E	PSO/PI [25]	PI	1.2142
F	*h*PSO-PS/fuzzy PI [26]	Fuzzy PI	0.1438
G	NSGA-II/PI [8]	PI	1.1785
H	NSGA-II/PIDF [8]	PIDF	0.387
I	GSA/PI [21]	PI	0.6659
J	GSA/PIDF [21]	PIDF	0.1174
K	FA/PI [32]	PI	0.8695
L	FA/PID [32]	PID	0.4714
M	DE/PI [15]	PI	1.2551
N	PSO/fuzzy PI [26]	Fuzzy PI	0.4470
O	PS/fuzzy PI [26]	Fuzzy PI	0.6334
P	PSO/PI [26]	PI	1.2142
Q	QOHSA/PID [22]	PID	0.1200
R	TLBO/PI [23]	PI	0.8597
S	TLBO/PID [23]	PID	0.2452
T	DSA/PI [24]	PI	0.1951
U	DSA/PID [24]	PID	0.1197
V	DSA/PIDF [24]	PIDF	0.1147
W	CLPSO/PI [24]	PI	0.3837
X	CLPSO/PID [24]	PID	0.1421
Y	EPSDE/PI [24]	PI	0.3319
Z	EPSDE/PID [24]	PID	0.1515
S1	SHADE/PI [24]	PI	0.4180
S2	SHADE/PID [24]	PID	0.1437
G1	GWO (Mentioned in present work by simulation)	2DOF-PID	0.1349
H1	hGWO-PS/2DOF-PID [28]	2DOF-PID	0.112
H2	**hGWO-PS/FO-2DOF-PID (Proposed)**	**FO-2DOF-PID**	**0.1082**

iii. *The sensitivity analysis carried out by varying T_{12} for wide range of [−50%, +50%] at keeping operating load condition, T_G and T_T constant.*

The system dynamic responses along with the ITAE value and ST are observed in each way of performance analysis aforementioned. The sensitivity/robustness analysis has been shown in the Figs. 29–33 and given in the Table 8 which shows that the variation in system dynamic responses, ITAE values and settling times are very close or near to the respective values of the system response obtained under nominal conditions, i.e. the system with the proposed approach *h*GWO-PS/FO-2DOF-PID is more robust.

5. Conclusions

A new hybrid/mixed algorithm: *h*GWO-PS has been used to determine and tune the parameters of FO-2DOF-PID controllers for AGC of interconnected thermal power system. The simulation results obtained by proposed approach; *h*GWO-PS/FO-2DOF-PID for the same have also been compared with the results obtained by some modern heuristic approaches. The robustness/sensitivity of the same with the proposed approach has also been carried out. It has been found that proposed approach provides minimum value of ITAE (0.1082) among all illustrated existing approaches. The dynamic responses of the same have also been studied for different SLPs at different locations and it has been found that the proposed approach; *h*GWO-PS/FO-2DOF-PID achieves better dynamic performance as compared to recently reported approaches. The % improvement in ITAE and STs carried by the proposed approach with respect to the existing approaches has also been calculated. The results show that the proposed approach gives better results than other well known approaches aforesaid. The results also conclude that *h*GWO-PS/FO-2DOF-PID provides more robust control for the same.

CRediT authorship contribution statement

Shailaja Yogesh Kanawade: . **Vikas Soni:** Conceptualization, Visualization.

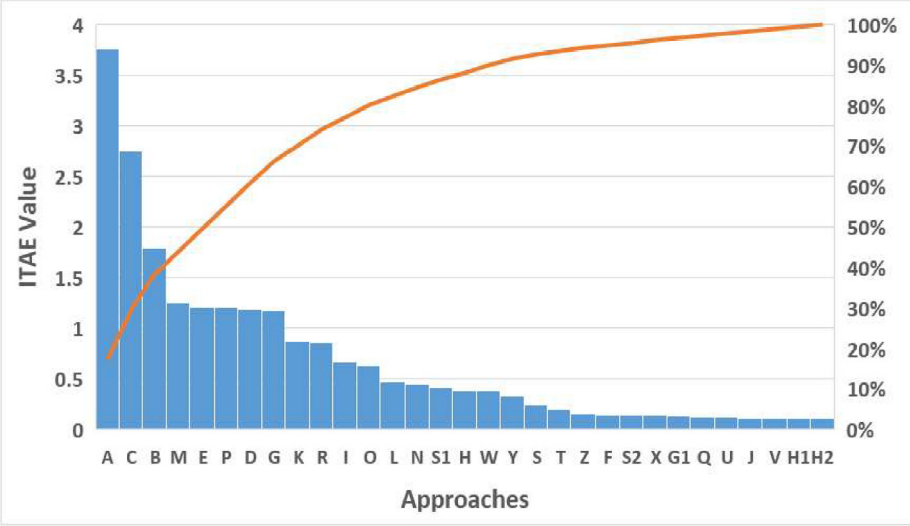


Fig. 9. Comparison of ITAE.

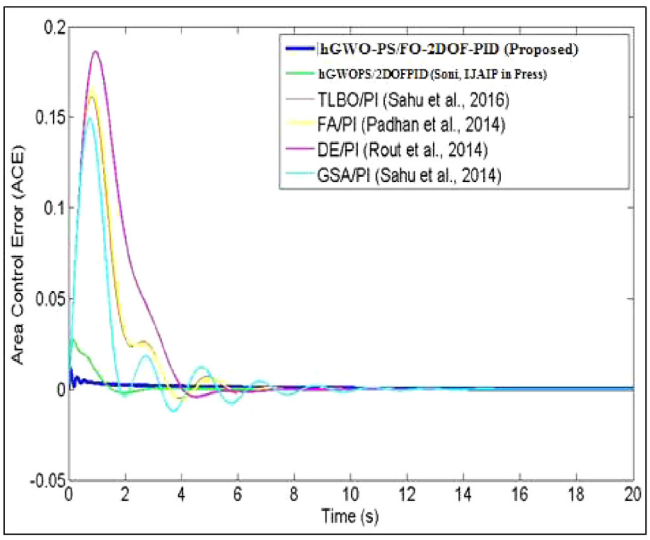


Fig. 10. ACE for *comparison 1.* (SLP_1 = 0.1p.u., SLP_2 = 0).

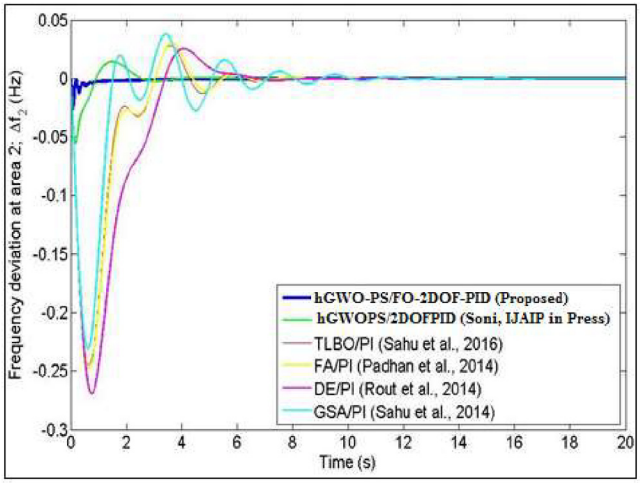


Fig. 11. Response of Δf_1 for *comparison 1.* (SLP_1 = 0.1p.u., SLP_2 = 0).

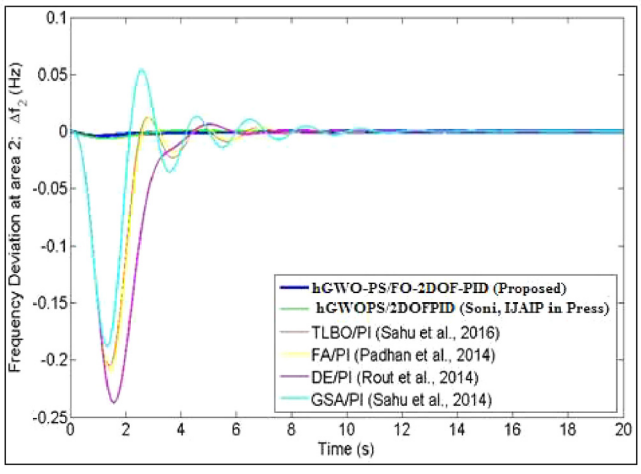


Fig. 12. Response of Δf_2 at area 2 for *comparison 1.* (SLP_1 = 0.1p.u., SLP_2 = 0).

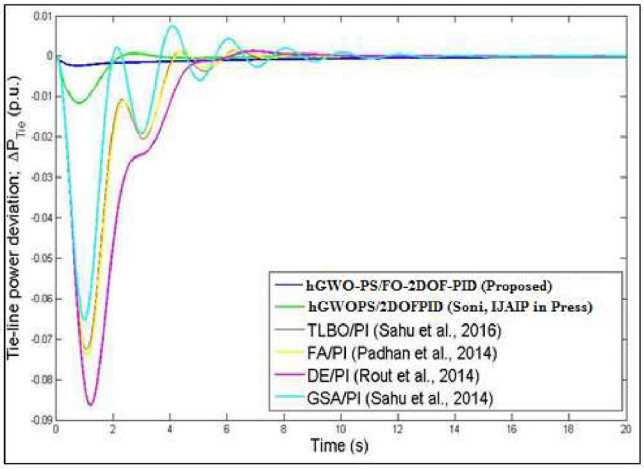


Fig. 13. Response of ΔP_{Tie} for *comparison 1.* (SLP_1 = 0.1p.u., SLP_2 = 0).

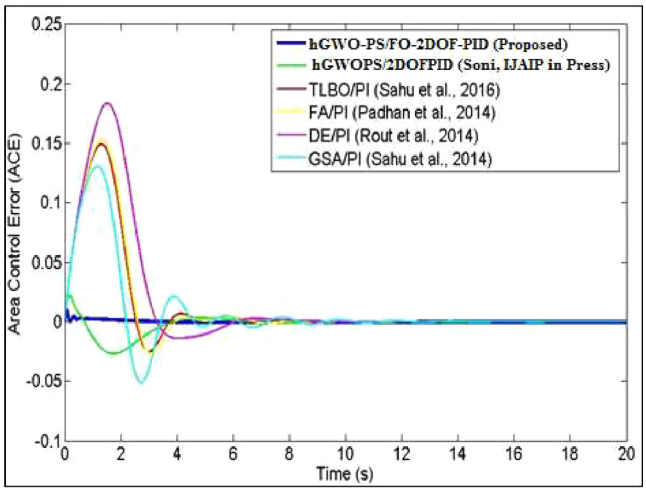


Fig. 14. ACE for *comparison 1.* (SLP_1 = 0.1p.u., SLP_2 = 0.2).

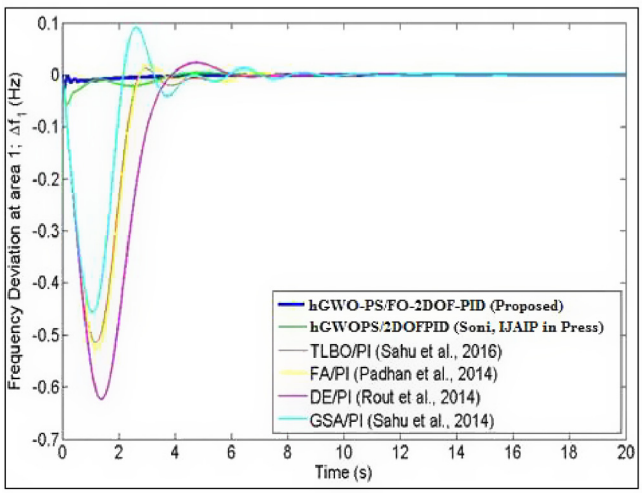


Fig. 15. Response of Δf_1 at area 1 for *comparison 1.* (SLP_1 = 0.1p.u., SLP_2 = 0.2).

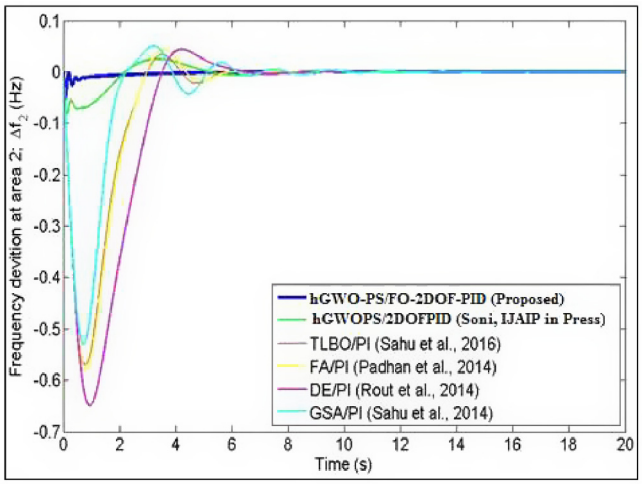


Fig. 16. Response of Δf_2 for *comparison 1.* ($SLP_1 = 0.1$p.u., $SLP_2 = 0.2$).

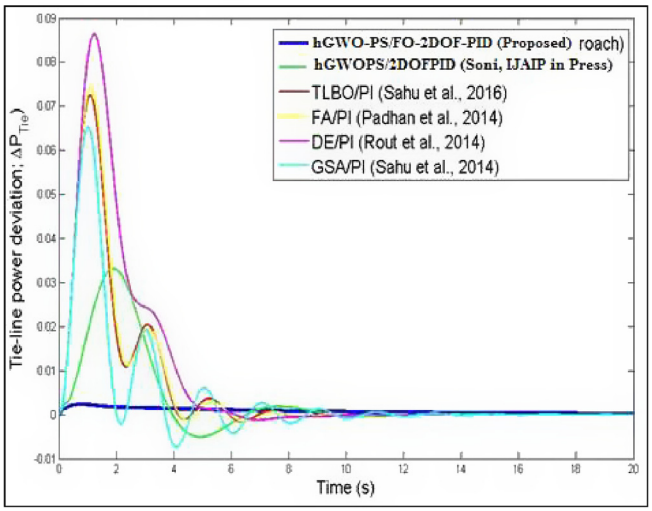


Fig. 17. Response of ΔP_{Tie} for *comparison 1.* ($SLP_1 = 0.1$p.u., $SLP_2 = 0.2$).

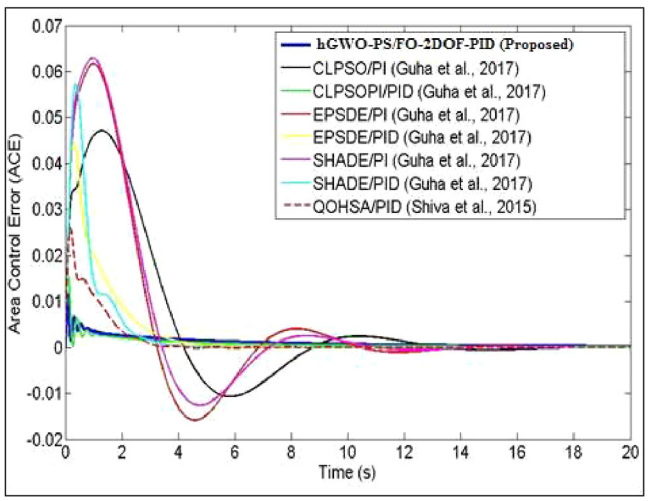


Fig. 18. ACE for *comparison 2.* ($SLP_1 = 0.1$p.u., $SLP_2 = 0$).

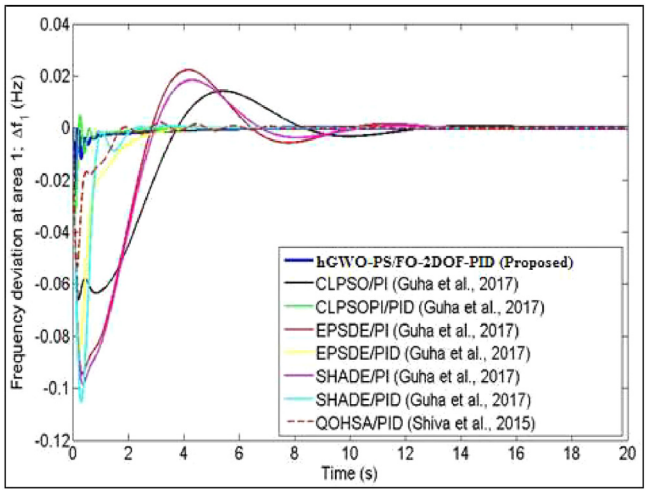


Fig. 19. Response of Δf_1 for *comparison 2.* ($SLP_1 = 0.1$p.u., $SLP_2 = 0$).

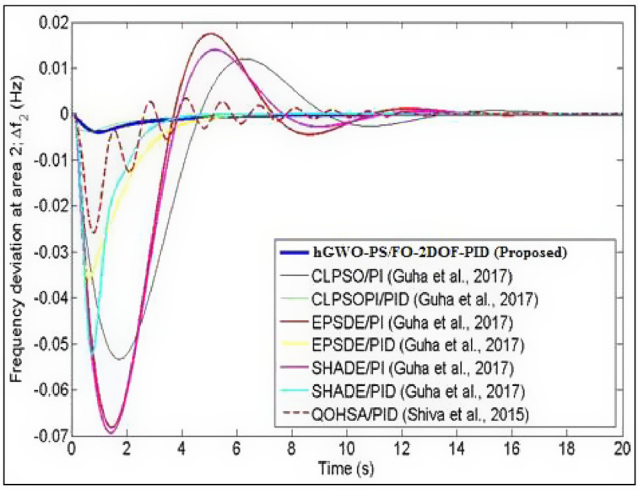


Fig. 20. Response of Δf_2 for *comparison 2.* ($SLP_1 = 0.1$p.u., $SLP_2 = 0$).

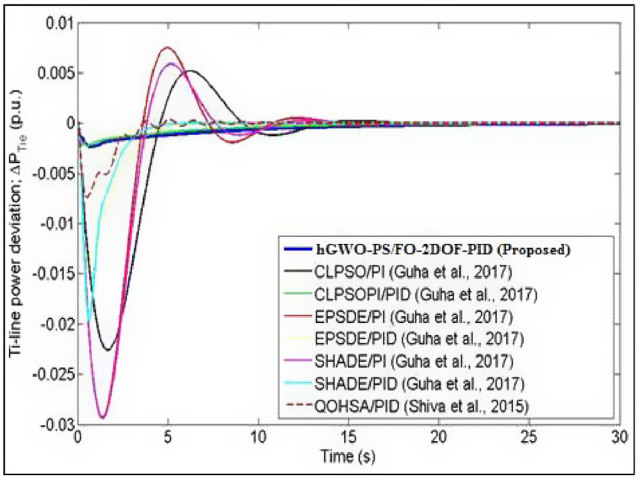


Fig. 21. Response of ΔP_{Tie} for *comparison 2.* ($SLP_1 = 0.1$p.u., $SLP_2 = 0$).

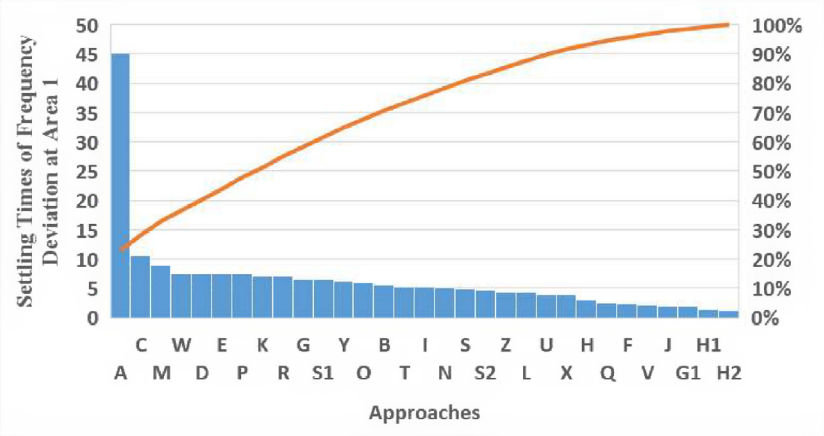


Fig. 22. Comparison of STs of Δf_1.

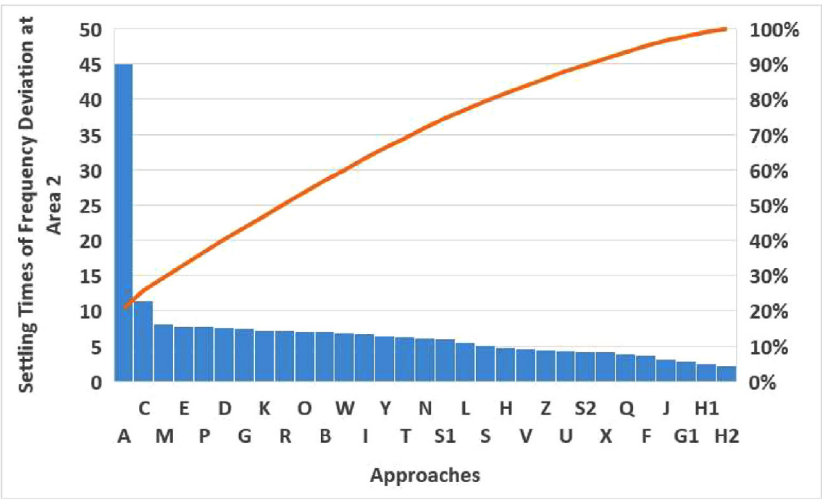


Fig. 23. Comparison of STs of Δf_2.

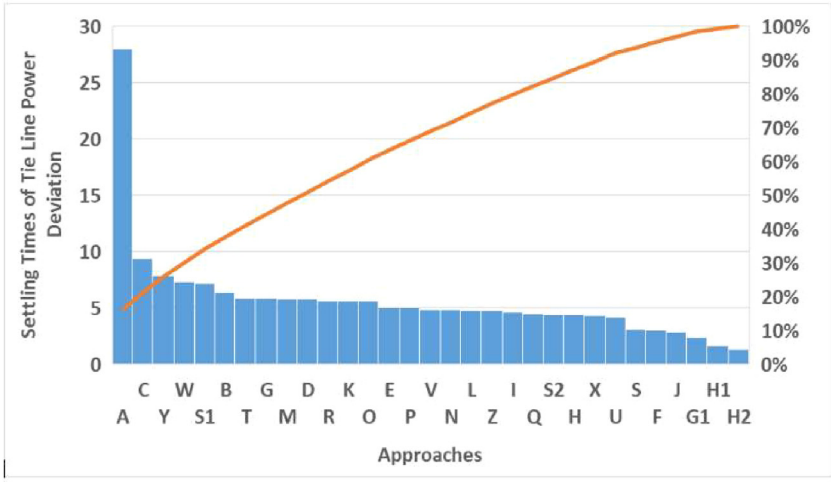


Fig. 24. Comparison of STs of ΔP_{Tie}.

Table 6
Comparison of settling times of Δf_1, Δf_2 & ΔP_{Tie}.

Abbreviation	***Approaches***	***Settling Times***		
		2% Tolerance Band		
		Δf_1	Δf_2	ΔP_{Tie}
A	Z-N/PI [12]	45	45	28
B	BFOA/PI [12]	5.52	7.09	6.35
C	GA/PI [12]	10.59	11.39	9.37
D	*h*BFOA- PSO [25]	7.39	7.65	5.73
E	PSO/PI [25]	7.37	7.82	5.0
F	*h*PSO-PS/fuzzy PI [26]	2.26	3.74	2.94
G	NSGA-II/PI [8]	6.49	7.54	5.79
H	NSGA-II/PIDF [8]	3.03	4.86	4.34
I	GSA/PI [21]	5.17	6.81	4.59
J	GSA/PIDF [21]	1.92	3.19	2.86
K	FA/PI [32]	7.11	7.22	5.62
L	FA/PID [32]	4.25	5.49	4.78
M	DE/PI [15]	8.96	8.16	5.75
N	PSO/fuzzy PI [26]	5.13	6.22	4.83
O	PS/fuzzy PI [26]	6.05	7.10	5.56
P	PSO/PI [26]	7.37	7.82	5.0
Q	QOHSA/PID [22]	2.55	3.82	4.43
R	TLBO/PI [23]	7.08	7.13	5.63
S	TLBO/PID [23]	4.84	5.13	3.03
T	DSA/PI [24]	5.27	6.25	5.84
U	DSA/PID [24]	4.04	4.31	4.17
V	DSA/PIDF [24]	2.05	4.57	4.86
W	CLPSO/PI [24]	7.50	6.96	7.26
X	CLPSO/PID [24]	3.99	4.13	4.28
Y	EPSDE/PI [24]	6.22	6.52	7.83
Z	EPSDE/PID [24]	4.29	4.45	4.72
S1	SHADE/PI [24]	6.48	5.95	7.15
S2	SHADE/PID [24]	4.67	4.14	4.35
G1	GWO/2DOF-PID	1.91	2.76	2.37
H1	*h*GWO-PS/2DOF-PID [28]	1.45	2.5	1.56
H2	***hGWO-PS/FO-2DOF-PID (Proposed)***	**1.25**	**2.2**	**1.256**

Table 7
Comparison of % improvement in ITAE and STs of Δf_1, Δf_2 & ΔP_{Tie} by hGWO-PS/FO-2DOF-PID with respect to various approaches.

Abbre-viation	***Approaches***	***ITAE***	***Settling Times***			***% reduct-ion in ITAE***	***% Reduction in Settling Times for Proposed Approach***		
			2% tolerance band				***2% Tolerance Band***		
			Δf_1	Δf_2	ΔP_{Tie}		Δf_1	Δf_2	ΔP_{tie}
A	Z-N/PI [12]	3.7568	45	45	28	97.11989	97.22222	95.11111	95.51429
B	BFOA/PI [12]	1.7975	5.52	7.09	6.35	93.98053	77.35507	68.97038	80.22047
C	GA/PI [12]	2.7475	10.59	11.39	9.37	96.06187	88.19641	80.68481	86.59552
D	*h*BFOA- PSO [25]	1.1865	7.39	7.65	5.73	90.88074	83.08525	71.24183	78.08028
E	PSO/PI [25]	1.2142	7.37	7.82	5.0	91.08878	83.03935	71.86701	74.88
F	*h*PSO-PS/fuzzy PI [26]	0.1438	2.26	3.74	2.94	24.75661	44.69027	41.17647	57.27891
G	NSGA-II/PI [8]	1.1785	6.49	7.54	5.79	90.81884	80.7396	70.82228	78.30743
H	NSGA-II/PIDF [8]	0.387	3.03	4.86	4.34	72.04134	58.74587	54.73251	71.05991
I	GSA/PI [21]	0.6659	5.17	6.81	4.59	83.75131	75.82205	67.69457	72.63617
J	GSA/PIDF [21]	0.1174	1.92	3.19	2.86	7.836457	34.89583	31.03448	56.08392
K	FA/PI [32]	0.8695	7.11	7.22	5.62	87.55607	82.41913	69.52909	77.65125
L	FA/PID [32]	0.4714	4.25	5.49	4.78	77.04709	70.58824	59.92714	73.72385
M	DE/PI [15]	1.2551	8.96	8.16	5.75	91.37917	86.04911	73.03922	78.15652
N	PSO/fuzzy PI [26]	0.4470	5.13	6.22	4.83	75.79418	75.63353	64.63023	73.99586
O	PS/fuzzy PI [26]	0.6334	6.05	7.10	5.56	82.91759	79.33884	69.01408	77.41007
P	PSO/PI [26]	1.2142	7.37	7.82	5.0	91.08878	83.03935	71.86701	74.88
Q	QOHSA/PID [22]	0.1200	2.55	3.82	4.43	9.833333	50.98039	42.40838	71.64786
R	TLBO/PI [23]	0.8597	7.08	7.13	5.63	87.41421	82.34463	69.14446	77.69094
S	TLBO/PID [23]	0.2452	4.84	5.13	3.03	55.87276	74.17355	57.11501	58.54785
T	DSA/PI [24]	0.1951	5.27	6.25	5.84	44.54126	76.28083	64.8	78.49315
U	DSA/PID [24]	0.1197	4.04	4.31	4.17	9.607352	69.05941	48.95592	69.8801
V	DSA/PIDF [24]	0.1147	2.05	4.57	4.86	5.666957	39.02439	51.85996	74.15638
W	CLPSO/PI [24]	0.3837	7.50	6.96	7.26	71.80089	83.33333	68.3908	82.69972
X	CLPSO/PID [24]	0.1421	3.99	4.13	4.28	23.85644	68.67168	46.73123	70.65421
Y	EPSDE/PI [24]	0.3319	6.22	6.52	7.83	67.39982	79.90354	66.25767	83.95913
Z	EPSDE/PID [24]	0.1515	4.29	4.45	4.72	28.58086	70.86247	50.5618	73.38983
S1	SHADE/PI [24]	0.4180	6.48	5.95	7.15	74.11483	80.70988	63.02521	82.43357
S2	SHADE/PID [24]	0.1437	4.67	4.14	4.35	24.70424	73.2334	46.8599	71.12644
G1	GWO /2DOF-PID	0.1349	1.91	2.76	2.37	19.79244	34.55497	20.28986	47.00422
H1	*h*GWO-PS/2DOF-PID [28]	0.112	1.45	2.5	1.56	3.392857	13.7931	12	19.48718
H2	**hGWO-PS/FO-2DOF-PID (Proposed Approach)**	**0.1082**	**1.25**	**2.2**	**1.256**				

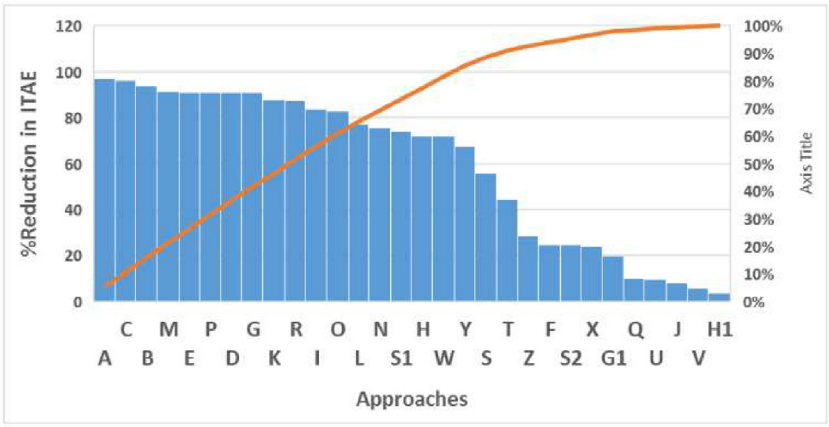


Fig. 25. % Improvement in ITAE carried by proposed approach: *h*GWO-PS/FO-2DOF-PID with respect to other approaches.

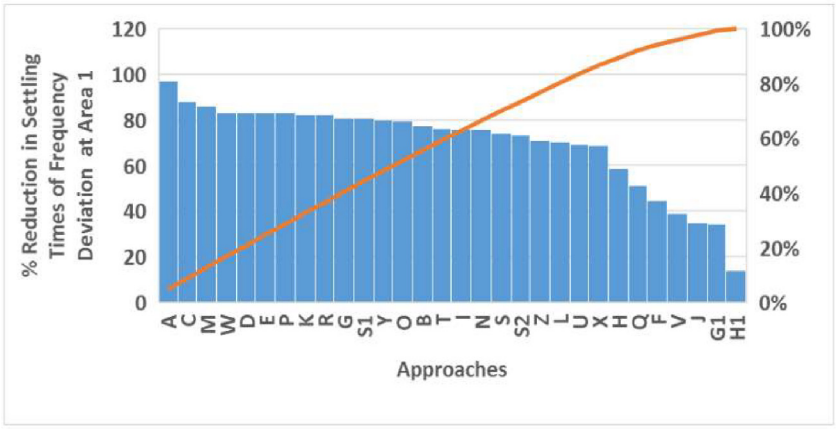


Fig. 26. % Improvement in STs of Δf_1 carried by proposed approach: *h*GWO-PS/FO-2DOF-PID with respect to other approaches.

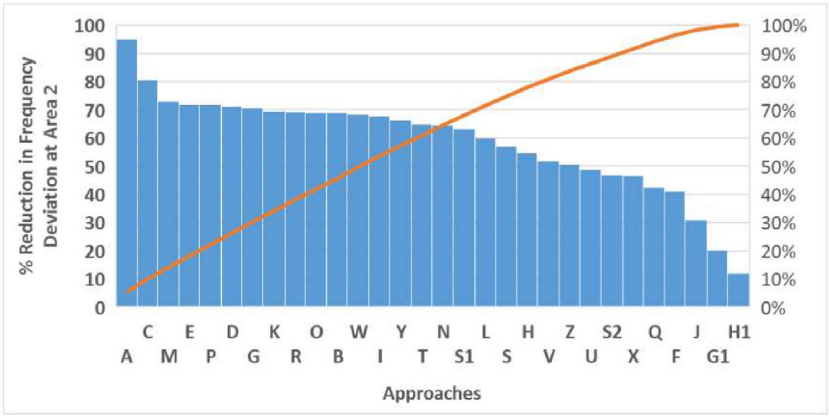


Fig. 27. % Improvement in STs of Δf_2 carried by proposed approach: *h*GWO-PS/FO-2DOF-PID with respect to other approaches.

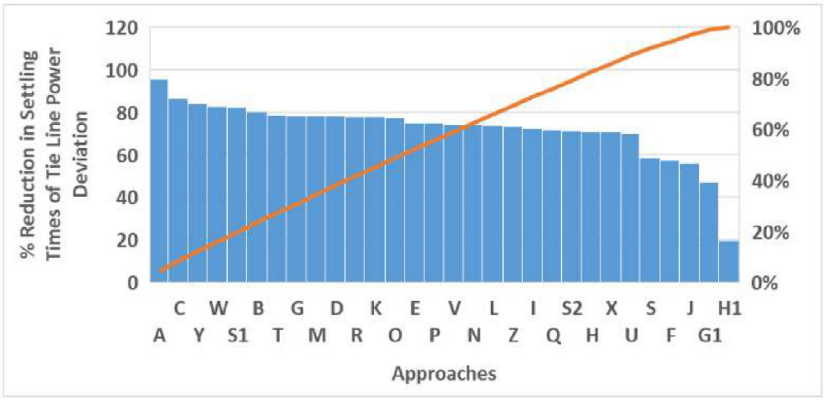


Fig. 28. % Improvement in STs of ΔP_{Tie} carried by proposed approach: *h*GWO-PS/FO-2DOF-PID with. respect to other approaches.

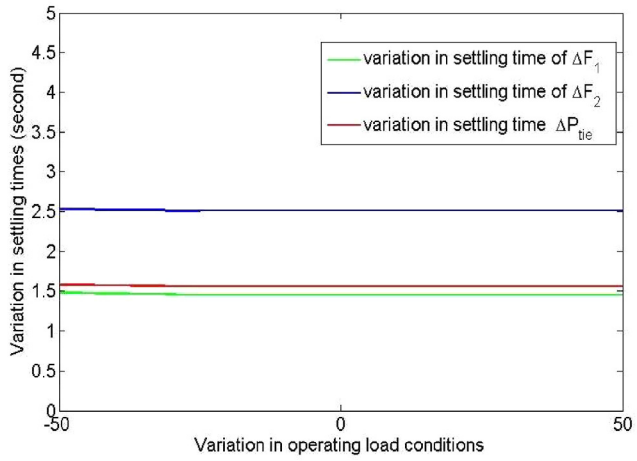


Fig. 29. Variation in settling times with variation in operating load. condition over the range [−50%, +50%].

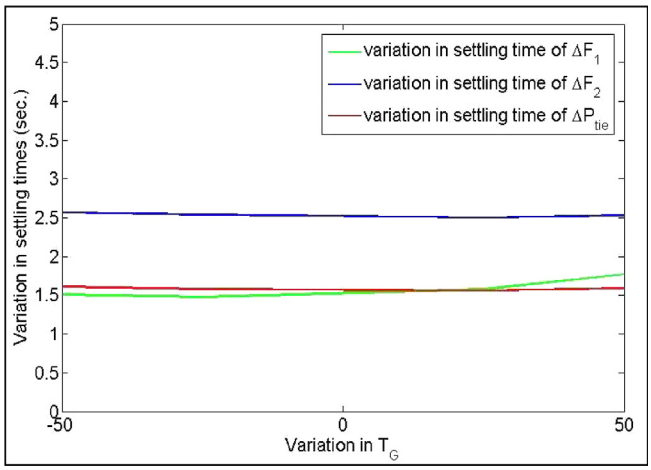


Fig. 30. Variation in settling times with variation in T_G over. the range [−50%, +50%].

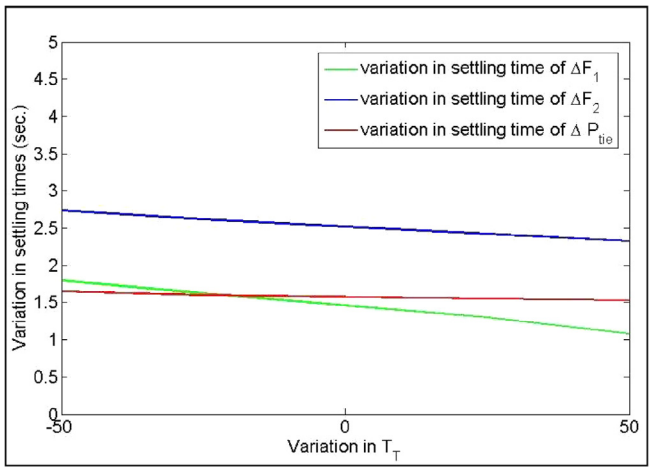


Fig. 31. Variation in settling times with variation in T_T. over the range [−50%, +50%].

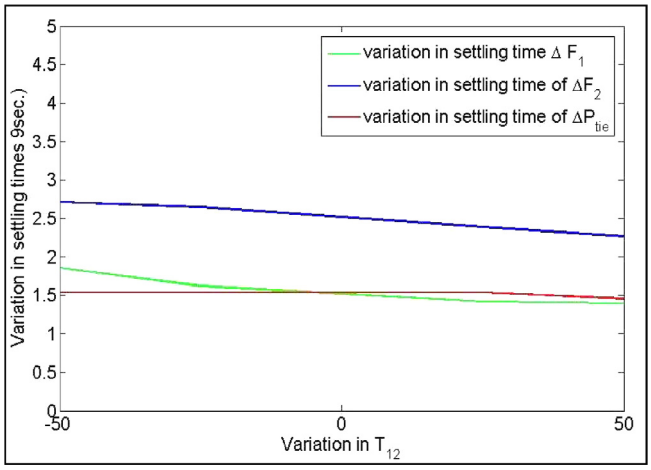


Fig. 32. Variation in settling times with variation in T_{12} over. the range [−50%, +50%].

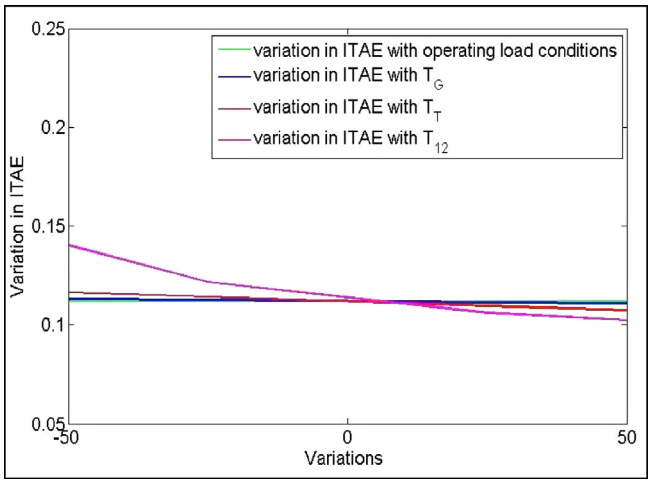


Fig. 33. Variation in ITAE with parametric variation. over the range [−50%, +50%].

Table 8
Robustness/Sensitivity analysis of interconnected. thermal power system.

Parameter variation	*% Change*	*Performance Index with Proposed hGWO-PS/FO-2DOF-PID Approach*			
		ST_s(Sec)			*ITAE value*
		Δf_1	Δf_2	ΔP_{Tie}	
Nominal	0	1.25	2.2	1.256	0.1082
Loading condition	+50	1.25	2.21	1.256	0.1082
	+25	1.25	2.21	1.256	0.1082
	−25	1.25	2.21	1.256	0.1082
	−50	1.28	2.23	1.258	0.1082
T_G	+50	1.47	2.253	1.259	0.1092
	+25	1.38	2.25	1.256	0.1092
	−25	1.28	2.254	1.258	0.1095
	−50	1.31	2.257	1.261	0.1092
T_T	+50	1.18	2.23	1.253	0.1076
	+25	1.32	2.32	1.255	0.1094
	−25	1.52	2.22	1.236	0.1132
	−50	1.82	2.24	1.265	0.1155
T_{12}	+50	1.34	2.28	1.36	0.1035
	+25	1.4	2.29	1.44	0.1052
	−25	1.31	2.25	1.44	0.1116
	−50	1.46	2.21	1.44	0.1303

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.

References

[1] M. Zribi, M. A1-Rashed, M. A1rifa, Adaptive Decentralized Load Frequency Control of Multi-Area Power Systems, Int. J. Electrical Power Energy Syst., 27 (6) (2005) 575–583.
[2] S. Selvakumaran, S. Parthasarathy, R. Karthigaivel, V. Rajasekaran, Optimal Decentralized Load Frequency in a Parallel AC-DC Interconnected Power System through HVDC Link Using PSO.
[3] R.C. Bansal, Optimization Methods for Electric Power Systems: An Overview, Int. J. Emerg. Electric Power Syst. 2 (l) (2005) Article 1021.
[4] P. Kundur, Power System Stability and Control, Tata McGraw Hill, New Delhi, 2009.
[5] O.I. Elgerd, Electric Energy Systems Theory-An Introduction, Tata McGraw Hill, New Delhi, 2000.
[6] P. Bhatt, S.P. Ghoshal, R. Roy, Load frequency stabilization by coordinated control of thyristor controlied phase shifters and superconducting magnetic energy storage for three types of interconnected two-area power systems, Int. J. Electr. Power Energy Syst. 32 (2010) 1111–1124.
[7] R.K. Sahu, S. Panda, U.K. Rout, DE optimized parallel 2-DOFPID controller for load frequency control of power system with governor dead-band nonlinearity, Int. J. Elect. Power & Energy Syst. 49 (2013) 19–33.
[8] S. Panda, N.K. Yegireddy, Automatic generation control of multi-area power system using multi-objective non-dominated sorting genetic algorithm-ll, Int. J. Electr. Power Energy Syst. 53 (2013) 54–64.
[9] L.C. Saikia, J. Nanda, S. Mishra, Performance comparison of several classical controllers in AGC for multi-area interconnected thermal system, Int. J. Electr. Power Energy Syst. 33 (2011) 394–401.
[10] S.A Taher, M.H. Fini, S.F. A1iabadi, Fractional order PlD controller design for LFC in electric power systems using imperialist competitive algorithm, Ain Shams Eng. J., 5 (2014) 121–135.
[11] S. Debbarma, L.C. Saikia, N. Sinha, Robust two-degree-of freedom controller for automatic generation control of multi area system, Int. J. Electr. Power Energy Syst. 63 (2014) 878–886.
[12] E.S. A1i, S.M. Abd-Elazim, Bacteria foraging optimization a1gorithm based load frequency controller for interconnected power system, Int. J. Elect. Power & Energy Syst., 33 (2011) 633–638.
[13] H. Gozde, M.C. Taplamacioglu, Automatic generation control application with craziness based particle swarm optimization in a thermal power system, Int. J. Elect. Power & Energy Systs. 33 (2011) 8–16.
[14] H. Shabani, B. Vahidi, M. Ebrahimpour, A robust PlD controller based on imperialist competitive algorithm for load frequency control of power systems, ISA Trans. 52 (2012) 88–95.
[15] U.K. Rout, R.K. Sahu, S. Panda, Design and analysis of differential evolution algorithm based automatic generation control for interconnected power system, Ain Shams Eng. J. 4 (3) (2013) 409–421.
[16] B. Mohanty, S. Panda, P.K. Hota, Controller parameters tuning of differential evolution algorithm and its application to load frequency control of multi-source power system, Int. J. Elect. Power & Energy Syst. 54 (2014) 77–85.
[17] A.K. Barisal, Comparative performance analysis of teaching learning based optimization for automatic load frequency control of multi-sources power systems, Int. J. Electr. Power Energy Syst. 66 (2015) 67–77.
[18] R. Farhangi, M. Boroushaki, S.H. Hosseini, Load frequency control of interconnected power system using emotional learning based intelligent controller, Int. J. Electr. Power Energy Syst. 36 (l) (2012) 76–83.
[19] K. Naidu, H. Mokhlis, A.H.A. Bakar, V. Terzija, H.A. Illias, Application of Firefly A1gorithm with Online Wavelet Filter in Automatic Generation Control of An Interconnected Reheat Thermal Power System, Int. J. Electr. Power Energy Syst. 63 (2014) 401–413.
[20] H. Gozde, M.C. Taplamacioglu, 1. Kocaarslan, Comparative performance analysis of artificial bee colony algorithm in automatic generation control for interconnected reheat thermal power system, Int. J. Elect. Power Energy Systs., 42 (2012) 167–178.
[21] R.K. Sahu, S. Panda, S. Padhan, Optimal gravitational search a1gorithm for interconnected power systems, Ain. Shams. Eng. J. 3 (5) (2014) 721–733.
[22] M. Shivaie, M.G. Kazemi, M.T. Ameli, A modified harmony search algorithm for solving load frequency control of non-linear interconnected hydrothermal power systems, Sustainable Energy Technol. Assess. 10 (2015) 3–62.
[23] R.K. Sahu, S. Panda, U.K. Rout, D.K. Sahoo, Teaching learning based optimization algorithm for automatic generation control of power system using 2-DOF-PID controller, Int. J Elect Power Energy Syst. 77 (2016) 287–301.
[24] D. Guha, P.K. Roy, S. Banerjee, Study of differential based automatic generation control of an interconnected thermal-thermal system with governor dead band, J. Appl. Soft Comput. (2017).
[25] S. Panda, B. Mohanty, P.K. Hota, Hybrid BFOA-PSO a1gorithm for automatic generation control of linear and nonlinear interconnected power systems, Appl. Soft Comput. 13 (12) (2013) 4718–4730.
[26] R.K. Sahu, S. Panda, G.T.C. Sekher, A novel hybrid PSO-PS optimized fuzzy PI controller for AGC in multi-area interconnected power system, Int. J. Electr. Power Energy Syst. 64 (2015) 880–893.
[27] V. Soni, G. Parmar, M. Kumar, S. Panda, Hybrid grey wolf optimization-pattern search (*h*GWO-PS) optimized 2DOF-PID controllers for load frequency control (LFC) in interconnected thermal power plants, ICTACT J. Soft Comp. 6 (3) (2016) 1244–1256.
[28] V. Soni, G. Parmar, M. Kumar, A hybrid grey wolf optimization and pattern search algorithm for automation generation control of multi area interconnected power systems, International Journal of Advanced Intelligence Paradigm (IJAIP-2017), An Inderscience Journal, 2017 [Article in press].
[29] V. Soni, G. Parmar, M. Kumar S. Panda, hGWO-PS optimized 2DOF-PID controller for non reheat two areas interconnected thermal power plants: a comparative study, in: 2016 IEEE conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES-2016), Delhi Technological University, Delhi, India, pp. 1-6, July 4-6, 2016.
[30] S. Mirjalili, S.M. Mirjalili, A. Lewis, Grey Wolf Optimizer, Adv. Eng. Softw 69 (2014) 46–61.
[31] E.D. Dolan, R.M. Lewis, V. Torczon, On the Local Convergence of Pattern Seareh, SIAM J. Optimization 14 (2003) 567–583.
[32] S. Padhan, R.K. Sahu, S. Panda, Application of firefly algorithm for load frequency control of multi-area interconnected power system, Electr. Power Compon. Syst. 42 (13) (2014) 1419–1430.
[33] C.K. Shiva, V. Mukharjee, A noval quazi oppositional harmony search algorithm for AGC optimization of three area multi unit power system after deregulation, EngSci. Technol. Int. J. 19 (2016) 395–420.

How COVID-19 Brought Indian Wedding Industry to a Halt?

Ravi Kumar Arya, Ritik Ranjan Gupta, Jatin Kumar, Tanay Shubham, Ravi Dugh

Abstract: ***COVID-19 has spread all over the world. Starting from China, it has reached mostly all countries of the world. It has affected major industries and some industries are at the brink of total collapse. Indian wedding industry was one of the most progressive industry in the world but COVID-19 has brought it to its heels. Lockdown due to COVID-19 has restricted major public functions and wedding industry has to follow these guidelines too. In this research, we take exhaustive study showing social and economic impacts of COVID-19 on Indian wedding industry.***

Keywords**: *Indian Wedding, Indian Marriage, Indian Market, Wedding Market***

I. INTRODUCTION

November and December are the busiest months for weddings in India. But due to COVID-19, the wedding industry has taken a toll as most other industries in the world did. The Indian wedding industry is one of the largest in the world and is a $50 billion industry [1]. It is also growing at a rate of 25% - 30% annually. It is the second-largest wedding market after the USA, which is at $70 Billion [2]. Every year 10 million marriages take place in India [3]. But this year has been different. COVID-19 has affected the marriage industry adversely. The wedding industry has been hit hard due to the government's COVID-19 restrictions to stop the spread of coronavirus disease. Currently, India has 9,392,919 total number of cases and 136, 696 people have died until now due to this pandemic. Every day these numbers are rising and putting extra pressure on different businesses. The wedding industry has many interlinked industries, and all are feeling the pinch. Caterers, designers, banquet/tent house owners, makeup artists, wedding vendors, or planners are all impacted by the sudden change of circumstances this year. Most of the weddings are either getting cancelled or delayed. The remaining ones are not attracting the crowds that they typically did due to the government guidelines. In these scenarios, these different businesses are collapsing badly. To understand the effects of COVID-19 on these businesses, we have an exhaustive study of different aspects of these interlinked businesses and investigate their social and economic impacts. In the end, we conclude our research by proposing some mitigation techniques that can help to save these businesses.

Following are the current rules to be followed for attending a wedding in India due to COVID-19 restrictions [4]:

- People should adhere to social distancing norms during the entire function.
- They should wear a mask during the ceremony.
- The guests should use hand-sanitizer.
- Ceremonies should be kept as short as possible.
- Not more than 50 people can attend a wedding.
- No drink or food is consumed during the ceremony until absolutely necessary.
- Many states set their own rules, including the number of guests allowed in a wedding function.

Because of so many rules and fear of COVID-19, people are avoiding weddings this year, or they are getting their weddings done in the simplest way also if they can do it without any function and guests. As it is common in Indian wedding functions, many people gather to celebrate it. But due to the safety, all have to follow physical distance from each other, which is not possible in the crowd, so most of the wedding functions are adversely affected or cancelled.

Indian weddings are all about extravaganza and if people get so uncertain about the future, every part of this industry will bleed and is bleeding currently. A typical wedding consists of many different expenses from different functions. Wedding cost mostly depends upon the socio-economic status of a family. Parents start planning for their child's marriage years in advance. A person typically spends one-fifth of their total accumulated lifetime wealth on wedding ceremonies [5]. Typical Indian wedding costs INR 5 lakh to INR 5 crore.

Major factors that decide overall wedding cost are:

- Number of guests
- Number of functions in a wedding
- Traditional activities
- Venue location, banquet hall decorations
- Food expenses

Some people are also realizing that money management is critical for survival [6] through this pandemic time and it is important to take care of extra or unwanted expenses. So, we will see the change in marriage expenses during COVID-19 as people are trying to keep their savings intact and try to spend the minimum to face an uncertain future.

II. INDIAN WEDDING INDUSTRY

A. Online Wedding Website/ Bureau

In Fig. 1, we see matrimony.com Limited share price change [7].

Revised Manuscript Received on January 15, 2021.

***** Correspondence Author

Ravi Kumar Arya*, Electrical and Communication, National Institute of Technology Delhi, New Delhi, India. Email: raviarya@nitdelhi.ac.in

Ritik Ranjan Gupta, Electrical and Electronics, National Institute of Technology Delhi, New Delhi, India. Email: 181230040@nitdelhi.ac.in

Jatin Kumar Electrical and Communication, National Institute of Technology Delhi, New Delhi, India. Email: jatin9562@gmail.com

Tanay Shubham, Electrical and Electronics, National Institute of Technology Delhi, New Delhi, India. Email: 171230004@nitdelhi.ac.in

Ravi Dugh, Goergen Institute Data Science, The University of Rochester, NY, USA. Email: ravi.dugh@gmail.com

We can see share prices started decreasing at the inception of COVID-19, and it was too low when COVID-19 was on the peak (March 20, April 20, May 20), then these prices started to increase as the weddings resumed. To cope with a sudden change in circumstances, online matrimonial websites/bureaus came with new offers and deals for their customers and also introduced new features on their app/website like online chatting/video calling features etc. Even they offered deals during the lockdown periods such as free premium membership. And that's the reason for the considerable increase in their share prices once people started to realize that the COVID-19 situation will stay for long.

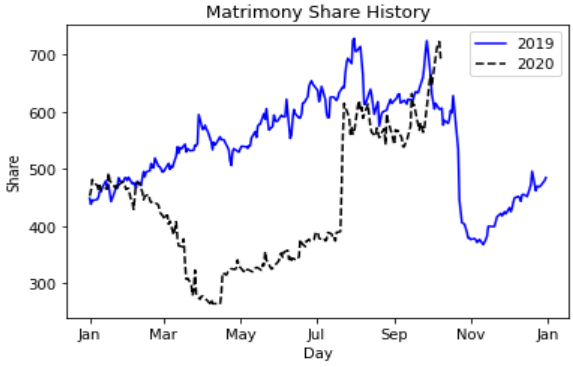


Fig. 1. Matrimony.com Ltd. Share Price.

According to The New York Times [8], there are over 1500 matrimony websites in India. According to a survey, traffic on matchmaking apps and sites increased during the lockdown. Jeevansaathi.com saw a 30% growth in new user acquisitions year-on-year during the period. Shaadi.com witnessed a traffic surge of 15-20% on its platform, compared to the pre-COVID period, while BharatMatrimony.com also reported a 30% increase in registrations. These are the top 3 matrimonial websites in India. The time spent on these platforms increased too. Jeevansaathi.com has seen a 20% increase in time spent on its platform, and engagement on Shaadi.com is up by 25%. But with strict social distancing norms in place during this time, these numbers may mean little. India's matchmaking market is valued at $184 million [9] and is expected to grow at a compound annual growth rate (CAGR) of 10% to reach $269 million by 2024. The website estimates that the market has about 8.2 million users in the country currently. As from the above data, we can say that because of the pandemic the population is more shifting towards the online facilities [10].

B. Jewellery Industry

"If India sneezes, the gold industry will catch a cold," says Ajay Mitra of the World Gold Council [11]. With the current situation, gold and related Jewellery industry have certainly caught a cold. India, the world's second-biggest gold consumer, experienced the lowest sales of gold in the past 25 years. Total purchases are estimated to plummet 30% in 2020 from the 690 tons last year. That would make it the smallest annual purchase since the 477 tons bought in 1995, according to data compiled by Bloomberg based on World Gold Council reports. As in India, Jewellery is sold mostly on occasions such as weddings. As COVID-19 has hit the country in 2020 almost all the wedding functions got cancelled or postponed and the demand for Jewellery has gone down. It also affected the price of gold and silver [12] as shown in Fig. 2.

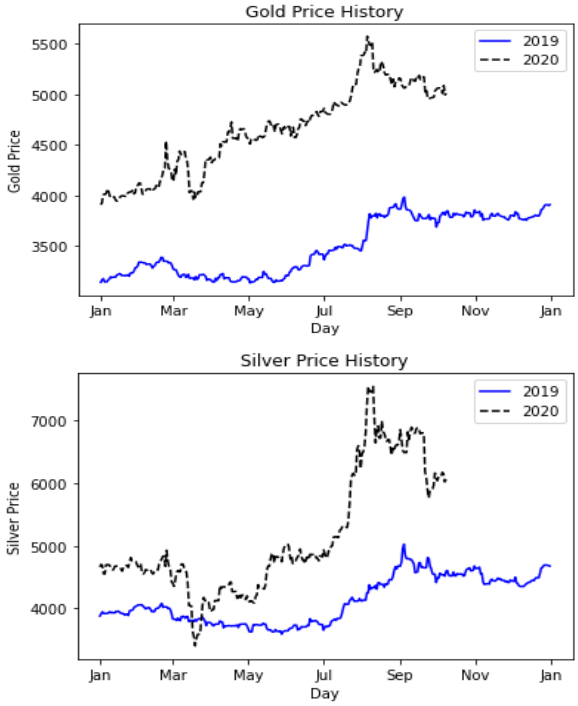


Fig. 3. Share price of Titan Corporation.

C. Wedding Photography

These days wedding photography not only includes photography during the wedding but pre-wedding shoots are also gaining traction. Due to the low count of wedding guests, people are also avoiding expenses on grand photography. Typical wedding shoots can cost INR 15, 000 to INR 100,000 [14]. Affording even that much money is getting difficult for families where they have lost jobs.i

D. Destination Wedding

There is no doubt that India has many beautiful locations but many people choose a destination wedding outside India to make it a memorable moment [15]. Due to COVID-19, the aviation sector has limited flight traffic. People are avoiding unnecessary travel and government restrictions are also keeping the air travel under control. People are also wary of changing scenarios as new studies are finding COVID-19 to be air-borne [16]. Different countries are implementing different lockdown conditions. Due to all these reasons, going for a destination wedding does not look like a healthy choice in these difficult times. All this has adversely impacted this sector economically and it will take some time to bring this earlier blooming sector back to its earlier glory.

E. Marriage Banquets and Tent House

According to the Hindustan Times, there are around 500 banquets and 5700 tent houses in Delhi [17]. Majority of banquets have 2-3 halls with a capacity of 350-500 guests each, but according to the government guidelines passed in Unlock 4 that from September 21, 2020 gathering of up to 100 people is allowed [18], and using these banquets at such low attendance is incurring big losses to owners of such places. And because of changing of rules in every unlock guidance, creating confusion among people which led to the cancellation of pre-booking which also became a blocker for the banquet's owners.

India hosts over 10 million weddings every year which gives employment to thousands of people as it requires lots of manpower and venues to host these events, but due to current situation it created a large number of employment and a huge loss to banquet's owners as they have to spend money on maintenance, electricity and staffs despite zero bookings. These constant expenses with no future income led owners of such places to start firing people from their jobs. Now, while the banquet owners have started wedding bookings, tent house owners are still facing problems with their business. Most of the tents are used for big weddings and other events that are held at parks and open areas. But with current restrictions, such activities in parks and open areas are not allowed and are affecting tent owners badly.

F. Band Artists and Caterers

The COVID-19 pandemic fear and the need to maintain social distancing have forced people to postpone their wedding ceremonies or some couples to opt for small intimate affairs instead of their big fat events, which has affected businesses of bands and caterers in the country. These industries are run only by manpower and because of no bookings, it created huge unemployment among these people. Most of the band artists and caterers come from the village to big cities but due to current situation, many are forced to seek another source of livelihood or to go back to the village as they don't have enough money to survive in the cities. The pandemic also created a phobia of not getting jobs in future as most of the small workers are afraid that now marriage won't happen the way they used to be. Most of the band groups have taken loans from a bank or private authorities to get instruments and band dresses which are lying useless at present, but they have to pay back the loan despite any income. Many caterers have taken Godowns on rent for storage which is also becoming a burden on them.

G. Marriage Retail Shopping

After the lockdown started, the offline retail market experienced a huge downfall. According to a KPMG report of 2018, the big Indian wedding market is often considered recession-proof [19]. But due to the pandemic, this huge market went down. As during the lockdown period, even some marriage functions were going on either online or in-house, the shops were closed which resulted in zero sale of wedding items. As the lockdown was lifted and shops started to open, there are still almost zero sale in the market due to changed wedding scenarios. People are preferring not to go to market due to COVID-19 phobia and they are trying to do online shopping as much as possible which is also affecting small retailers with a huge loss. After the unlock started the shopkeepers started offering heavy discounts to clear their stocks and get some money as many of them filled their shops with stock just before COVID-19 spread by borrowing money and they have to pay them back which is a challenge for them in current situations.

III. FACTORS EFFECTING WEDDING AFTER COVID-19

In this section, we list different factors affecting the weddings after COVID-19 spread. Most of the weddings are getting postponed or even get cancelled due to change in people's situation after COVID-19.

A. Employment

As the COVID-19 hit, it led to an economic crisis in the country. Many lost their jobs and many went into the huge loss which adversely effected the employment rate of the country. During the lockdown period (April, 2020 – May, 2020) the unemployment rate went to peak in just one month [20] as shown in Fig. 4. As many people started losing their jobs, it also created uncertainty about their marriage. Due to no job, people started cancelling the wedding or postponed them.

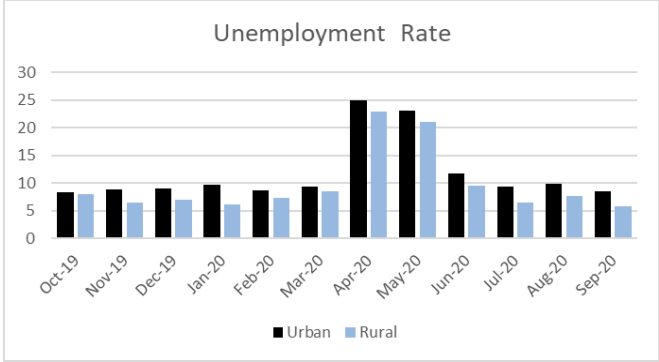


Fig. 4. Unemployment rate in India.

B. Government Rules and Regulation

As the unlock started with tons of rules and regulations by the government, it becomes very difficult to have a function as it used to be before the COVID-19 spread. In India, wedding is the dream of the parents, they always plan to do as much as possible in their children's wedding function. India has a tradition of inviting family members and close ones to their weddings and celebrate with all. But due to these rules and regulations it is almost impossible to have a function like before. So, because of so many ever-changing rules and fear of COVID-19, most of the people are avoiding weddings this year or they are getting their weddings done in the simplest way they can without any function and guests.

C. Savings

In India, most of the parents save a part of their hard-earned money for their children and their marriage. But during the lockdown, many families are forced to spend their savings to fulfil their daily requirement as the market was closed so there was no source of income which created a huge burden on parents and families.

D. Patient Phobia

After the COVID-19 spread, the people's mindset has changed too as everyone is treating each other as a source of virus because of their own safety. This mindset is also followed in weddings. Recently, there are frequent news of virus spread in wedding functions. In one of the wedding, groom died just after two days of wedding by COVID-19 and 95 guests also got COVID-19 positive [21], further creating phobia for people that weddings are spreading coronavirus and can be deadly to attend. Apart from the industry collapse, there are also other social and economic factors to be considered.

Many people have lost jobs and it is getting difficult to get new jobs. This has also lowered the chances for young people to get married. Most parents of the bride want their future son-in-law to have future job security as it gives better financial stability. Losing a job in such trying times is also bringing the marriage count low. For many parents, savings have almost exhausted; therefore, they will not be able to afford weddings of their children for some time.

E. India's GDP

Almost every industry got affected due to COVID-19 which led to huge drop in the India's GDP as shown in Fig. 5. There was a huge drop in the first quarter of 2020 when the lockdown was implemented in the country.

Table- I: GDP rate of past 5 Years

Quarter	Year				
	2016-17	*2017-18*	*2018-19*	*2019-20*	*2020-21*
Q1	8.68	5.78	7.10	5.25	-23.92
Q2	9.67	6.47	6.20	4.42	
Q3	8.58	7.64	5.59	4.08	
Q4	6.29	8.18	5.67	3.09	

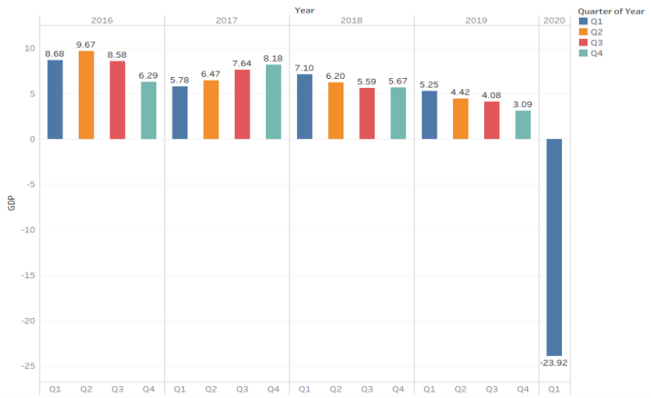
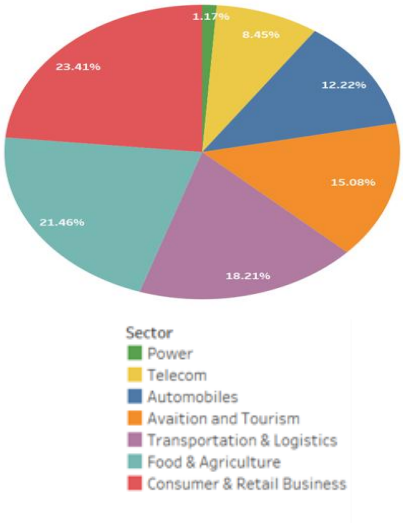


Fig. 5. India GDP Trend.

On further investigation, we can come to the conclusion that somehow the different sectors of the GDP are directly related to the marriage industry. In Fig. 6, we see the share of each sector effecting the GDP. [22]



IV. IS ONLINE WEDDING THE NEXT FUTURE?

There are many ways in which technology can be used today for online weddings. Dance performances can be recorded and played during the sangeet (music) ceremony. The wedding can be broadcast on online live video platforms. Guests can dress up in finery and participate through video calls from their drawing rooms. And when the time comes for the feast, delivery services can deliver the exact same menu to each of the invited guests. Imagine having a small wedding in a truly blissful location with a few people in attendance and then hundreds of others participating through 3D virtual reality goggles. Drones can deliver amazing camera angles and swoop around the beautiful landscape. Flowers, food, scents — all the other wedding accoutrements can be brought to those participating remotely through a wedding planning service. The money saved on travel, hospitality, and transportation can be utilized to ensure a truly memorable in-home wedding experience. Even using augmented reality, the guests can be virtually present at the wedding while sitting in their drawing rooms. Not only that, people can have wedding anniversaries to just get the feel of the real weddings once the COVID-19 vaccine comes and we feel much safer.

V. CONCLUSION

The key question is if the wedding industry is impacted, then what is the future of the wedding industry in India? From the above sections, we have shown the new normal that the industry is witnessing. These changes will stay for some time and related industries will have to change accordingly. We will also see a digital wedding taking away bigger pie in coming times as it adheres better to these new times. Even if it is not fully digital, we will definitely see significantly cheaper weddings in India. Businesses like online matchmaking will have new opportunities to make money. We will also see a major part of the existing workforce changing careers or using new second career to feed themselves as it will be difficult for such a large workforce to learn new skills in such a short time with so many restrictions in place. It is going to be a difficult time for the Indian wedding industry and only time will tell how we cope with this new normal.

REFERENCES

1. The recession-proof Big Fast Indian Wedding just got more corporatized. cnbctv18.com. Accessed October 04, 2020 https://www.cnbctv18.com/retail/the-recession-proof-big-fast-indian-wedding-just-got-more-corporatised-1012991.htm
2. Pandit, Virendra. "Big fat Indian wedding market has foreign 'suitor' Zankyou lining up", accessed October 5, 2020, https://www.thehindubusinessline.com/news/national/big-fat-indian-wedding-market-has-foreign-suitor-zankyou-lining-up/article9622383.ece
3. India's love affair with gold, CBS News, February 13, 2012, https://web.archive.org/web/20120213212435/http://www.cbsnews.com/8301-18560_162-57376057/indias-love-affair-with-gold/
4. Getting married during the coronavirus outbreak? Check state-wise wedding rules here, accessed on October 5, 2020, https://www.timesnownews.com/india/article/getting-married-during-the-coronavirus-outbreak-check-state-wise-wedding-rules-here/621868

5. How much does a typical upper-middle-class wedding cost? accessed October 4, 2020 https://www.aegonlife.com/insurance-investment-knowledge/how-much-does-a-typical-upper-middle-class-wedding-cost/
6. Spend cautiously to sail through the coronavirus pandemic, accessed on October 5, 2020, https://m.economictimes.com/wealth/spend/spend-cautiously-to-sail-through-the-coronavirus pandemic/articleshow/78329758.cms
7. NSE, matrimony. Com Limited https://www.nseindia.com/get-quotes/equity?symbol=MATRIMONY
8. Websites in India Put a Bit of Choice Into Arranged Marriage, accessed on October 5, 2020
9. https://www.nytimes.com/2015/04/26/world/asia/india-arranged-marriages-matrimonial-websites.html
10. How matrimonial apps have turned to organize weddings during the pandemic, , accessed on October 5, 2020
11. https://www.financialexpress.com/brandwagon/how-matrimonial-apps-have-turned-to-organising-weddings-during-the-pandemic/1999113/
12. https://www.financialexpress.com/brandwagon/how-matrimonial-apps-have-turned-to-organising-weddings-during-the-pandemic/1999113/
13. Precious metal: India's love affair with gold, CBS News, accessed on Feb. 13, 2012 https://web.archive.org/web/20120213212435/http://www.cbsnews.com/8301-18560_162-57376057/indias-love-affair-with-gold/
14. India braces for lowest gold sales in 25 years as virus hits, accessed on October 5, 2020, https://economictimes.indiatimes.com/markets/commodities/news/india-braces-for-lowest-gold-sales-in-25-years-as-virus-hits/articleshow/74795271.cms
15. NSE,TITAN https://www.nseindia.com/get-quotes/equity?symbol=TITAN
16. "One wedding photo shoot. Make it viral please!" accessed October 4, 2020 https://timesofindia.indiatimes.com/life-style/spotlight/one-wedding-photo-shoot-make-it-viral-please/articleshow/70710938.cms
17. "Indian celebs who married outside India" accessed October 4, 2020 https://www.orissapost.com/indian-celebs-who-married-outside-india/
18. Morawska, Lidia, and Donald K. Milton. "It is time to address the airborne transmission of COVID-19." Clin Infect Dis 6 (2020): ciaa939.
19. "Covid-19 shadow over big, fat weddings in Delhi." https://www.hindustantimes.com/delhi-news/covid-19-shadow-over-big-fat-weddings-in-delhi/story-O5NucA9dhDoYbJSvIO9ipI.html
20. "Govt allows sports gathering of up to 100 people from September 21"
21. http://timesofindia.indiatimes.com/articleshow/77826097.cms
22. "The recession-proof Big Fast Indian Wedding" https://www.cnbctv18.com/retail/the-recession-proof-big-fast-indian-wedding-just-got-more-corporatised-1012991.htm
23. CMIE, https://unemploymentinindia.cmie.com/
24. "Bihar: Groom dead, 95 guests test Covid-19 positive" https://timesofindia.indiatimes.com/city/patna/bihar-groom-dead-95-guests-test-covid-19-positive/articleshow/76706537.cms
25. "THE IMPACT OF COVID 19 IN INDIAN ECONOMY – AN EMPIRICAL-STUDY", http://www.iaeme.com/MasterAdmin/Journal_uploads/IJEET/VOLUME_11_ISSUE_3/IJEET_11_03_023.pdf

AUTHORS PROFILE



Ravi Kumar Arya received the Ph.D. degree in Electrical Engineering from the Pennsylvania State University, USA in 2017. In 2019, he joined NIT Delhi, India, where he is currently an Assistant Professor in the Department of Electronics and Communication Engineering. In 2017, he worked with Isotropic Systems, MD, USA as RF Optical Design Engineer where he developed various antennas for satellite communication systems, wrote various analysis tools for lens and array design, implemented and managed Amazon Web Services (AWS) computer resources for the company. He received his M. Tech in RF and Microwave Engineering from Indian Institute of Technology, Kharagpur, India in 2006 and B. E. from Delhi Technological University, Delhi, India in 2003 respectively. After completing B. E., he worked with Electronics Corporation of India Limited, India for 6 months and then with Centre for Development of Telematics (C-DOT), India from 2006 to 2010. He has published several technical papers in highly ranked journals and refereed conference proceedings, including three book chapters. His research interests include RF circuit design, antenna design, lens design and the analysis of frequency-selective surfaces



Ritik Ranjan Gupta is pursuing his under graduation in Electrical and Electronics Engineering from National Institute of Technology and will complete it in 2022. He has done projects in Data Science and Analytics. His Major area of Interest is Management and Engineering. He has participated in International conference conducted by NIT Jalandhar. His research interests include Data Science, Statistical Modelling, Data Analysis. He has done various internships in the field of Data Science and Web Development. He is also a Social worker, Volunteered in Many NGOs and Social Groups.



Jatin Kumar is a final year student in Electronics and Communication branch at National Institute of Technology, Delhi. He has done several projects in machine learning and data science like Covid-19 face mask detection app, Hindi hand written text analyzer, housing price predicting website, sports celebrity classifier webpage, sales insights generating Power BI dashboard.



Tanay Shubham is a under graduate student Electrical and Electronics Engineering, National Institute of technology Delhi. His minor project in 7th semester (4th year) was on skin cancer detection using computer vision. His summer internship in June 2019 in National Hydro Power Corporation was on hydropower generation also in June 2020 he did his summer internship on smart grid requirement in Indian power system under Dr. Amit Kumar Singh, Assistant professor, NIT Delhi.



Ravi Dugh has over 15 years of diverse experience and has worked in various industries, including energy, telecom, professional services, and building materials. Ravi has worked with many leading companies, including Halliburton and Lafarge Holcim.
Ravi has a Bachelor of Engineering from Delhi Technological University (formerly Delhi College of Engineering, University of Delhi) in Electronics and Communications. Ravi also has a Master of Business Administration (MBA) degree from the Rotman School of Management, University of Toronto. Recently, he earned a graduate-level certification in Financial Technology from Seneca College. Ravi is currently pursuing his Master of Science degree in Data Science at the University of Rochester and gaining skills in computational methods to offer innovative solutions to both engineering and management problems.
Ravi has a passion for business, technology, and research. His research interests include Digital Transformation, Data Science for Business, Strategy & Innovation, Artificial Intelligence, Technology Management.

J Mater Sci: Mater Electron

Improvement in photovoltaic response of bismuth ferrite by tuning its ferroelectric and bandgap properties

Arti[1,2], Reema Gupta[2], Renuka Bokolia[3], and Vivek Verma[2,*]

[1] *Department of Physics and Astrophysics, University of Delhi, Delhi, India*
[2] *Department of Physics, Hindu College, University of Delhi, Delhi, India*
[3] *Department of Applied Physics, Delhi Technological University, Delhi, India*

Received: 6 October 2020
Accepted: 19 November 2020

© Springer Science+Business Media, LLC, part of Springer Nature 2021

ABSTRACT

Pristine and Mn-doped bismuth ferrite samples were prepared by sol–gel method. XRD patterns showed rhombohedral to orthorhombic phase transition with Mn doping in $BiFeO_3$ sample. The surface morphology of the samples was carried out by scanning electron microscope which exhibited a decrease in grain size with Mn doping. Magnetic hysteresis curves exhibit that manganese doping induces modification in magnetic properties of bismuth ferrites. Significant modification was observed in dielectric with Mn doping and maximum value was recorded for $BiFe_{0.8}Mn_{0.20}O_3$ sample. *P–E* hysteresis curves were recorded to evaluate the ferroelectric properties of all samples at room temperature. A significant improvement was observed in ferroelectric properties in doped samples. Bandgap values of $BiFe_{1-x}Mn_xO_3$ in UV–Vis region decreases with Mn doping and minimum for $BiFe_{0.80}Mn_{0.20}O_3$ sample as evaluated from Tauc's plots. *I–V* studies of pristine and Mn-doped samples were carried out with light and without light. Enhanced photocurrent density was observed in Mn-doped BFO samples. Hence, these modifications in ferroelectric and photovoltaic properties of $BiFe_{1-x}Mn_xO_3$ samples make bismuth ferrites a good candidate for photovoltaic applications.

1 Introduction

Multiferroics are the materials which exhibit ferroic (ferroelectric, ferromagnetic and ferroelastic) properties simultaneously. $BiFeO_3$ (BFO) is an important multiferroic material and has applications in wide range of fields like memory storage devices, spintronics and nowadays photovoltaics [1]. BFO has gained attention not only because of its ferroelectric and magnetic properties but also because of its good optical response. BFO has narrow bandgap (~ 2.7 eV) comparative to other ferroelectric materials which lies in ultraviolet–visible (UV–vis) range [2]. It is a great advantage to achieve good

Address correspondence to E-mail: vermavivek.neel@gmail.com; vivekverma@hinducollege.ac.in

https://doi.org/10.1007/s10854-020-04925-z
Published online: 02 January 2021

photovoltaic response because maximum range of the light spectrum can be used during absorption and more of the photons get absorbed due to its bandgap. These incident photons generate electron–hole pairs instantaneously and intrinsic electric field of ferroelectric material is used to separate these photogenerated electrons and holes. Therefore, good ferroelectricity and optical bandgap are essential parameters for ferroelectric-based photovoltaic materials [3]. The bulk ferroelectric photovoltaic (BFPV) effect in ferroelectric materials is mainly due to the built in polarization which is induced due to non-centrosymmetry of the crystal structure [4]. Therefore, in ferroelectric materials the bulk ferroelectric photovoltaic effect can be observed without any p–n junction or Schottky barrier as in silicon-based solar cells. Photovoltaic cells, based on silicon and other semiconductors, have drawback that the bandgap of these materials is within very short range of wavelength of electromagnetic spectrum, due to which most of the light get wasted. However, maximum market is captured by the silicon-based solar cells but further modification in efficiency is restricted due to bandgap limit and high cost, harsh preparation conditions, and serious environmental pollution are also serious issues. However, perovskite solar cells especially ferroelectric-based solar cells have attracted extensive attention from researchers worldwide and have been developed rapidly in recent years due to low cost and tunable bandgap [5]. Bismuth ferrite is one of the suitable materials for ferroelectric-based solar cell due its ferroelectricity and tunable bandgap. However, BFO-based ferroelectric cells exhibits low photocurrent values and power conversion efficiency. Worldwide researchers are looking bismuth ferrites as a promising candidate for ferroelectric-based solar cells and trying to modify its efficiency by numerous techniques [6]. Therefore, our group is also working to overcome the limitations of BFO-based photovoltaic cells. In our previous work, we reported that Pr doping in BFO enhanced its ferroelectric and photovoltaic properties. It was observed that a small amount of Mn doping in BFO changes the leakage current significantly [7]. Ianculescu et al. reported the magnetic and dielectric properties of Mn-doped bismuth ferrites with leakage current studies [8]. Detailed investigations have been done on structural, dielectric, magnetic and optical properties by various authors [9, 10]. Athena et al. reported magnetic and optical properties of Mn-doped $Bi_{0.9}Ho_{0.1}FeO_3$ (BHoFO) nanoparticles [11]. In their work, they have reported structural and enhanced magnetization properties with Mn doping. It is understood that leakage current may help in photovoltaic current up to some extend and this fact motivated us to perform a thorough study of Mn-substituted bismuth ferrites. In this work, pristine and Mn-doped BFO ceramics have been prepared using sol–gel method. A detailed investigation has been made on structural, magnetic, ferroelectric and photovoltaic properties of pristine and Mn-doped bismuth ferrites.

2 Experimental techniques

Pristine and Mn-doped BFO samples were prepared by low temperature sol–gel method as reported in previous work [12]. High-grade chemicals, bismuth nitrate [$Bi(NO_3)_3{\cdot}5H_2O$], iron nitrate [$Fe(NO_3)_3{\cdot}9H_2O$], and manganese acetate [$Mn(CH_3COO)_2{\cdot}4H_2O$], were used as starting materials. First, bismuth nitrate was dissolved in 40 ml of distilled water, and then iron nitrate was added to it while maintaining the magnetic stirring to make a homogeneous solution. Small amount (2–3 ml) of nitric acid was then added drop wise to make the solution homogeneous. After the precursor solution become transparent, oxalic acid as a chelating agent was added in 1:1 molar ratio with metal ions under constant stirring. The as-obtained light brownish solution was kept on hot plate till it became xerogel. On continuous heating, dried resin auto-ignited and released large amounts of gases and became a dark brownish powder, which was then grinded well in agate mortar and pestle followed by annealing at 500 °C for 4 h to get pure phase. Then pellets of different samples were prepared by mixing PVA in powder and applying hydraulic pressure of 10 ton. The pellets were sintered at 500 °C for four hours for further studies. Both sides of pellets were polished with silver paste for electric characterization. The structural analysis of prepared samples was carried out by X-ray diffractometer (CuKα, λ = 1.5406 Å). The dielectric measurements of all samples were performed at room temperature with the help of Precision 6500B Impedance Analyzer. The ferroelectric behavior of samples was recorded using the Automatic *P–E* Loop Tracer (Marine India). The room temperature *M–H* curves were recorded with

vibrating sample magnetometer (Microsense EV-9, Magnetic field: − 10 kOe to + 10 kOe). *U–V* measurements of the powder samples were obtained using U–V spectrometer in absorbance mode for calculation of bandgap. *I–V* characteristics of the pure and doped samples were carried out with and without light to observe photovoltaic response.

3 Results and discussion

3.1 Structural analysis

The XRD patterns of pristine and Mn-doped BFO are shown in Fig. 1. The obtained peak positions of pure $BiFeO_3$ correspond to the crystal planes as reported in JCPDS No. 82-1254. BFO comprises rhombohedral distorted perovskite structure with phase group of R3c.

Impurity phases $^{*}Bi_{25}FeO_{29}/Bi_2O_3$ residing with $BiFeO_3$ peaks are visible in the 2θ range between 25 and 30. These impurity peaks are suppressed with small concentration of Mn in BFO. The incorporation of Mn^{2+} ion in crystal structure can be attributed to the merging of two intense diffraction peaks at 31.74 degree corresponding to lattice planes (104) and (110) and thus peak shift appears at 32.03°. The peaks of BFO corresponding to plane (1 0 4), (1 1 0) and (0 0 6), (2 0 2) shift toward higher angle with increase in the doping concentration of Mn, which may leads in structural transformation. This change in structure of BFO may be due to substitution of Fe^{3+} cations with Mn^{2+} cations. Peaks (104) and (110) are shifted toward higher angles, which may be due to transition taking place from perovskite R3c structure to orthorhombic pnma structure [13]. Substitution of smaller ionic radius cation (Fe^{3+}) with larger ionic radius cation (Mn^{2+}) leads to increase in average B-site cationic radius. This change in the cationic radius can be verified by calculating the value of tolerance factor *t* as given in Table 1. From Table 1 it can be depicted that the value of tolerance factor is increasing with Mn doping due its larger ionic radii which leads to change in octahedral inclination and variations in iron-oxygen bond distance, Iron-oxygen-Iron bond angle. The tolerance factor is used to quantify the structural stability of perovskite compounds. The tolerance factor for $BiFe_{1-x}Mn_xO_3$ can be written as, $t = R_A + R_O/1.44\ (R_B + R_O)$, where R_A, R_B and R_O are the effective ionic radii of Bi^{3+}, Fe^{3+}, and O^{2-} ions, respectively. When the value of tolerance factor is less than one, the compressive strain acts on the Fe–O bonds and hence on Bi–O bonds which induce lattice distortion and leads to the evolution of higher symmetric crystallographic phase such as orthorhombic or tetragonal [14]. These results correspond to the lattice distortion in rhombohedral structure of BFO occurring with the increase in the substitution concentration which leads to rhombohedral to orthorhombic crystal phase transition [15, 16].

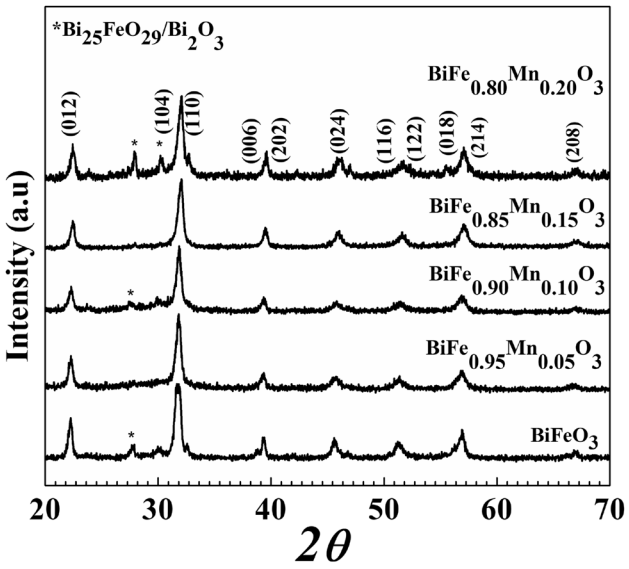


Fig. 1 XRD pattern of pristine and Mn-doped BFO samples

The surface morphology of pristine and Mn-doped BFO samples have been evaluated by scanning electron microscopy. Grain distribution and average grain size of the samples were estimated using image j software. With increase in Mn content in BFO, grain size decreases and results uniformly distributed grains with highly dense morphology (~ 81.30 nm to 48.63 nm) as shown in Fig. 2. The suppression of

Table 1 Structural parameters of pristine and Mn-doped bismuth ferrite samples

Sample name	Tolerance factor (t)	Grain size (SEM) (nm)
$BiFeO_3$	0.8872	81.30
$BiFe_{0.95}Mn_{0.05}O_3$	0.8933	59.45
$BiFe_{0.90}Mn_{0.10}O_3$	0.8995	54.73
$BiFe_{0.85}Mn_{0.15}O_3$	0.9058	50.60
$BiFe_{0.80}Mn0._{20}O_3$	0.9302	48.63

grain growth is may be due to difference in ionic radii of Mn^{2+} and Fe^{3+} and variation in bond strength with oxygen. Kirkendall effect may be the other reason for reduction in grain size due to doping which arise due to diffusion rates of constituting element of the compound [17, 18]. Elemental analysis of the pristine and Mn-doped powder samples were investigated by Energy Dispersive X-ray (EDX) analysis using Scanning Electron microscope. Figure 2 displays a typical EDX spectrum of all the samples sintered at 500 °C. EDX analysis of pristine and Mn-doped samples shows the chemical homogeneity of all the samples through spectra with peaks corresponding to all different elements present in the elements. From the EDX spectrum, it is clearly observed that all the pristine and Mn-doped powdered samples

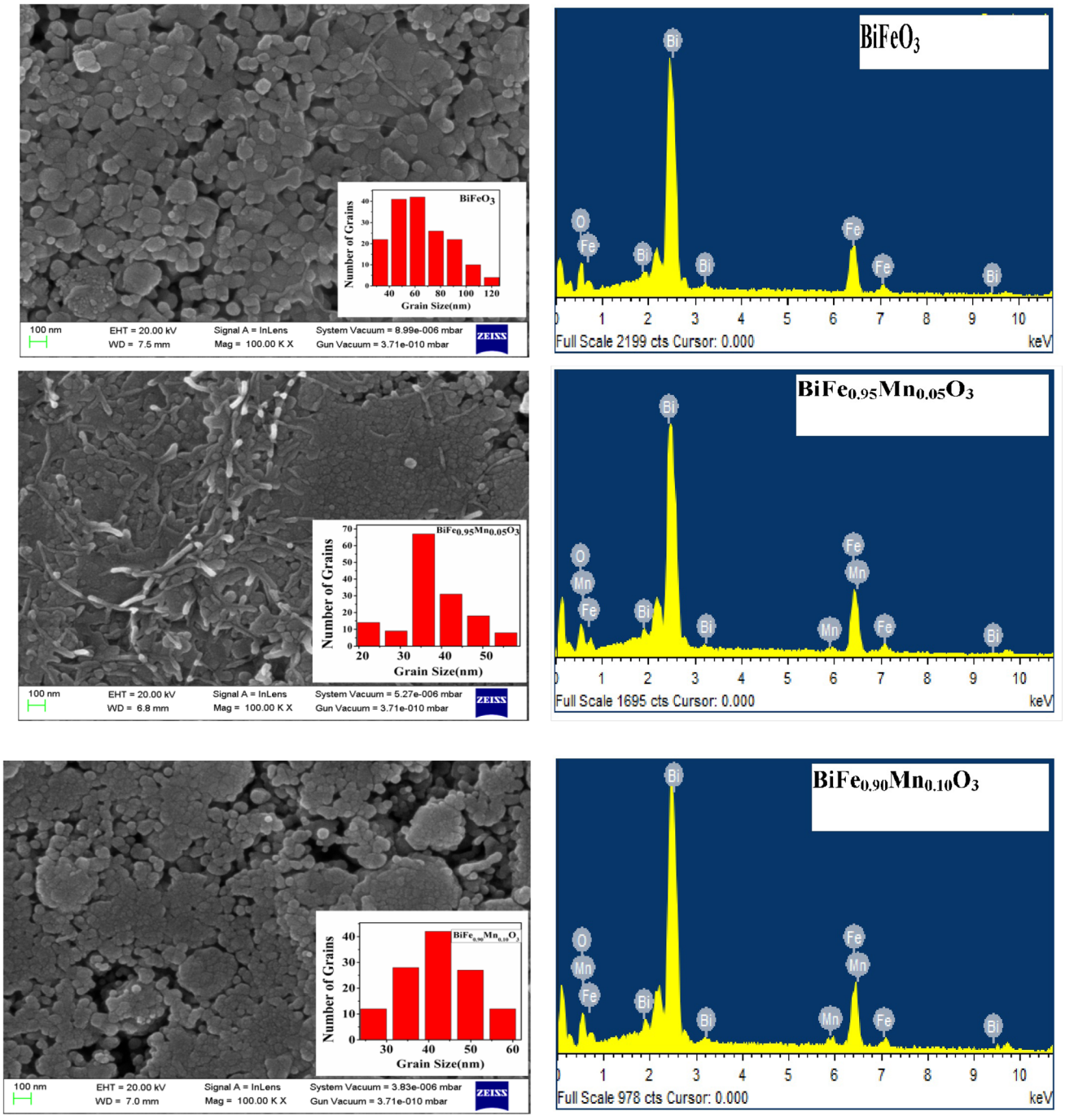


Fig. 2 SEM and EDX images of pristine and Mn-doped BFO samples

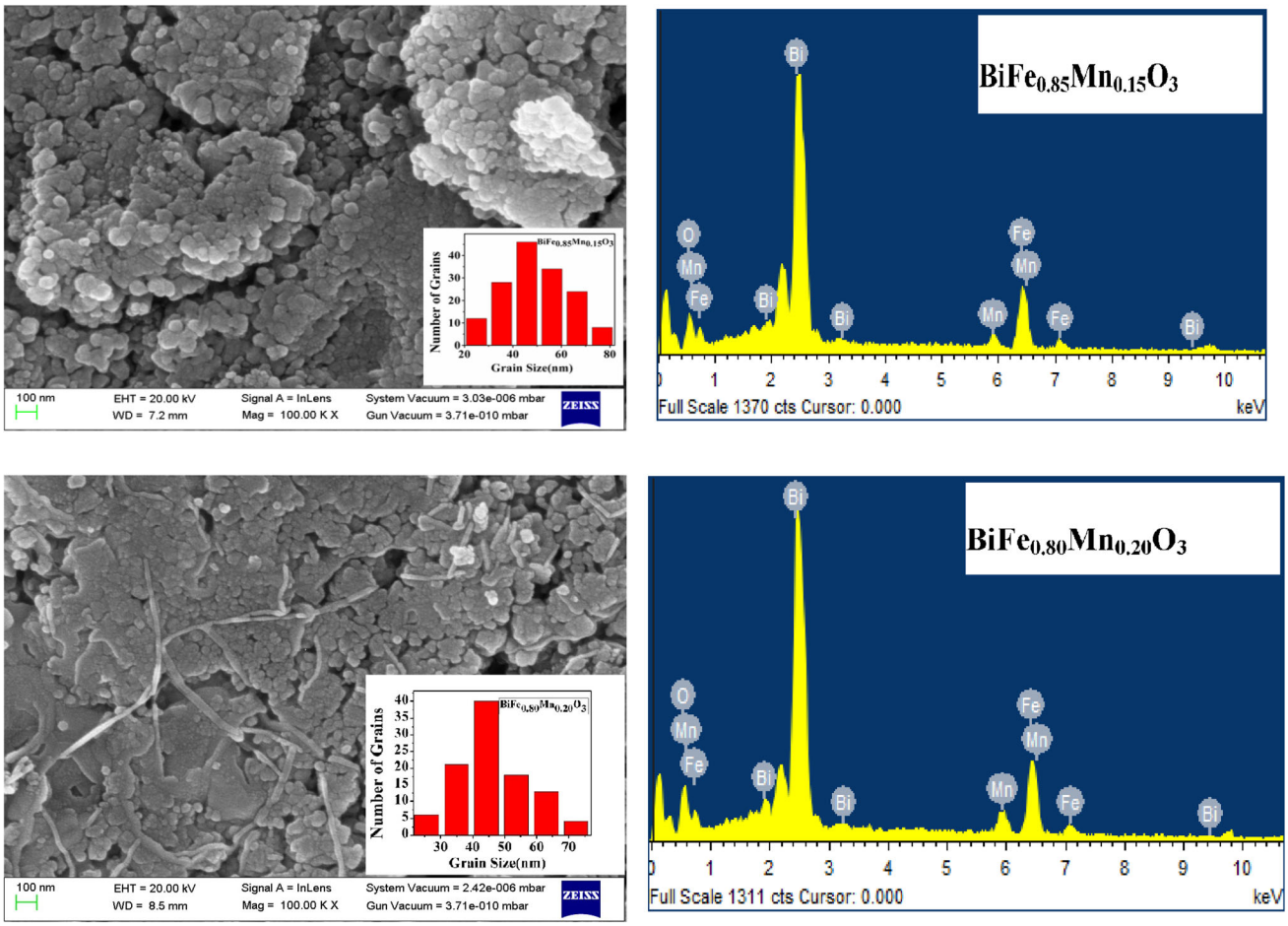


Fig. 2 continued

comprised Bi, Fe, Mn and O elements, this confirms the presence of natural constituents of BFO. The EDX results as shown in Fig. 2 and in Table 2 confirm the presence of essential elements for the successful synthesis of the pristine and Mn-doped BFO samples [19].

3.2 Ferroelectric properties

The room temperature *P–E* hysteresis loops for pristine and Mn-doped BFO samples are shown in Fig. 3. Significant modifications have been observed in ferroelectric properties with Mn doping in BFO.

Table 2 EDX analysis of pristine and Mn-doped bismuth ferrite samples

Element	Sample									
	$BiFeO_3$		$BiFe_{0.95}Mn_{0.05}O_3$		$BiFe_{0.90}Mn_{0.10}O_3$		$BiFe_{0.85}Mn_{0.15}O_3$		$BiFe_{0.85}Mn_{0.15}O_3$	
	Weight%	Atomic%	Weight%	Atomic%	Weight%	Atomic%	Weight%	Atomic%	Weight%	Atomic%
O K	15.90	62.13	14.83	58.13	15.47	60.01	15.20	59.64	17.19	62.00
Fe K	15.51	17.36	18.80	21.16	16.51	18.35	15.44	17.35	15.83	16.36
Bi M	68.58	20.51	65.36	19.61	66.30	17.70	66.78	19.70	62.94	17.38
Mn K	–	–	1.01	1.16	1.72	1.94	2.59	2.95	4.05	4.26
Total	100		100		100		100		100	

Hysteresis loops show the enhanced values of saturation polarization (Ps) and remnant polarization (Pr) but decrease in the value of coercive electric field which could be attributed to effect of grain size and structural change [20]. The hysteresis loop of pristine BFO sample shows a small value of saturation polarization (Ps = 0.380 μC/cm^2) and remnant polarization (Pr = 0.151) with lossy behavior attributed to high leakage current density and low ferroelectricity. Polarization increases with Mn doping in BFO with maximum value of Ps = 5.608 μC/cm^2 and Pr = 3.967 μC/cm^2 for Mn = 0.20 as shown in Table 3. This modification in the value of saturation and remnant polarization of the samples may be due

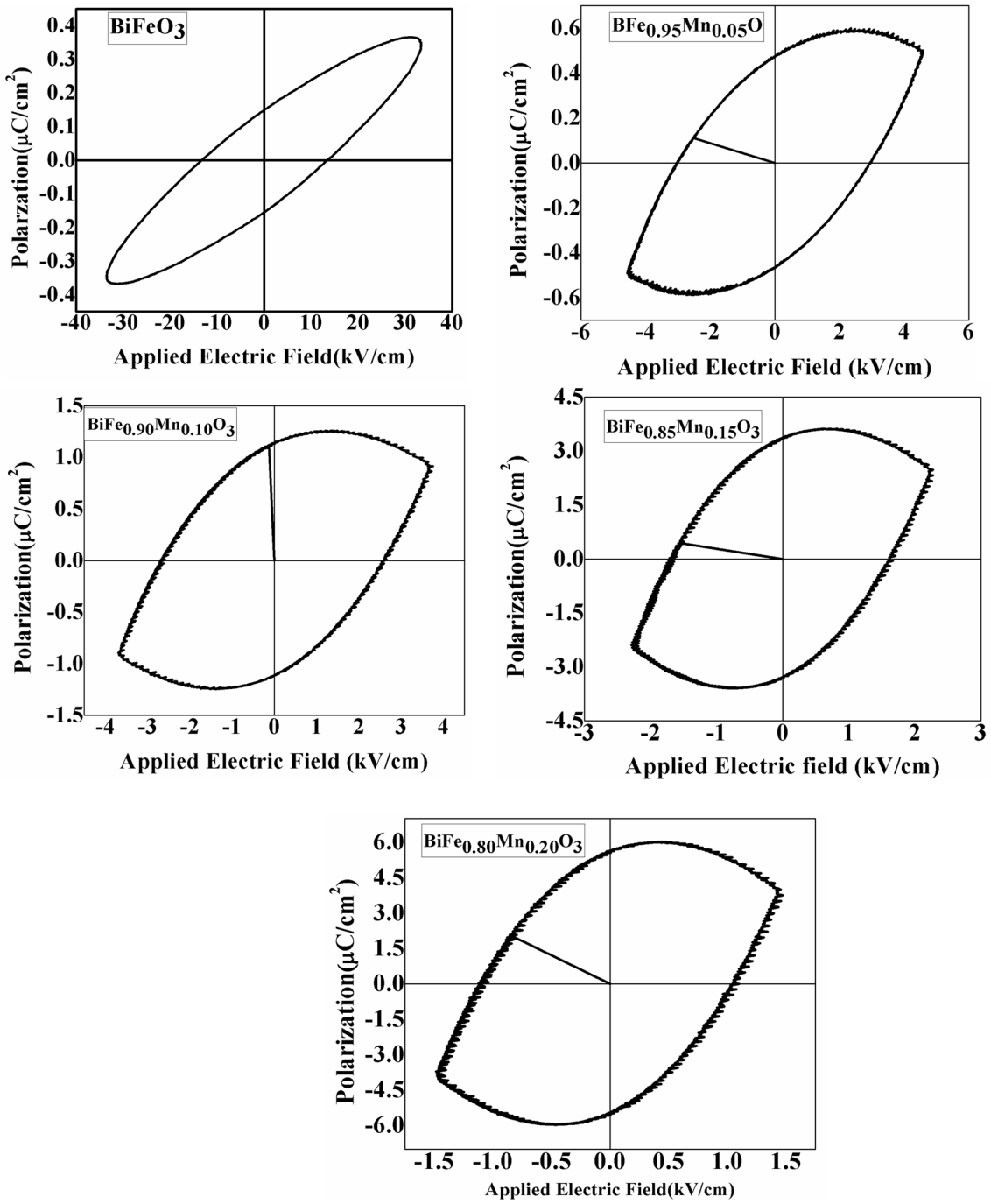


Fig. 3 *P–E* hysteresis loops for pristine and Mn-doped BFO samples at room temperature

Table 3 Magnetic and ferroelectric properties for pristine and Mn-doped BFO samples

Sample name	Saturation polarization (Ps) (μC/cm^2)	Remnant polarization (Pr) (μC/cm^2)	Coercivity (E_c) (kV/cm)	Magnetization (M_s) (emu/g)
$BiFeO_3$	0.356	0.151	13.30	0.158
$BiFe_{0.95}Mn_{0.05}O_3$	0.493	0.478	2.91	0.175
$BiFe_{0.90}Mn_{0.10}O_3$	0.910	1.134	2.59	0.161
$BiFe_{0.85}Mn_{0.15}O_3$	2.469	3.381	1.61	0.165
$BiFe_{0.80}Mn_{0.20}O_3$	3.967	5.608	1.04	0.168

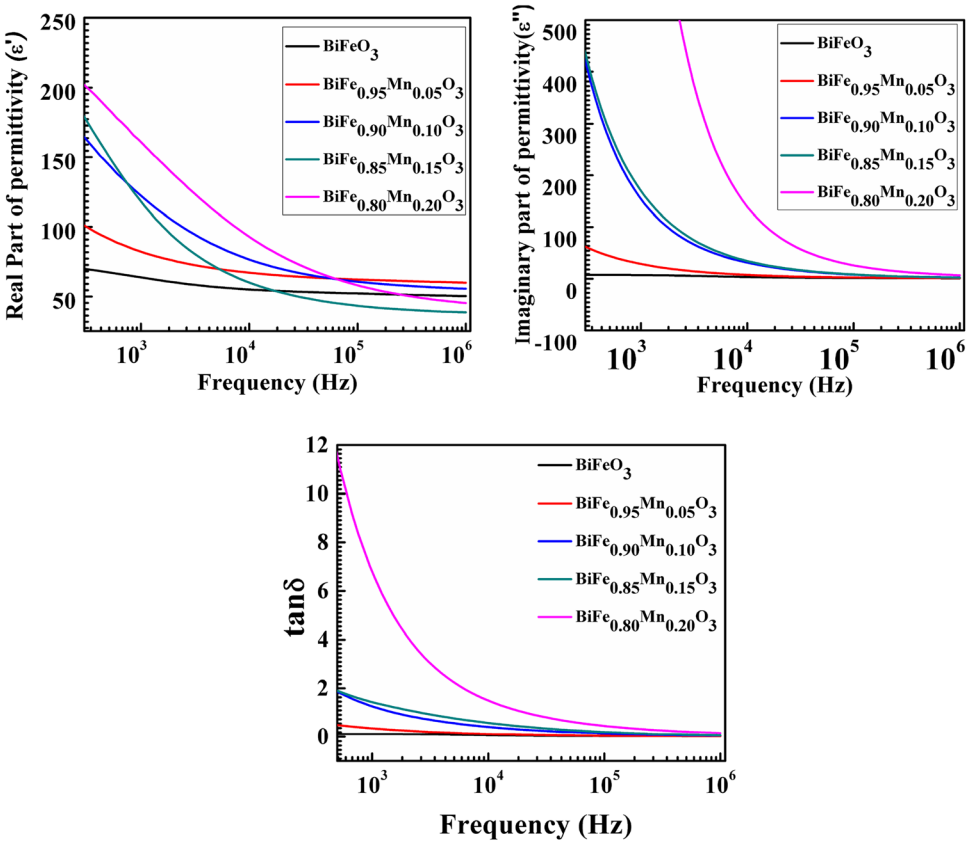


Fig. 4 Dielectric response with frequency for pristine and Mn-doped BFO samples

different reasons. Mainly it is due to the substitution of Mn^{2+} for Fe^{3+} on B sites leads to the lattice distortion of BFO samples which can be verified through XRD analysis. It can also be observed that Mn^{2+} ions with larger ionic radius enter into oxygen octahedron in perovskite structure to cause deformation and enhance the structural transformation from rhombohedral to orthorhombic or tetragonal. The spontaneity of internal electric polarization may be affected due this transformation of structure from one phase to another [21, 22]. This enhanced ferroelectric behavior of bismuth ferrites with Mn doping may be useful for better response in photovoltaic applications.

3.3 Dielectric properties

Frequency-dependent response of dielectric constants (ε' and ε'') and loss tangent (tanδ) for pristine and Mn-doped BFO has been investigated at room temperature in the frequency range of 100 Hz–1 MHz as shown in Fig. 4. Dielectric constants and

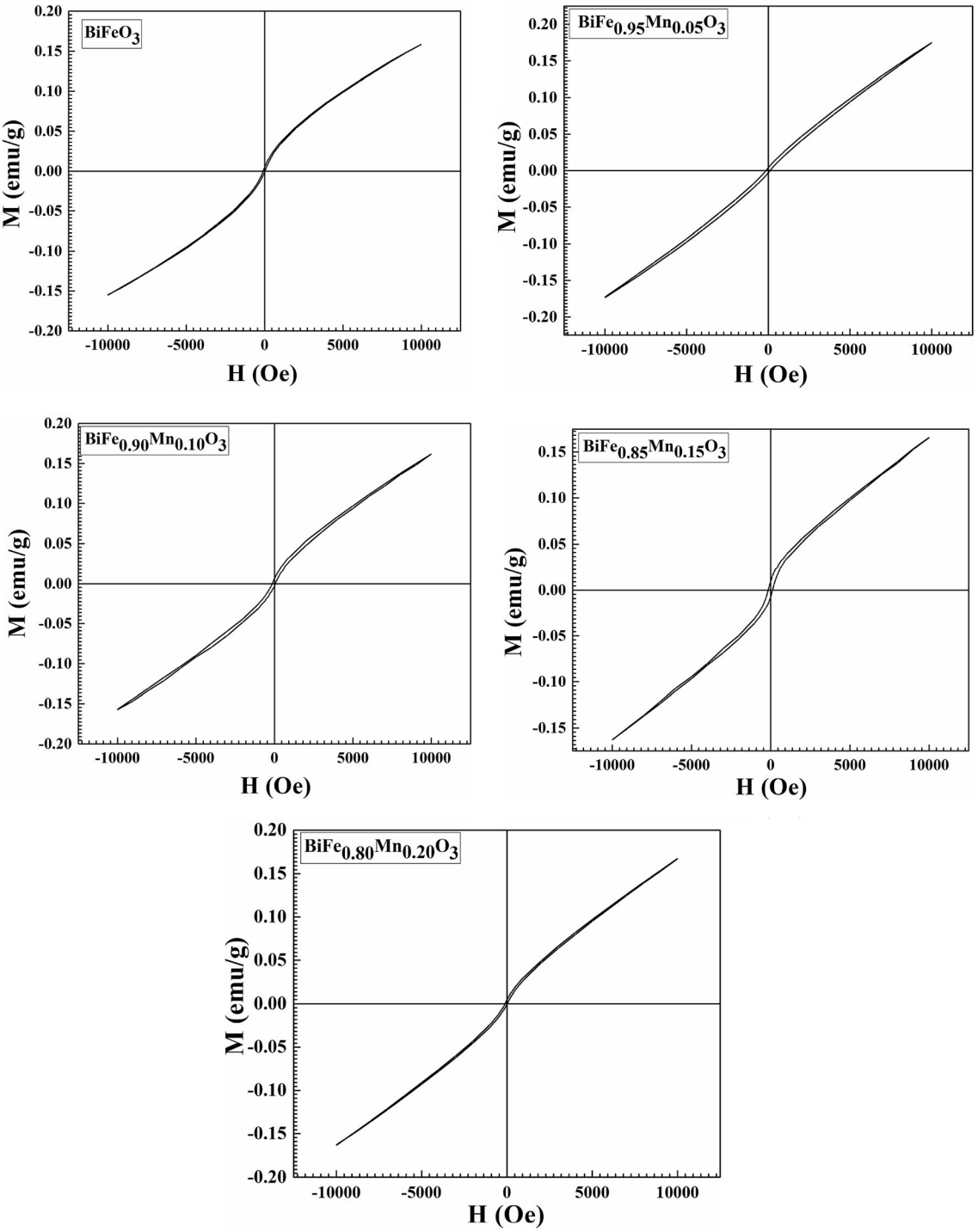


Fig. 5 *M–H* loops for pristine and Mn-doped BFO samples measured at room temperature

loss tangent for all the samples shows decreasing trend with the increase in frequency and becomes independent at higher frequencies. This decrease in dielectric constant and loss tangent values occurs mainly due to space charge contributions toward polarization due to accumulation of charge carriers at grain boundary. Maxwell–Wagner effect explained low-frequency dispersion or interfacial polarization in agreement with Koop's theory [23, 24]. This phenomenon of decrement in values of dielectric constants can also be related to the hopping of ions from Fe^{+2} to Fe^{+3}. Hopping of electron is started from Fe^{+2}

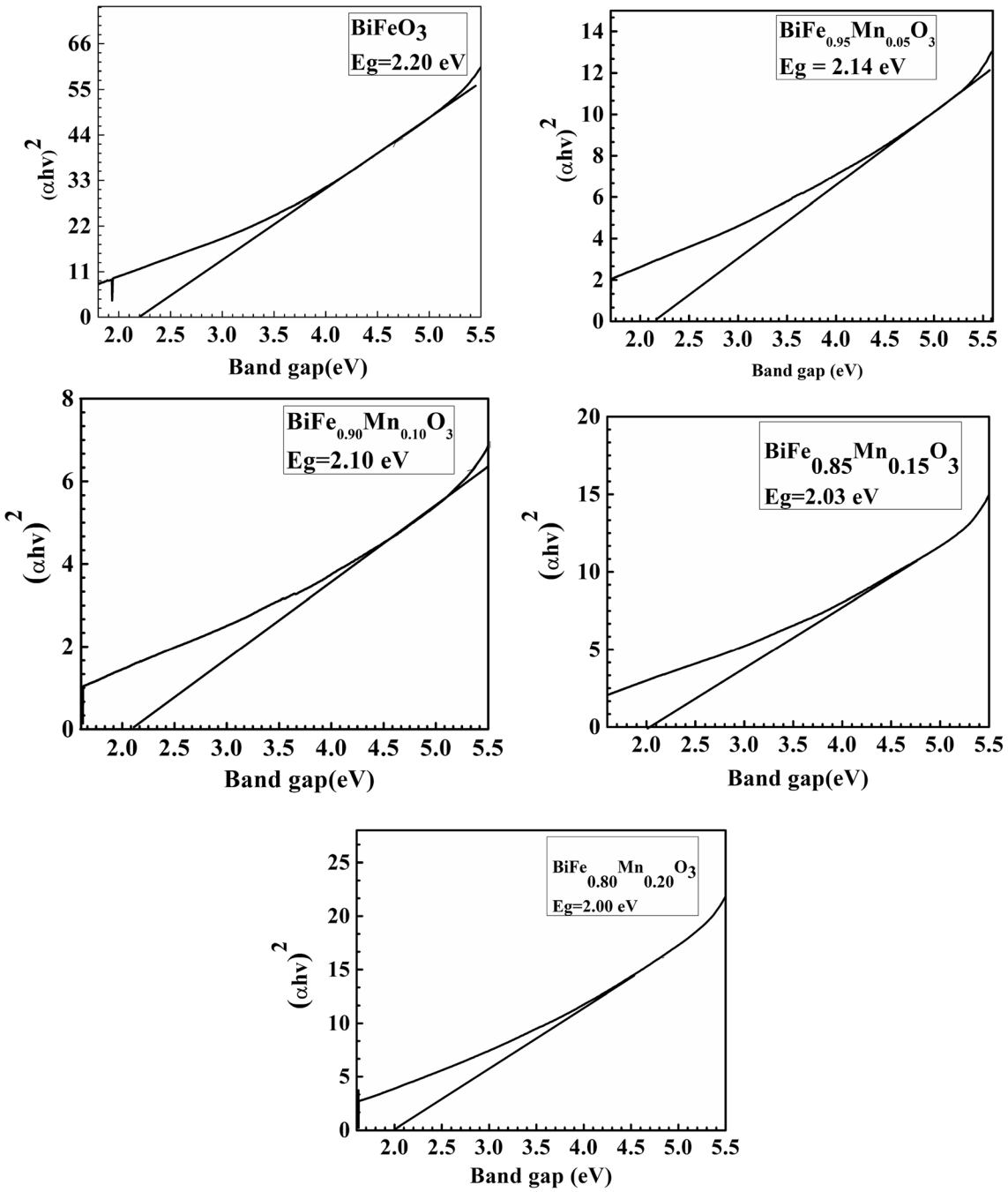


Fig. 6 Bandgap spectra: $(\alpha h\nu)^2$ versus hν response of pristine and Mn-doped BFO samples

to Fe^{+3} ions when these electrons get sufficient energy, which is at higher frequency, since in low-frequency region electrons does not have sufficient energy to hop. Therefore, with the increase in frequency the conductivity of the dielectric materials increases which in results decrease the dielectric constant and loss tangent values [25–27].

3.4 Magnetic properties

M–H hysteresis curve for pristine and Mn-doped BFO samples were recorded at room temperature and shown in Fig. 5. It is observed that all the samples show weak ferromagnetic behavior. Pure BFO exhibits G-type anti-ferromagnetic behavior due to local spin ordering of Fe^{3+} at room temperature [28]. The

helical order with a periodicity of about 62 nm that may be suppressed due decrease in particle size in nanorange results in higher magnetization, which was calculated in earlier reports [29]. BFO with the rhombohedral distorted perovskite exhibits weak ferromagnetic ordering due to some amount of impurity phases. The increased magnetization in BFMO samples may be due to the structural phase transition from rhombohedral to orthorhombic phase resulting in destruction of spiral spin structure [30, 31]. Another reason enhanced magnetization would be the ionic size of Mn^{2+} which is larger than Fe^{3+} which leads to tilt in octahedral and forms canted structure. Therefore the existence of Mn^{2+} ions in BFO affects Fe–O–Fe bond angle and Fe–O length [32]. There is no significant change in saturation magnetization of the samples which may be due to similar magnetic moments of Fe and Mn ions.

3.5 Photovoltaic properties

3.5.1 *Bandgap*

Bismuth ferrite is a potential candidate for photovoltaic properties for solar cell applications due to its relatively narrow bandgap. The study of absorption spectrum is the most direct and simplest method for inquiring the band structure of semiconductors or insulators. The absorption curve rise which comes with the excitation of light in the visible range 400–600 nm can be used to determine the energy bandgap of the semiconductor. The effect of Mn doping on optical properties of BFO has been investigated by absorption optical spectra at room temperature. The bandgap of pristine and Mn-doped BFO was observed in the UV–Vis wavelength range of 400–600 nm. The absorption edge shifts toward the higher wavelength region with the increase in the Mn content in pristine BFO samples. The corresponding energy bandgap of doped samples can be calculated from the tangent line in the graph, according to classical Tauc's relation for direct bandgap $(\alpha h\nu)\alpha(h\nu - E_g)^n$ with $n = 2$, where α is the absorption coefficient, h is the Planck's constant, ν is the light's frequency. The plots of $(\alpha h\nu)^2$ versus hν give energy bandgap plots that are shown in Fig. 6 and the extrapolations of the linear region of these plots to $(\alpha h\nu) = 0$ give the corresponding energy bandgap values of these samples. The energy bandgap of BFO decreases with Mn doping from ~ 2.37 to 2.00 eV as can be depicted from Fig. 6 and Table 4. This narrowing of bandgap can be due to changes in the lattice constants and particle size which may alter the energy band structure of BFO. Sub-bandgap states appear due to lowering of adjacent 3d levels of Fe because of oxygen vacancies present in the samples as reported by Ju et al. First principle studies on optical properties show that on increasing the concentration of oxygen vacancies, the absorption edge gets smoother and shifted toward lower energy levels

Table 4 Bandgap values for pristine and Mn-doped BFO samples

S. no	Sample name	Value of energy bandgap (E_g) (eV)
1	$BiFeO_3$	2.20
2	$BiFe_{0.95}Mn_{0.05}O_3$	2.14
3	$BiFe_{0.90}Mn_{0.10}O_3$	2.10
4	$BiFe_{0.85}Mn_{0.15}O_3$	2.03
5	$BiFe_{0.80}Mn_{0.20}O_3$	2.00

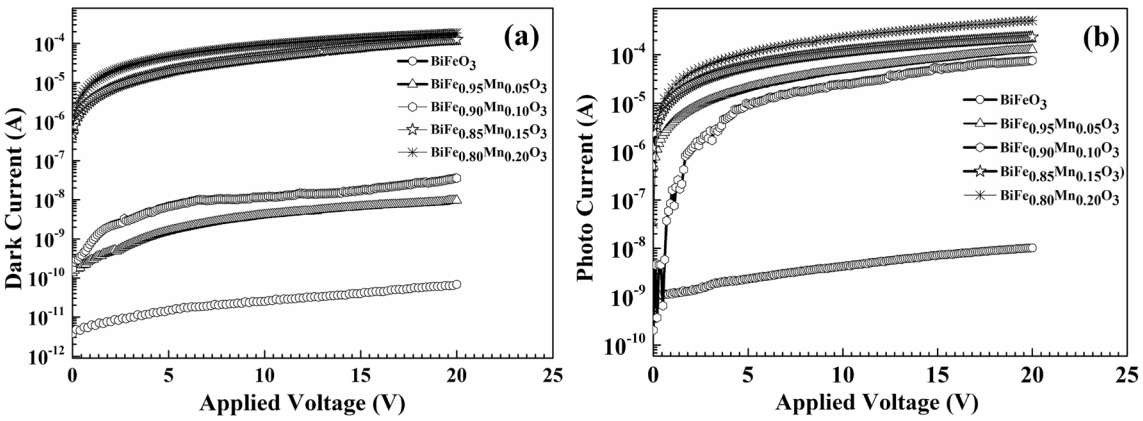


Fig. 7 Photovoltaic response of pure and Mn-doped BFO samples **a** dark current and **b** under illumination (405 nm)

[33]. There are many factors which may cause lowering of bandgap such as decrease in particle size, dislocations, antiphase domain boundaries, and grain surface relaxation. [34–38].

3.5.2 *I–V characteristics*

Figure 7 shows the *I–V* characteristics of pristine and Mn-doped BFO samples. The dark current (without light) measurements have been observed for all the samples at room temperature as shown in Fig. 7a. Dark current for pure BFO is in the range of 10^{-11}–10^{-12} A, and with Mn doping magnitude of dark current increases with Mn doping as can be depicted from Fig. 7a. The results of dark current shown here are well consistent with the foregoing *P–E* loop behaviors. Ferroelectric studies exhibited that lossy nature of samples increases with Mn doping which in turn help in conductivity of the samples. *I–V* characteristics under illumination were studied to investigate the photoconductivity of the samples as shown in Fig. 7a. The laser with maximum power of 5mW of wavelength 405 nm was chosen for photoexcitation-induced current in samples. A significant modification in the magnitude of photocurrent with the illumination of laser light is observed. The photocurrent of pure BFO sample is in the order of 10^{-8}–10^{-9} which is about 10^3 higher order than dark current observed for same sample, although this modification in photocurrent of pure BFO sample is less as compare to our earlier reported work [6] which may be due to some impurity phase present in the sample. These observed modifications in conductivity upon illumination may be due to the result of excitation of charge carriers across a Schottky barrier between ITO and BFO samples.

4 Conclusion

Pure and Mn-doped BFO samples were prepared at low temperature by sol–gel method. Structural transformation from rhombohedral to orthorhombic phase has been observed from XRD spectra. The enhanced electrical and magnetic properties were observed which were in agreement with XRD and SEM data. Pure and Mn-doped BFO samples exhibited the weak ferromagnetic behavior. Dielectric constants and loss tangent for all the samples shows decreasing trend with the increase in frequency and becomes independent at higher frequencies. Decrease in the value of bandgap of BFO with Mn doping has been observed which is in the favor of photovoltaic application. Enhanced ferroelectric properties and modification in photovoltaic current indicate that Mn-doped BFO samples may tune the required properties. Significant decreased value of bandgap and modification in *I–V* characteristics of pure and Mn-doped samples may be useful for optical and photovoltaic applications.

Acknowledgements

Authors of this work are grateful to the Principal, Hindu College, University of Delhi and Head, Department of Physics and Astrophysics, University of Delhi. We acknowledge the University Science Instrumental Centre (USIC) and Prof. Vinay Gupta, University of Delhi for the characterization facilities and valuable suggestions regarding the present work.

References:

1. A. Manzoor, A.M. Afzal, N. Amin, M.I. Arshad, M. Usman, M.N. Rasool, M.F. Khan, Ceram. Int. **42**, 11447–11452 (2016)
2. Z. Peng, J. Ma, J. Wang, Q. Zhao, Y. Wang, Physica Status Solidi **210**, 819 (2013)
3. G. Chena, J. Chena, W. Peia, Y. Lua, Q. Zhanga, Q. Zhangb, Y. Hea, Mater. Res. Bull. **110**, 39 (2019)
4. M.M. Seyfouri, D. Wang, Crit. Rev. Solid State Mater. Sci. (2020). https://doi.org/10.1080/10408436.2019.1708700
5. I. Grinberg, D.V. West, M. Torres, G. Gou, D.M. Stein, L. Wu, G. Chen, E.M. Gallo, A.R. Akbashev, P.K. Davies, J.E. Spanier, A.M. Rappe, Nature **503**, 509 (2013)
6. S. Kumar, P. Kumar, R. Walia, V. Verma, Results Phys. **14**, 102403 (2019)
7. V. Verma, J. Alloys Compd. **641**, 205 (2015)
8. A. Ianculescu, F.P. Gheorghiu, P. Postolache, O. Oprea, L. Mitoseriu, J. Alloys Compd. **504**, 420–426 (2010)
9. A.D. Sharma, H.B. Sharma, Integrated Ferroelectr. **203**, 81–90 (2019)
10. S. Chauhan, M. Kumar, S. Chhoker, S.C. Katyal, H. Singh, M. Jewariya, K.L. Yadav, Solid State Commun. **152**, 525–529 (2012)
11. F. FarlinAthena, T. Saha, M. Islam, M.A. Matin, A. Sharif, Mater. Sci. Eng. **438**, 012030 (2018)

12. A. Bismibanu, M. Alagar, J.S. Mercy Jebaselvi, C. Gayathri, Int. J. Res. Appl. Sci. Eng. Technol. (IJRASET) **06**, 4299 (2018)
13. M. Halder, A.K. Meikap, J. Appl. Phys. **127**, 174104 (2020)
14. V. Srinivas, A.T. Raghavender, K. Vijaya Kumar, Phys. Res. Int. **4835328**, 1–5 (2016)
15. V. Raghavendra Reddy, A. Gupta, D.M. Phase, N. Lakshmi, S.K. Deshpande, A.M. Awasthi, J. Phys. **19**, 136202 (2007)
16. A.T. Raghavender, N.H. Hong, J. Magn. **16**, 19–22 (2011)
17. P.C. Sati, M. Arora, S. Chauhan, S. Chhoker, M. Kumar, J. Appl. Phys. **112**, 094102 (2012)
18. B. Dhanalakshmi, K. Pratap, B. Parvatheeswara Rao, P.S.V. SubbaRao, J. Alloys. Compd. **676**, 193 (2016).
19. Nisha, D.K. Shukla, S. Kumar, AIP Conf. Proc. **2006**, 030037 (2018).
20. J.-Z. Huang, Y. Shen, M. Li, C.-W. Nan, J. Appl. Phys. **110**, 094106 (2011)
21. G. Arya, A. Kumar, M. Ram, N.S. Negi, Int. J. Adv. Eng. Technol. **5**, 245 (2013)
22. B.K. Vashisth, J.S. Bangruwa, A. Beniwal, S.P. Gairola, A. Kumar, N. Singh, V. Verma, J. Alloys. Compd. **698**, 699 (2017)
23. C. Elissalde, J. Ravej, Mater. Ceram. Int. **39**, 8527 (2013)
24. S. Godara, B. Kumar, Ceram. Int. **41**, 6912 (2015)
25. P. Godara, A. Agarwal, N. Ahlawat, S. Sanghi, R. Dahiya, J. Alloy. Comp. **594**, 175 (2014)
26. C.T. Munos, J.P. Rivera, H. Schmid, Ferroelectrics **55**, 235 (2011)
27. M.A. Khan, T.P. Comyn, A.J. Bell, Appl. Phys. Lett. **92**, 072908 (2008)
28. Y.P. Wang, L. Zhou, M.F. Zhang, X.Y. Chen, J.M. Liu, Z.G. Liu, Appl Phys Lett. **84**, 1731 (2004)
29. J.P. Tae, C. Georgia, J.V. Arthur, M.R. Arnold, S.W. Stanislaus, Nano. Lett. **7**, 766 (2007)
30. M.M. Kumar, A. Srinivas, S.V. Suryanarayana, J. Magn. Magn. Mater. **188**, 203 (1998)
31. T.J. Park, G.C. Papaefthymiou, A.J. Viescas, A.R. Moodenbaugh, S.S. Wong, Nano Lett. **7**, 766 (2007)
32. J.P. Ge, J. Wang, H.X. Zhang, X. Wang, Q. Peng, Y.D. Li, Adv. Funct. Mater. **15**, 303 (2005)
33. S. Ju, T.Y. Cai, Appl Phys Lett **95**, 231906 (2009)
34. S.G.V. Rao, C.N.R. Rao, Appl. Spectrosc. **24**, 436 (1970)
35. H.-M. Xu, H. Wang, J. Shi, Y. Lin, C. Nan, Nanomaterials **6**(11), 215 (2016)
36. A. Singh, Z.R. Khan, P.M. Vilarinho, V. Gupta, R.S. Katiyar, Mater. Res. Bull. **49**, 531 (2014)
37. R.Q. Guo, L. Fang, W. Dong, F.G. Zheng, M.G. Shen, J. Phys. Chem. C **114**, 21390 (2010)
38. F. Gao, X.Y. Chen, K.B. Yin, S. Dong, Z.F. Ren, F. Yuan, T. Yu, Z.G. Zou, J.M. Liu, Adv. Mater. **19**, 2889 (2007)

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

Vegetos
https://doi.org/10.1007/s42535-020-00169-8

RESEARCH ARTICLES

In silico analysis of plumbagin against cyclin-dependent kinases receptor

Arpita Roy[1] · **Karanpreet Singh Bhatia**[2]

Received: 6 September 2020 / Revised: 8 October 2020 / Accepted: 24 October 2020
© Society for Plant Research 2021

Abstract
Cancer is an abnormal growth of cells which could migrate from its place of origin to other parts of the body. It is one of the most serious disease on which widespread research work has been going on. Computer aided drug designing has helped in the identification of potential leads that can be used for the development of a drug. Plumbagin is a naphthoquinone derivative from *Plumbago zeylanica* roots which possess strong anticancer properties. Plumbagin has been shown to produce inhibitory effects on multiple cancer-signaling proteins. However, binding mechanism and molecular interactions have not been elucidated yet for most of the target proteins. In this investigation, an attempt was done to explore the binding mechanism of plumbagin against cyclin-dependent kinases (CDK) receptor using molecular docking. The least binding energy of plumbagin with CDK6 was found to be − 6.18 kcal/mol. The molecular simulation suggests that plumbagin has potential binding affinities with CDK6 and its interactions with CDK6 was quite stable during the whole period of simulation run. It was also found that plumbagin obeys Lipinski's Rule of 5 and has drug likeness proved by ADMET analysis. As plumbagin is a natural compound, it has reduced side effects and these results would be useful for cancer treatment.

Keywords Cancer · Plumbagin · CDK6 · Molecular docking · Dynamics

Introduction

In-silico method is extensively used in pharmacology and chemistry to find out the information at the molecular level. Computational results help in explaining the molecular interactions and suggest probable mechanisms that are involved in the process. Structure-based virtual screening and post-screening analysis methods are an important step in hit identification and development of a lead molecule. Nowadays cancer is one of the major causes of death globally. Changes in lifestyle, and consumption of junk foods have immensely contributed to the burgeoning cancer cases (Jemal et al. 2011). The onset of cancer has been linked with various mutations in signalling proteins, molecular biomarkers and important genes associated with normal cell proliferation. A range of signal transduction pathways are involved in cell growth regulation. Therefore, any minute changes in these pathways might lead to tumor pathogenesis. The most significant reversible mechanism to inhibit the specific protein activity in a signaling pathway is phosphorylation. This elementary process is achieved through different protein kinases activity. In recent years, significant development has been done to understand the cancer progression and treatment. The therapeutic arsenal of tumor treatment includes radiation therapy, chemotherapy, surgery, immunotherapy and hormonal therapy and their possible combinations as well. However, these treatment methods possess harmful side effects, therefore there is a demand for finding more influential and lesser toxic anticancer agents.

Medicinal plants are widely used for the treatment of various diseases across the world since ages in Ayurveda and indigenous cultures of different countries, therefore research groups often screen a library of medicinal plant compounds for the development of new drugs. They contain a wide range of phytocompounds which include alkaloids, flavonoids, sterols, glycosides, etc. Medicinal plants can suppress proliferation of cancerous cells due to the presence of various cancers inhibiting compounds like taxols, phytosterols, naphthoquinones, etc. Over 3000 plants have been

✉ Arpita Roy
arbt2014@gmail.com

[1] Department of Biotechnology, School of Engineering & Technology, Sharda University, Greater Noida, India

[2] Department of Biotechnology, Delhi Technological University, Delhi, India

Published online: 11 January 2021

recognized which showed anticancer properties (Roy et al. 2017). Additionally, significance of plant-derived drugs for cancer treatment is growing at a rate of 10–40% (Molassiotis et al. 2006). Approximately 60% of the medications that are available for cancer are attributed to plant sources (Roy et al. 2018). Utilization of plant secondary metabolites having anticancer properties is a preferred choice for cancer therapy (Grover et al. 2012).

Plumbago zeylanica is a medicinal plant, also known as "Chitrak", belongs to the family '*Plumbaginaceae*'. Roots of this plant are commonly used for the treatment of piles, anasarca, diarrhea, cancer, rheumatism and leprosy (Roy and Bharadvaja 2017a, b). This plant possesses antibacterial, anti-fungal, anti-cancer, anti-plasmodial, anti-tumour, stimulatory activity, anti-inflammatory, anti-hyperglycemic, central nervous system and hepatoprotective activity (Roy and Bharadvaja 2017a, b, 2018a, b). Medicinal properties of this plant are due to the presence of a compound i.e. plumbagin. Experimental studies suggest that due to its anti-inflammatory and anti-tumor effect, plumbagin may show protection against various cancers such as oral, colon, blood, breast cancer, ovarian cancer, leukemia, prostate cancer, etc. (Jamal et al. 2014; Tripathi et al. 2019). Inhibitory role of plumbagin is due to alterations of various signaling pathways which play a vital role in cancer cell proliferation, survival, invasion, and metastasis (Jamal et al. 2014). The chemopreventive effect of plumbagin on in vivo cancer model was also studied (Sakpakdeejaroen et al. 2019).

Current situation demands more effective therapies for cancer. Various molecular targets that are responsible for cancer have been recognized, one of the potential targets to inhibit various diseases is cyclin-dependent kinases (CDKs) which is a serine/threonine kinases and extensively studied due to their vital roles in cell cycle division, neuronal functions, transcription, differentiation and apoptosis (Li et al. 2015). These kinases get activated only in the presence of a specific cyclin partner. CDK6 has an important function in the regulation of cell cycle and up-regulation of CDK6 is associated with the development of numerous cancers like colon, breast, bladder, pancreatic and oral cancers (Liu et al. 2015; Kawasaki et al. 2016; Li et al. 2017). Expression of CDK6 is more in cancer cells and low in normal cells. These studies showed that CDK6 has a particular oncogenic function in cancer treatment that can be useful for designing potential anti-cancer drugs having low toxicity (Tadesse et al. 2015). CDK6 has a vital role in the proliferation of cells and cancer. Therefore, development of inhibitors for CDK6 is required to treat cancer. In this study we performed in silico analysis comprising molecular docking of plumbagin against uncoupled CDK6 protein, validation of Lipinski's Rule of 5, ADMET analysis and molecular dynamic simulation to analyse the substantial outcome in support of the proposed idea that plumbagin could inhibit CDK6 function in cell cycle and its mechanism of action was also elucidated.

Materials and methods

Preparation of protein and ligand

Three-dimensional structure of CDK6 (PDB: 3NUP) was retrieved from Protein Data Bank (PDB). The structure of plumbagin (ligand) (CID: 10,205) was retrieved from NCBI—PubChem.

Active site prediction

Understanding the reaction of catalytic residues is essential for correct function of an enzyme. Information regarding active site of CDK6 was reported earlier by Zhang et al. (2018) and was used in this study.

Preparation of receptors

To prepare receptors for docking analysis, co-crystallized water molecules, nonpolar hydrogens, lone pairs, small molecules and non-standard residues were omitted, and polar hydrogens and Gasteiger charges were added. Hydrogens were added to reasonably generate protonation states at physiological pH.

Molecular docking

AutoDock 4.2 is a commonly utilized tool for docking studies, it has the efficiency to predict fast and precise ligand to targets binding confirmations. It is a grid-based process which is used for rapid estimation of binding energy of test confirmations. Docking protocol included the processing of ligand by removing coordinates and water molecules from its PDB file. For conformational searching, Lamarckian genetic algorithm was used. Main residues were placed in a grid of 60 Å × 60 Å × 60 Å (x, y, z). Finally, protein–ligand interaction was observed in discovery studios (2019).

Molecular dynamic simulation (MDS)

MDS of the target CDK6 protein with Plumbagin was performed using non-commercial Desmond version 2019.4 (D. E. Shaw Research, New York, 2019; Schrödinger). For the description of water molecules, TIP3P model was used. The neutralization of each complex was done using a salt concentration of 0.15 M. The standard relaxation protocol and other standard settings given in the Desmond module package were used. For production runs of 1 ns, 5 ps time steps with NPgT ensemble at 300 K were deployed for simulating each

system prepared. For each simulation, Root mean square deviation (RMSD), Potential energy, and Root mean square fluctuations (RMSF) values were recorded and observed for analysis of the stability of the complex.

Drug likeliness and bioavailability

Any compound needs to follow Lipinski's rule of five to be formulated as a drug. SWISSADME online tool is used to analyse drug likeness and bioavailability. It provides LogP, mass, hydrogen bond acceptor, hydrogen bond donor and molar refractivity of a compound as well as other pharmacokinetic parameters and bioavailability.

Results and discussion

Molecular docking

CDK6 has a bilobal fold with 326 amino acid residues in which 1–100 amino acid residues comprises the N-terminal lobe which has 5 antiparallel beta sheet strands and the alpha C-helix. C terminal lobe is quite large which contains residues 101–326. A hinge connects the both lobular folds. The location of the ATP binding site is at the lobal interface and the hinge forms one of its edges. A highly conserved DFG-motif and the activation loop (or T loop) comprises residues 163–189 are present in the C-terminal lobe part of the ATP binding site. Glu61, Lys43 and Asp163 are the key catalytic residues which are not properly positioned in inactive CDK6 due to obstruction by T loop. On binding with cyclin partner, Glu61 is positioned near to Lys43 and Asp163 which helps in the proper orientation of ATP for the nucleophilic attack and also Thr177 is exposed after binding with cyclin partner which is phosphorylated by CDK activating kinases (Schulze-Gahmen and Kim 2002). Palbociclib (PD0332991) has been shown to inhibit CDK6 by binding to the ATP binding site. The major interactions observed in case of PD0332991 was hydrogen bond with Val101, Asp102 and Asp163 and the substituents in Palbociclib were blocking the important Phe98 residue as well (Davies et al. 2002).

Molecular docking score of plumbagin with CDK6 was observed to be − 6.18 kcal/mol. For comparison of docking score, inhibitor that is already present in the PDB structure of 3nup was used. 4-(Pyrazol-4-yl)-pyrimidine was the inhibitor and docking score for this complex was -6.48 kcal/mol. So, plumbagin is quite close to the already established inhibitor of CDK6. Plumbagin in the docked complex formed H-bond with Glu-99 and Val101, hydrophobic interactions with Ile19, Lys43, His-100 and Asp163, and showed interactions with other important residues like Phe98, Val27 and others (Fig. 1). Favorable interactions with all these residues clearly demonstrate that Plumbagin occupies the ATP binding site as Palbociclib and moreover its interactions

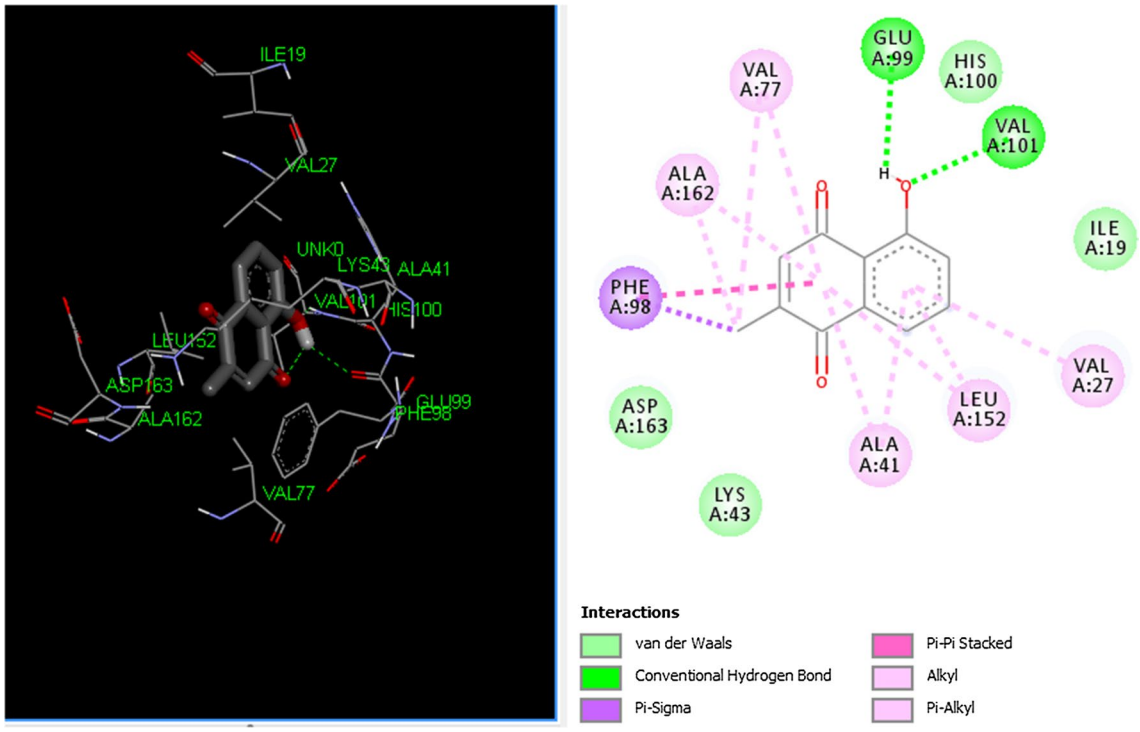


Fig. 1 Docking of CDK6 with Plumbagin

with catalytic residues Lys43 and Asp163 could also help in the inhibition of CDK6 which could be further explored. Hydrophobic interaction with His-100 might give plumbagin a selectivity towards CDK4/6 as it has been reported that interaction with His100 could lead to 26-fold selectivity against CDK 4/6 (Cho et al. 2010).

CDK has a significant role in the regulation of cell cycle and administers cellular transitions from growth phases to linked with DNA replication and mitosis (Bertoli et al. 2013). Due to the differential expression of CDKs, checkpoint control could be affected which would result in neoplastic conversion. A report suggested an enhanced expression of CDK6 in acute lymphoblastic leukemia and its function in pathogenesis (Aleem and Arceci 2015). A study reported unusual proliferation in acute lymphoblastic leukemia cells due to enhanced CDK6 expression and they suggested that CDK6 protein inhibition can serve as a significant therapeutic strategy (Agirre et al. 2009). Cho et al. (2010) reported that ligand (4-(pyrazol-4-yl)-pyrimidines) interact with Lys 43 and Val 101 residues which indicates that these residues are important binding site residues. Jayaraman and Jamil (2014) also found that ligands interact with the Ile 19, Lys 43, Gly 25, Glu 99, Phe 98, Val 101, His 100, Asp 104, Asp 102, Lys 147, Asp 145, Gln 149, Asp 163 residues of CDK6 receptor. Zhang et al. (2018) reported that the binding efficiency of flavonoids against CDK6 were primarily through hydrogen bond and van der waal interactions with Ile19, Val27, Ala41, Asp163, Phe98, Glu61, Gln103, and Leu152 residues. In our docking study also, we found that plumbagin interacts with these residues through H-bonds. This shows its importance as a potential inhibitor of the target molecule.

Molecular dynamic simulation

Throughout the simulation period, prominent hydrogen bonds were formed with residue VAL101 and ASP163 direct as well as through water bridges. Besides this, ILE19, GLN149 and ASN150 also interact through water bridges. Hydrophobic interactions were shown by ILE19, VAL27, ALA41, PHE98, LEU152 and ALA162 (Figs. 2, 3). In our study, VAL101 and ASP163 form hydrogen bonds with plumbagin for most of the simulation time which is similar to the previously conducted study by Hernandez Maganhi et al. (2017) where palbociclib was used as a potential CDK6 inhibitor. The binding residues obtained were mostly similar to our study. The average RMSD value for the plumbagin-CDK6 complex was 1.65 A which was approximately similar to the study done by Hernandez Maganhi et al. (2017). In another study, flavonoids were investigated for their inhibitory effects against CDK6-cyclin D complex (Zhang et al. 2018). They also obtained similar molecular dynamics simulations results where binding residues, interactions and RMSD scores were well in conjunction with our results (Zhang et al. 2018). These major interactions in the binding pocket and the active site of these molecules clearly show the relevance of simulations results. Hydrogen bonding with important residues in the active site of these molecules shows the relevance of results and the potential effect of plumbagin as an inhibitor of these cancer targets.

Drug likeliness and ADMET prediction

Lipinski's Rule of Five is utilized to check the drug likeness that is to assess if a compound shows the desired pharmacological and physiological properties needed to make it orally active drug candidate for humans. According to Lipinski's rule of 5, molecular weight of a drug should be < 500 Da, should have log P (ratio of drug concentration in octanol: water) value < 5, number of H bond donor < 5, number of H bond acceptor < 10 and Molar refractivity should be in between 40–130. SWISSADME analysis of plumbagin reported a molecular weight of 188 Da, log P value of 1.72 and Molar refractivity of 51.07 (Fig. 4). The compound has one hydrogen bond donor and three hydrogen bond acceptors. Therefore, plumbagin follows all the parameters of Lipinski's Rule of 5 and possess drug likeliness.

SWISSADME also predicted pharmacokinetic parameters of plumbagin which includes high GI absorption, blood brain barrier permeability, substrate or inhibitor of CYP1A2, no P-glycoprotein substrate, and not a inhibitor of CYP2C19/CYP2C9/CYP2D6/CYP3A4. As plumbagin is not a substrate of these Cytochrome P450 proteins, its metabolism would be lesser which could be further explored in in vivo *assays.* When the count of rotatable bonds is increased, the compound exhibits more flexibility and is better suitable for efficient ligand protein interaction in a précised binding pocket (Sliwoski et al. 2014). More TPSA (Topological Polar Surface Area) indicates hydrophilic compound and less TPSA represent hydrophobic compound. Ideally, TPSA should be less than or equal to 140 Å (Wipf et al. 2015). Lesser TPSA value is related to more drug distribution and easy permeation of the cell membrane barrier as lower value corresponds to more hydrophobicity. TPSA of plumbagin is 54.37 Å which is in the lower range of PSA values thus its distribution across the cell membrane would be good. All of these observations clearly accomplished that plumbagin can efficiently act as a drug.

Conclusion

Present study showed that inhibition of CDK6 a potential cancer target can be done by a natural compound plumbagin and its mechanism of action has been elucidated. Plumbagin was binding in the ATP binding site and showed interactions

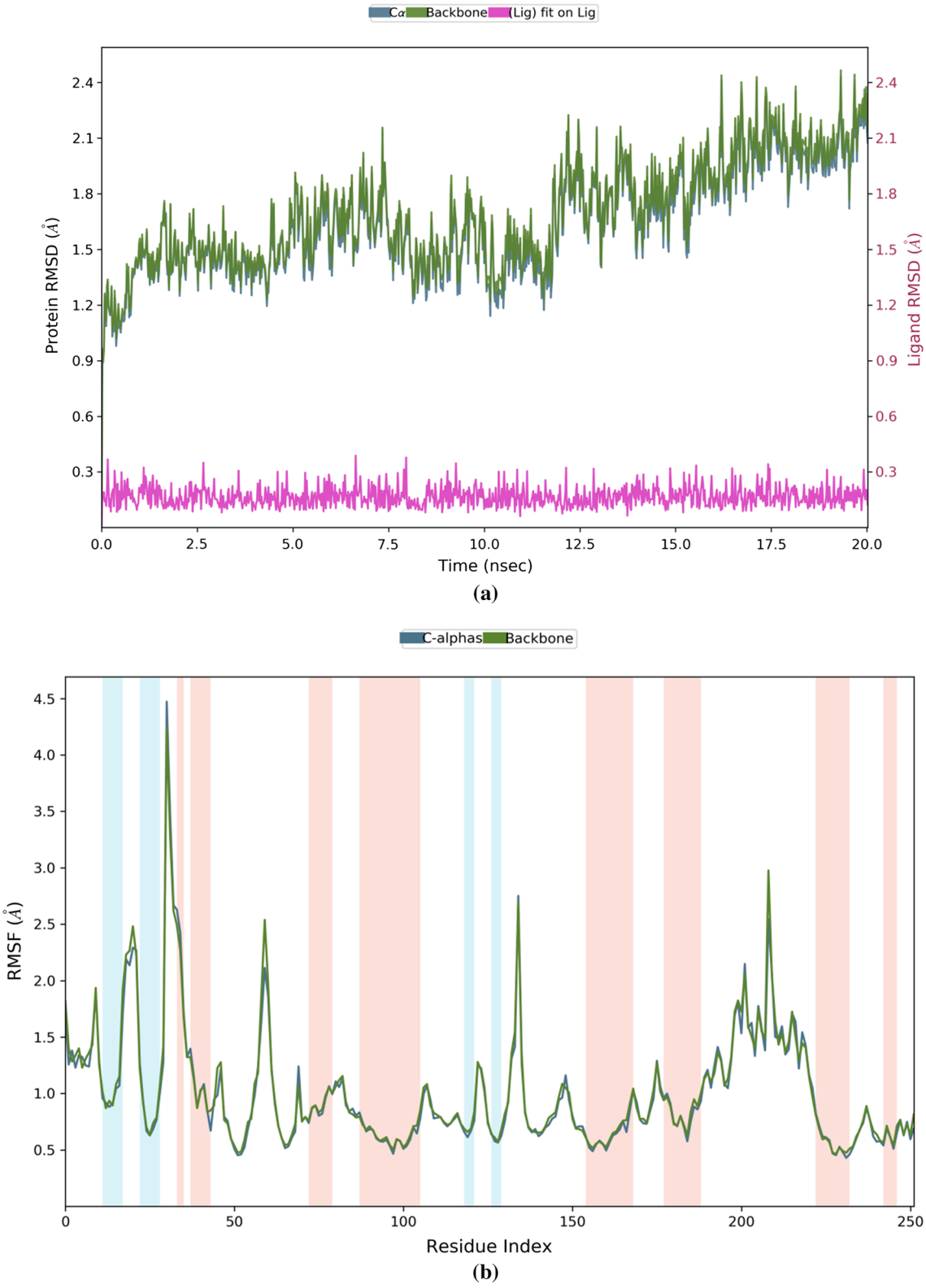


Fig. 2 (**a** RMSD plot for CDK6, **b** RMSF plot for CDK6 with secondary structural features

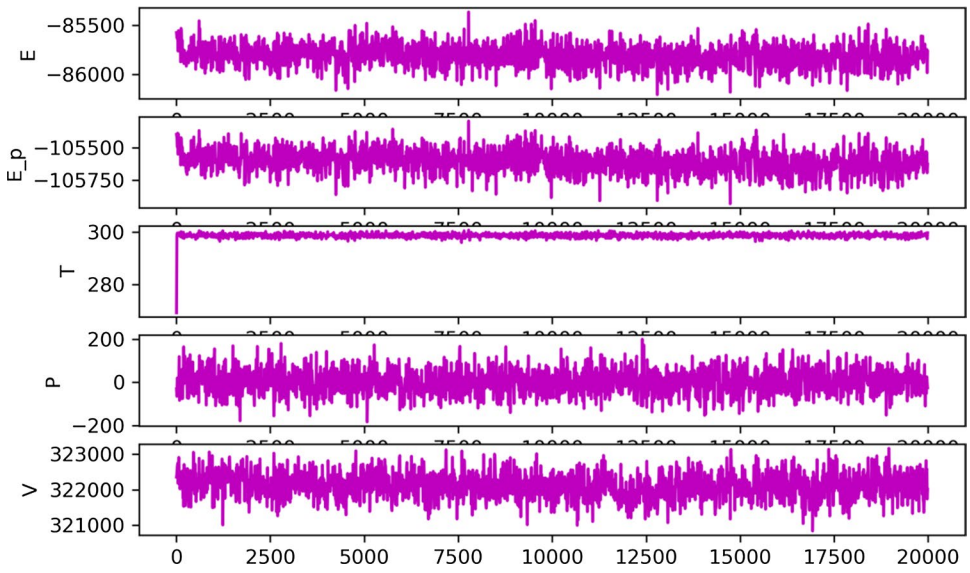


Fig. 3 Simulation quality analysis for CDK6

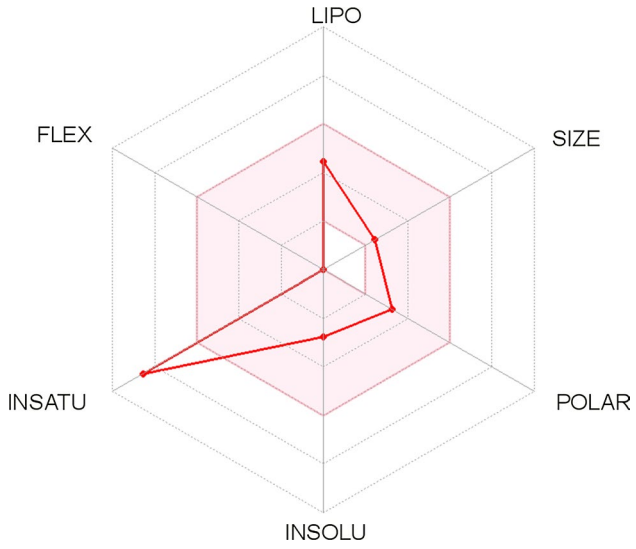


Fig. 4 Radar plot for Plumbagin after SWISSADME analysis

with important catalytic residues of CDK6. MD results confirm the stability of the docked complex and putatively predicted the inhibition of CDK6 active site. Plumbagin also obeys essential parameters to act as drug particularly follows Lipinski's Rule of 5. Present study explains the mechanism of action of plumbagin and it also suggest that it could act as a potential inhibitor of CDK6. Further evaluation of pharmacokinetics and pharmacodynamics of plumbagin can also be done.

Compliance with ethical standards

Conflict of interest Author declare 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

Agirre X, Vilas-Zornoza A, Jiménez-Velasco A, Martin-Subero JI, Cordeu L, Gárate L et al (2009) Epigenetic *silencing of the tumor suppressor microRNA Hsa-miR-124a regulates CDK6 expression and confers a poor prognosis in acute lymphoblastic leukemia. Can Res 69(10):4443–4453

Aleem E, Arceci RJ (2015) Targeting cell cycle regulators in hematologic malignancies. Front Cell Dev Biol 3:16

Bertoli C, Skotheim JM, De Bruin RA (2013) Control of cell cycle transcription during G1 and S phases. Nat Rev Mol Cell Biol 14(8):518–528

Cho YS, Borland M, Brain C, Chen CH, Cheng H, Chopra R, Chung K, Groarke J, He G, Hou Y, Kim S, Kovats S, Lu Y, O'Reilly M, Shen J, Smith T, Trakshel G, Vögtle M, Xu M, Xu M et al (2010) 4-(Pyrazol-4-yl)-pyrimidines as selective inhibitors of cyclin-dependent kinase 4/6. J Med Chem 53(22):7938–7957

Davies TG, Pratt DJ, Endicott JA, Johnson LN, Noble ME (2002) Structure-based design of cyclin-dependent kinase inhibitors. Pharmacol Ther 93(2–3):125–133

Grover A, Singh R, Shandilya A, Priyandoko D, Agrawal V, Bisaria VS et al (2012) Ashwagandha derived withanone targets TPX2-Aurora A complex: computational and experimental evidence to its anticancer activity. PLoS ONE 7(1):e30890

Hernandez Maganhi S, Jensen P, Caracelli I, Zukerman Schpector J, Fröhling S, Friedman R (2017) Palbociclib can overcome mutations in cyclin dependent kinase 6 that break hydrogen bonds between the drug and the protein. Protein Sci 26(4):870–879

Jamal MS, Parveen S, Beg MA, Suhail M, Chaudhary AG, Damanhouri GA et al (2014) Anticancer compound plumbagin and its molecular targets: a structural insight into the inhibitory mechanisms using computational approaches. PLoS ONE 9(2):e87309

Jayaraman A, Jamil K (2014) Drug targets for cell cycle dysregulators in leukemogenesis: in silico docking studies. PLoS ONE 9(1):e86310

Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D (2011) Global cancer statistics. CA Cancer J Clin 61(2):69–90

Kawasaki Y, Komiya M, Matsumura K, Negishi L, Suda S, Okuno M et al (2016) MYU, a target lncRNA for Wnt/c-Myc signaling, mediates induction of CDK6 to promote cell cycle progression. Cell Rep 16(10):2554–2564

Li Y, Zhang J, Gao W, Zhang L, Pan Y, Zhang S, Wang Y (2015) Insights on structural characteristics and ligand binding mechanisms of CDK2. Int J Mol Sci 16(5):9314–9340

Li J, Li J, Cai G, Shen L, Lu F (2017) Proapoptotic and growth-inhibitory effects of plumbagin on human gastric cancer cells via suppression of signal transducer and activator of transcription 3 and protein kinase B. Altern Ther Health Med 23(7):42–48

Liu X, Feng J, Tang L, Liao L, Xu Q, Zhu S (2015) The regulation and function of miR-21-FOXO3a-miR-34b/c signaling in breast cancer. Int J Mol Sci 16(2):3148–3162

Molassiotis A, Panteli V, Patiraki E, Ozden G, Platin N, Madsen E et al (2006) Complementary and alternative medicine use in lung cancer patients in eight European countries. Complement Ther Clin Pract 12(1):34–39

Roy A, Bharadvaja N (2017a) Medicinal plants in the management of cancer: a review. Int J Complement Alt Med 9(2):00291

Roy A, Bharadvaja N (2017b) Establishment of the shoot and callus culture of an important medicinal plant *Plumbago zeylanica*. Adv Plants Agric Res 7(5):00274

Roy A, Bharadvaja N (2018a) Effect of various culture conditions on shoot multiplication and GC–MS analysis of Plumbago zeylanica accessions for plumbagin production. Acta Physiol Plant 40(11):190

Roy A, Bharadvaja N (2018b) Biotechnological approaches for the production of pharmaceutically important compound: plumbagin. Curr Pharm Biotechnol 19(5):372–381

Roy A, Ahuja S, Bharadvaja N (2017) A review on medicinal plants against cancer. J Plant Sci Agric Res 2:008

Roy A, Jauhari N, Bharadvaja N (2018) Medicinal plants as a potential source of chemopreventive agents. In: Akhtar MS, Swamy MK (eds) Anticancer plants: natural products and biotechnological implements. Springer, Singapore, pp 109–139

Sakpakdeejaroen I, Somani S, Laskar P, Mullin M, Dufès C (2019) Transferrin-bearing liposomes entrapping plumbagin for targeted cancer therapy. Journal of Interdisciplinary Nanomedicine 4(2):54–71

Schulze-Gahmen U, Kim SH (2002) Structural basis for CDK6 activation by a virus-encoded cyclin. Nature Structural Biology 9(3):177–181

Sliwoski G, Kothiwale S, Meiler J, Lowe EW (2014) Computational methods in drug discovery. Pharmacol Rev 66(1):334–395

Tadesse S, Yu M, Kumarasiri M, Le BT, Wang S (2015) Targeting CDK6 in cancer: state of the art and new insights. Cell Cycle 14(20):3220–3230

Tripathi SK, Panda M, Biswal BK (2019) Emerging role of plumbagin: cytotoxic potential and pharmaceutical relevance towards cancer therapy. Food Chem Toxicol 125:566–582

Wipf P, Skoda EM, Mann A (2015) Conformational restriction and steric hindrance in medicinal chemistry. In: Wermuth CG, Aldous D, Raboisson P, Rognan D (eds) The practice of medicinal chemistry. Academic Press, Burlington, pp 279–299

Zhang J, Zhang L, Xu Y, Jiang S, Shao Y (2018) Deciphering the binding behavior of flavonoids to the cyclin dependent kinase 6/cyclin D complex. PLoS ONE 13(5):e0196651

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

2020 IEEE International Conference for Innovation in Technology (INOCON)
Bengaluru, India. Nov 6-8, 2020

Increasing Accuracy of Software Defect Prediction using 1-dimensional CNN with SVM

Hitendra Singh Yadav
Department of Computer Science and Engineering,
Delhi Technological University,
Shahbad Daulatpur, Main Bawana Road, Delhi 110042, India
hitendradba@gmail.com

***Abstract*- Precise imperfection expectation computations help to find more deformities with limited effort. To improve the exactness of deformity forecast, in this paper, we propose a significant learning-based strategy for effort careful without a second to save imperfection expectation. The key idea of the proposed approach is that neural system and significant learning could be mishandled to pick important features for deformity expectation since they have been shown amazing at picking accommodating features for plan and backslide. In this paper, we execute a proficient for programming deformity expectation in ventures utilizing 1-Dimentional convolution neural network (CNN) with support vector machine (SVM) and improve exactness in the model.**

Keywords - SVM; artificial neural network; Software defect; detection of bug.

I. INTRODUCTION

Programming imperfection expectation has been an excitement of assessment period considering the way that predicting absconds in starting stages improves programming quality with reduced expense and convincing programming the board. Authorities from different spaces are contributing their undertakings to propose a procedure that satisfactorily and capably help in such way.[1] Diverse artificial intelligence (AI) strategies have been applied to evacuate superfluous what's more, shortcoming information from defect inclined modules and numerous approaches, systems, techniques and models have been proposed utilizing diverse datasets, measurements, and assessment methodologies. [2]

Typically, software has shortcomings/mistakes during development stages and endeavors are expected to diminish these. Early prediction of defects is exceptionally important to diminish the expense. [3] It encourages the professionals to dedicate extra assets and time to the non-defective software modules. In the time of Information Technology, software development is getting intricate step by step what's more, with expanded intricacy, the likelihood of flaws is getting higher. Software is made out of numerous modules and during this procedure modules can be set apart as prone to be defective or non-defective. [4] At the point when a module is anticipated to be defected, more endeavors furthermore, consideration is paid to it. [5] Numerous scientists from various areas are proposing assortment of systems, models and methods for the defect prediction. Scientists are likewise upgrading existing methods and models for better prediction. [6] In spite of the considerable number of endeavors being performed, software development process (SDP) research region despite everything has numerous ambiguities. [7]

To foresee the Software abandons, scientists can likewise apply an AI approach on ongoing Software frameworks. Models incorporate telecontrol/telepresence, apply autonomy and mission arranging systems.[8]-[10] There are number of methods That are utilized like Statistical models and AI models for all the informational indexes. But there is no such a method, that can gives perfect outcome. [11]-[13]

In this paper, we propose a 1D CNN based support vector machine (SVM) based method to anticipate software defects. The problem statement is to develop a CNN based architecture improve the accuracy of defect finding by minimizing the original data loss during the process, this accuracy to be measured in term of F-measure(F1-score), for a set of JAVA based open source projects. The proposed procedure has been actualized into a straightforward yet successful graphical user interface (UI) to let the client handily input the necessary boundaries and foresee the quantity of software defects that may emerge later on. This recognition prompts an advantageous as we proceeds forward in the software development cycle, where the client can roll out fundamental improvements in the software or potentially the technique being applied with the goal that the flaws can be kept away from at later stages.

II. RELATED WORK

```
1. static void myFunc (Queue myQueue) {
2.     int i;
3.     for (i = 0; i < 10; i++) {
4.         // insert i to the tail of the queue
5.         myQueue.add(i);
6.         myQueue.remove();
7.         // remove the head of the queue
8.     }
9. }
```

File1.java

```
1. static void myFunc (Queue myQueue) {
2.     int i;
3.     for (i = 0; i < 10; i++) {
4.         // remove the head of the queue
5.         myQueue.remove();
6.         myQueue.add(i);
7.         // insert i to the tail of the queue
8.     }
9. }
```

File2.java

Fig. 1. Java Example for Defect Detection [1]

In figure 1, a java based example is shown in which a sample function is taken with remove and add functions in queue. It shows how a disparity comes in code when a function is used without declaration.

III. BACKGROUND

Initially for software defect detection various techniques in machine learning are used. To overcome the issues of lower accuracy in software defect prediction CNN based system are implemented, which consists for some number of layers for self-learning and higher accuracy as shown in fig. 2.

978-1-7281-9744-9/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:11:55 UTC from IEEE Xplore. Restrictions apply.

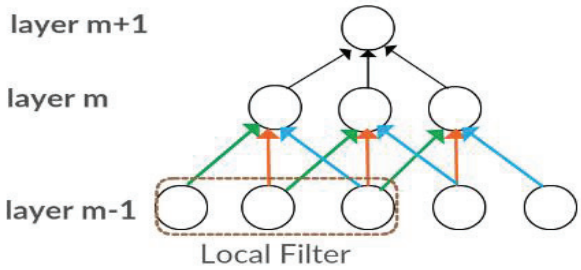


Fig. 2. CNN layers basic model [1]

Reason behind using SVM instead of Logistic Regression:

SVM is used as classifier in proposed solution because SVM attempts to finds the "best" edge (separation between the line and the help vectors) that isolates the classes and this decreases the danger of mistake on the information, while calculated relapse doesn't, rather it can have diverse choice limits with various loads that are close to the ideal point.

IV. METHODOLOGY

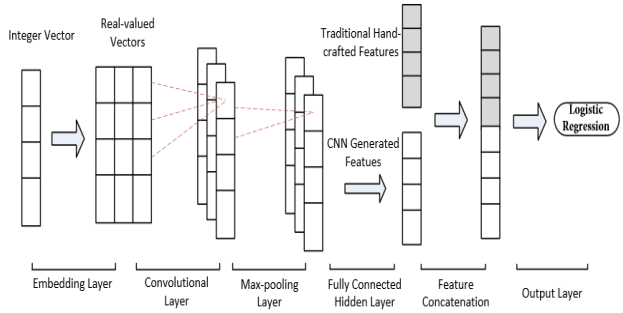


Fig. 3. Li's CNN model[1]

The existing system consists of the CNN process which is shown in fig 3. To which we are going to apply 1D CNN with SVM as per our proposed methodology. In table 1 description of the Dataset is shown. Main dataset is shown in table 1 and table 2. Total 41 .csv files of 12 java-based projects are taken into account for performance evolution.

TABLE I. BASIC PARAMETERS CONSIDERED IN DATASET[1]

Feature Names	Symbols	Description
Weighted Methods per Class	WMC	The number of methods in the class.
Depth of Inheritance Tree	DIT	Indicates the position of the class in the inheritance tree.
Number of Children	NOC	The number of immediate descendants of the class.
Coupling Between Object classes	CBO	The value increases when the methods of one class access services of another.
Response for a Class	RFC	Number of methods invoked in response to a message to the object.
Lack of Cohesion in Methods	LCOM	Number of pairs of methods that do not share a reference to an instance variable.
Lack of Cohesion in Methods, different from LCOM	LCOM3	If m, a are the number of methods, attributes in a class number and $\mu(a)$ is the number of methods accessing an attribute, then $\text{lcom3} = \left(\left(\frac{1}{a}\sum_j^a \mu(a_j)\right) - m\right)/(1-m)$.
Number of Public Methods	NPM	The number of all the methods in a class that are declared as public.
Data Access Metric	DAM	Ratio of the number of private (protected) attributes to the total number of attributes.
Measure of Aggregation	MOA	The number of data declarations (class fields) whose types are user defined classes.
Measure of Function Abstraction	MFA	Number of methods inherited by a class plus number of methods accessible by member methods of the class.
Cohesion among Methods of class	CAM	Summation of number of different types of method parameters in every method divided by a multiplication of number of different method parameter types in whole class and number of methods.
Inheritance Coupling	IC	The number of parent classes to which a given class is coupled.
Coupling Between Methods	CBM	Total number of new/redefined methods to which all the inherited methods are coupled.
Average Method Complexity	AMC	The number of JAVA byte codes.
Afferent couplings	Ca	How many other classes use the specific class.
Efferent couplings	Ce	How many other classes is used by the specific class.
Maximum McCabe	Max (CC)	Maximum McCabe's Cyclomatic Complexity values of methods in the same class.
Average McCabe	Avg (CC)	Average McCabe's Cyclomatic Complexity values of methods in the same class.
Lines of Code	LOC	Measures the volume of code.

TABLE II. SOURCE CODE (PSC) DATASET DESCRIPTION[2]

Project	Version	#Files	#Defects	Buggy Rate (%)
	1.3	124(–1)	20(0)	16.0(+0.1)
	1.4	177(–1)	40(0)	22.6(+0.1)
Ant	1.5	278(–15)	29(–3)	10.4(–0.5)
	1.6	350(–1)	92(0)	26.3(+0.1)
	1.7	741(–4)	166(–1)	22.4(0)
	1.0	339(–0)	13(0)	3.8(0)
Camel	1.2	595(–13)	216(0)	36.3(+0.8)
	1.4	847(–25)	145(0)	17.1(+0.5)
	1.6	934(–31)	188(0)	20.1(+0.6)
	1.1	111(0)	63(0)	56.8(0)
Ivy	1.4	241(0)	16(0)	6.6(0)
	2.0	352(0)	40(0)	11.4(0)
	3.2	260(–12)	90(0)	34.6(+1.5)
	4.0	281(–25)	67(–8)	23.8(–0.7)
JEdit	4.1	266(–46)	67(–12)	25.2(–0.1)
	4.2	355(–12)	48(0)	13.5(+0.4)
	4.3	487(–5)	11(0)	2.3(0)
	1.0	119(–16)	34(0)	28.8(+3.4)
Log4j	1.1	104(–5)	37(0)	35.6(+1.6)
	1.2	194(–11)	186(–3)	95.9(+3.7)
	2.0	186(–9)	91(0)	48.9(+2.3)
Lucene	2.2	234(–13)	143(–1)	61.1(+2.8)
	2.4	330(–10)	203(0)	61.5(+1.8)
Pbeans	1.0	26(0)	20(0)	76.9(0)
	2.0	51(0)	10(0)	19.6(0)
	1.5	235(–2)	141(0)	60.0(+0.5)
Poi	2.0	309(–5)	37(0)	12.0(+0.2)
	2.5	380(–5)	248(0)	65.3(+0.8)
	3.0	438(–4)	529(0)	64.2(+0.6)
	1.0	157(0)	16(0)	10.2(0)
Synapse	1.1	205(–17)	55(–5)	26.8(–0.2)
	1.2	256(0)	86(0)	33.6(0)
	1.4	195(–1)	147(0)	75.4(+0.4)
Velocity	1.5	214(0)	142(0)	66.4(0)
	1.6	229(0)	78(0)	34.1(0)
	2.4	676(–47)	110(0)	16.3(+1.1)
Xalan	2.5	754(–49)	379(–8)	50.3(+2.1)
	2.6	875(–10)	411(0)	47.0(+0.5)
	Initial	162(0)	77(0)	47.5(0)
Xerces	1.2	436(–4)	70(–1)	16.1(–0.1)
	1.3	446(–7)	68(–1)	15.2(0)
Total	-	14,066(–289)	6542(–77)	31.4(+0.6)

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:11:55 UTC from IEEE Xplore. Restrictions apply.

Promise source code (PSC) data set description : Description of the dataset given in table 2 incorporates nearly 14,066 records from 41 sub-versions of 12 open source java based projects, and data incorporates diverse adaptation, quantities of documents, and carriage rate[2]. The source code files are available at conserved Github repositories. The data from source code further processes to get 20 matrix parameters described in table 1 above. The SPSC and PSC dataset are accessible at *https://github.com/penguincwarrior/CNN-WPDP-APPSCI2019.*

V. IMPLEMENTATION

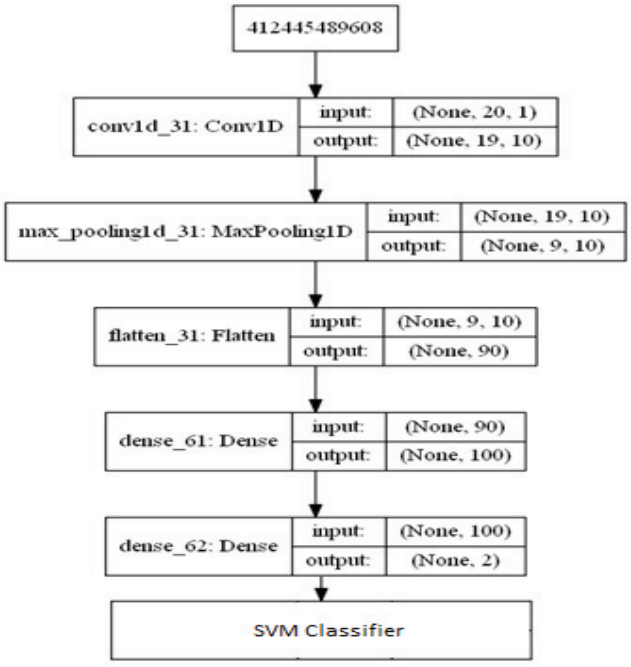


Fig. 4. Proposed 1-D CNN model

We are comparing the 1-D CNN mode with logistic regression somewhat similar to Li's CNN [1] model with the one dimensional CNN model with SVM . The 1-D CNN with SVM is implemented as per fig. 4.

- Implementation framework used: Python 3.1 setup with Keras 2.2.4 and tenserflow 1.14 for building model is used.
- All of our experiments were run on windows 8.1 pro with Intel(R) core(TM) i5-7200u cpu @ 2.30ghz and having RAM of GB.
- ReLU is used as the activation function except in the last layer, which uses SVM for classification.
- The model epoch value is kept at 30 rounds and resut is average of output reacvied by every itration. Number of repetition rounds kept limited due to resorce limitaions.
- 1-D CNN model with SVM clasifier is used for processing the data set to get the desire results.

The proposed model is one dimensional model with having input layer with 20 nodes. The PSC dataset [2] is taken as input, it consists of 20 evolution parameters listed in table 2. Out of input layer given to next layer i.e. convolutional layer. Output of convoltional layer work as input to next building block i.e. max-pooling layer. Further some more model blocks used as shown in figure 4. SVM classifier classifies the final output as buggy or not buggy. The code implementation is as per the algorithm described for the model.

The algorithm of the implementation is shown below:

m No. of samples

n No. of features

0 Label for clear data

1 Label for Defective data

mT No. of training samples

mt No. of testing samples

$X_T^{mT\times n}$ Training samples

$Y_T^{mT\times 1}$ Training labels

$X_t^{mt\times n}$ Testing samples

$Y_t^{mt\times 1}$ Testing labels

$Y_{pred}^{mt\times 1}$ Predicted labels

1 Prepared dataset $X_{m\times n}$ and label $Y_{m\times 1}$ where $Y \in \{0,1\}$

2 Split data $X_{m\times n}$ and label $Y_{n\times 1}$ into 70% training and 30% testing sets

3. $(X_T^{mT\times n}, Y_T^{mT\times 1})$ is the training set and $(X_t^{mt\times n}, Y_t^{mt\times 1})$ is the testing set

4. Set $batch\ size = 64$ *and* $epochs = 30$

5. for $epoch = 1,2,3,\ldots\ldots\ldots\ldots, epochs$

5.1 Set $count = 0$

5.2 Repeat until $count + batch\ size < mT$

5.2.1 Train classifier on $X_T^{count+1\ to\ count+batch\ size}, Y_T^{count+1\ to\ count+batch\ size})$

5.2.2 Calculate accuracy

5.2.3 Set $count = batch\ size$

6. end for

7. Perform testing on $X_t^{mt\times n}$ and find $Y_{pred}^{mt\times 1}$

8. Compare $Y_{pred}^{mt\times 1}$ and $Y_t^{mt\times 1}$ and calculate

8.1 Precision

8.2 Recall

8.3 F1-Score

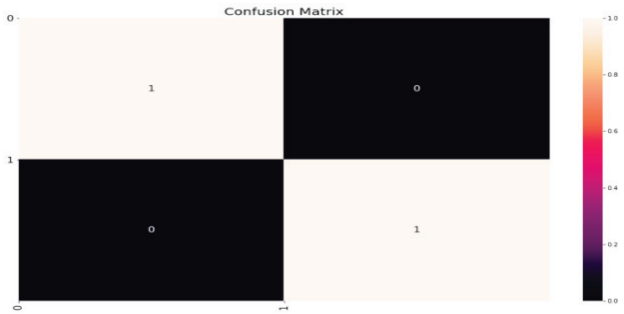


Fig. 5. Confusion Matrix

High review infers that not many positives are misclassified as negatives. High exactness suggests not many negatives are misclassified as positives. There's an exchange

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:11:55 UTC from IEEE Xplore. Restrictions apply.

off here. In the event that model is fractional towards positives, we'll end up with high review however low exactness. It model favors negatives, we'll end up with low review and high exactness.

In table 3 the parameters and its values are mentioned. The parameters are number of filters used, which indicates the CNN filters. Kernel sizes used for filters and batch size is mentioned, the non-linearity function ReLU is also used.

TABLE III. PARAMETER TABLE

Parameter	Value
Number of Filters	10
Kernal Sizes	2
Batch Size	256
Non-linearity function	ReLU
Verbose Size	1
Pool Size	2

TABLE IV. COMPARISON BETWEEN TWO MODELS

	CNN mode with logistic regression somewhat similar to Li's CNN [1] model	1-D CNN model with SVM
Input layer of the model	yes	yes
Convolutional Layer	1	1
Max-Pooling layer	1	1
Activating function used	ReLU	ReLU
Classification model used	Liner regression Function used	SVM classifier used

Table 4 shows the comparison of existing and proposed implementation for software defect prediction.

VI. RESULTS

The evaluation metrics used are precision, Recall and F1 score for better understanding of the proposed model. The values are averages for each individual version of project taken after 30 runs. The higher values are highlighted in bold for the F-1 Score. The formulas for which are given below:

$$\text{F1 Score} = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}$$

$$Precision = \frac{T_p}{T_p + F_p}$$

$$Recall = \frac{T_p}{T_p + T_n}$$

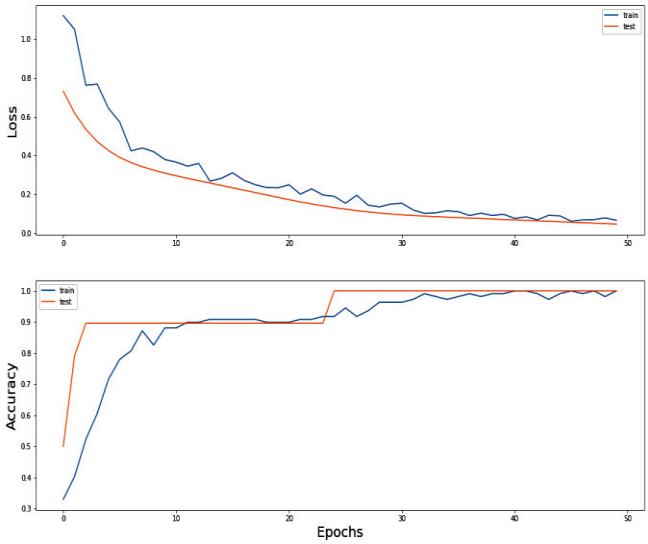


Fig. 6. Accuracy and Loss Graph

As we increase the running instances of model over every project the loss of model is reduced and the accuracy increases and it became flat after certain number of repetitive runs carried out, in our case we can say 30 runs. It can be understood by fig. 6. In this section, results are mentioned in table 5 on basis of F1 score. The better values are mentioned in bold given.

TABLE V. PERFORMANCE MEASURE OF 1-D CNN MODE WITH LOGISTIC REGRESSION AND PROPOSED 1-D CNN WITH SVM ON BASIS OF F-1 SCORE

Project Name	Version	1-D CNN mode with logistic regression	Proposed 1-D CNN with SVM
Lucene	2	0.998	**1**
	2.2	**0.995**	0.933
	2.4	**0.998**	0.895
Synapse	1	0.033	**0.594**
	1.1	0.964	**0.999**
	1.2	**0.998**	0.996
Poi	1.5	0.995	**0.998**
	2	0.033	**0.998**
	2.5	1	1
	3	0.996	**0.999**
JEdit	3.2	0.930	**0.996**
	4	0.631	**0.997**
	4.1	0.630	**0.998**
	4.2	0.033	**0.992**
	4.3	0	**0.999**
Camel	1	0	**0.033**
	1.2	**0.995**	0
	1.4	**0.997**	0.993
	1.6	0.999	**1**
Xerces	Initial	0.997	0.997
	1.2	0.9	**0.994**
	1.3	**0.833**	0.364
Log4j	1	0.8	**0.990**
	1.1	0.964	**1**
	1.2	**0.998**	0.994
Ivy	1.1	0.993	**0.995**
	1.4	0	**0.997**
	2	0	**1**
Xalan	2.4	0.732	**0.997**
	2.5	**1**	0.166
	2.6	**1**	0.997
Ant	1.3	0.066	**1**
	1.4	0.3	**1**

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:11:55 UTC from IEEE Xplore. Restrictions apply.

	1.5	0	**0.966**
	1.6	**0.765**	0.764
	1.7	**0.992**	0
Pbeans	1	0.997	**1**
	2	0.433	**0.997**
Velocity	1.4	0.993	**1**
	1.5	1	1
	1.6	0.993	**0.997**
Average		0.869	**0.988**

Table 6 shows the performance of proposed and existing on the basis of recall value. Also precision values are shown.

TABLE VI. 1-D CNN MODE WITH LOGISTIC REGRESSION AND PROPOSED 1-D CNN WITH SVM MODEL COMPARISON ON BASIS OF CALCULATION OF PRECISION AND RECALL

Project Name	Version	1-D CNN mode with logistic regression		Proposed 1-D CNN with SVM model	
		Precision values	Recall values	Precision values	Recall values
Lucene	2	0.997	1	1	1
	2.2	0.99	1	0.933	0.933
	2.4	0.997	1	0.891	0.9
Synapse	1	0.033	0.033	0.589	0.6
	1.1	0.960	0.966	0.998	1
	1.2	0.997	1	0.992	1
Poi	1.5	0.991	1	0.997	1
	2	0.033	0.033	0.996	1
	2.5	1	1	1	1
	3	0.995	1	0.998	1
JEdit	3.2	0.927	0.933	0.994	1
	4	0.628	0.633	0.995	1
	4.1	0.627	0.633	0.997	1
	4.2	0.033	0.033	0.985	1
	4.3	0	0	0.999	1
Camel	1	0	0	0.033	0.033
	1.2	0.992	1	0	0
	1.4	0.994	1	0.987	1
	1.6	0.998	1	0.999	1
Xerces	Initial	0.994	1	0.996	1
	1.2	0.9	0.9	0.995	1
	1.3	0.833	0.833	0.363	0.366
Log4j	1	0.8	0.8	0.984	1
	1.1	0.963	0.966	1	1
	1.2	0.997	1	0.988	1
Ivy	1.1	0.987	1	0.991	1
	1.4	0	0	0.993	1
	2	0	0	1	1
Xalan	2.4	0.731	0.733	0.995	1
	2.5	0.998	1	0.166	0.166
	2.6	0.998	1	0.994	1
Ant	1.3	0.066	0.066	0.998	1
	1.4	0.3	0.3	1	1
	1.5	0	0	0.945	1
	1.6	0.763	0.766	0.762	0.766
	1.7	0.986	1	0	0
Pbeans	1	0.995	1	1	1
	2	0.433	0.433	0.994	1
Velocity	1.4	0.987	1	1	1
	1.5	1	1	1	1
	1.6	0.986	1	0.994	1
Average		0.705	0.709	0.867	0.872

The results are shown in bar graph form in fig. 7, fig. 8 and fig.9. it improves the existing software defect prediction technique by using SVM classifier.

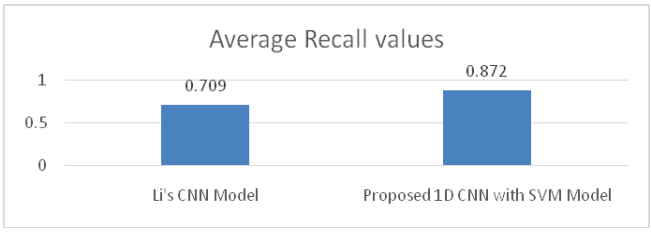


Fig. 7. Recall Comparison Graph

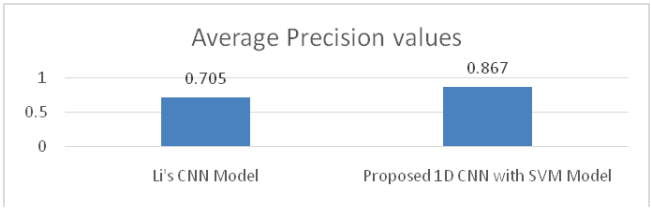


Fig. 8. Precision Comparison Graph

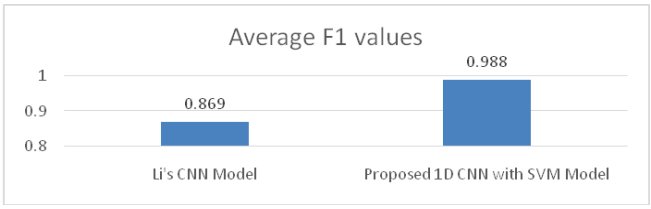


Fig. 9. F1 score Comparison Graph

VII. CONCLUSION

Hence, we proposed 1D CNN based SVM method for software defect prediction and contrasted and various kinds of AI strategies so as to characterize consequently produced crash reports with almost no area information. The outcomes unequivocally demonstrate that a lot of data can be gained from the accident reports utilizing our proposed strategies. Besides, our outcome show that it could be achievable to utilize a comparable sort of approach so as to make a choice help device, which can be utilized to both lessen the time it takes to classify crash reports just as expanding how much such a gathering is made. The accident reports utilized change in both structure and data. All things considered, we endeavored exceptionally broad method of grouping crash reports utilizing a long, text-based methodology. To mimic a progressively reasonable situation, all data that could be utilized so as to gather crash reports coming from the equivalent fundamental issue was utilized while isolating the dataset into a training and approval set. There is high difference improvement in F1 score values, accuracy/precision and recall values. The proposed method is evaluated for some Java based open source projects. Future work may be carry forwarded for the other programming language based projects like C/C++ and others, results for them can vary and may be other complex models can outperform the proposed model on basis of matrix parameter taken or on basis of evolution of projects based on other programming languages.

REFERENCES

[1] J. Li, P. He, J. Zhu and M. R. Lyu, "Software Defect Prediction via Convolutional Neural Network," 2017 IEEE International Conference on Software Quality, Reliability and

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:11:55 UTC from IEEE Xplore. Restrictions apply.

Security (QRS), Prague, 2017, pp. 318-328, doi: 10.1109/QRS.2017.42.

[2] Pan, C.; Lu, M.; Xu, B.; Gao, H. An Improved CNN Model for Within-Project Software Defect Prediction. Appl. Sci. 2019, 9, 2138

[3] S. R. Chidamber and C. F. Kemerer, "A metrics suite for object oriented design," in IEEE Transactions on Software Engineering, vol. 20, no. 6, pp. 476-493, June 1994, doi: 10.1109/32.295895.

[4] Jureczko, Marian & Madeyski, Lech. (2010). Towards identifying software project clusters with regard to defect prediction. Proceedings of the 6th International Conference on Predictive Models in Software Engineering. 1-10.

[5] Menzies, J. Greenwald, and A. Frank, "Data mining static codeattributes to learn defect predictors,"IEEE Transactions on SoftwareEngineering, vol. 33, no. 1, pp. 2–13, 2007

[6] A. N. Lam, A. T. Nguyen, H. A. Nguyen, and T. N. Nguyen, "Combiningdeep learning with information retrieval to localize buggy files forbug reports (n)," inASE'15: Proc. of the International Conference onAutomated Software Engineering, 2015.

[7] R. Jayanthi and L. Florence, "Software defect prediction techniques using metrics based onneural network classifier," Cluster Computing, pp. 1-12, 2018

[8] Koru, A. G., & Liu, H. (2005). Building effective defect-prediction models in practice. Software, IEEE, 22(6), 23-29.

[9] Okutan, A., &Yıldız, O. T. (2014). Software defect prediction using Bayesian networks. Empirical Software Engineering, 19(1), 154-181.

[10] Jureczko, M., &Madeyski, L. (2010, September). Towards identifying software project clusters with regard to defect prediction. In Proceedings of the 6th International Conference on Predictive Models in Software Engineering (p. 9).ACM

[11] X. Huo, M. Li, and Z.-H. Zhou, "Learning unified features fromnatural and programming languages for locating buggy source code,"inProceedings of IJCAI'2016

[12] Menzies, Z. Milton, B. Turhan, B. Cukic, Y. Jiang, and A. Bener,"Defect prediction from static code features: current results, limitations,new approaches,"Automated Software Engineering, vol. 17, no. 4, pp.375–407, 2010

[13] X.-Y. Jing, S. Ying, Z.-W. Zhang, S.-S. Wu, and J. Liu, "Dictionarylearning based software defect prediction," inICSE'14: Proc. of theInternational Conference on Software Engineering, 2014

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:11:55 UTC from IEEE Xplore. Restrictions apply.

Iranian Polymer Journal
https://doi.org/10.1007/s13726-020-00897-3

ORIGINAL RESEARCH

Influence of cross-linkers on the properties of cotton grafted poly (acrylamide-*co*-acrylic acid) hydrogel composite: swelling and drug release kinetics

Preeti Gupta[1] · **Roli Purwar**[1]

Received: 8 October 2020 / Accepted: 23 December 2020
© Iran Polymer and Petrochemical Institute 2021

Abstract
Cross-linker plays a crucial role in monitoring water holding and drug release properties of a hydrogel system, an essential requirement for smart wound dressings. Present study is focused on the influence of cross-linkers poly ethylene glycol (PEG) and *N*,*N*′-methylene bisacrylamide (MBAAm) on the properties of poly (acrylamide-*co*-acrylic acid) hydrogel grafted over the cotton fabric to form composite for medicated dressings. Fourier transform infrared spectroscopy (FTIR) confirms the grafting of hydrogel on the cotton fabric. Uniform hydrogel layer on the cotton surface is obtained under scanning electron microscopy (SEM). Swelling of the composite prepared using PEG follow first-order kinetics at acidic and neutral pH whereas second-order kinetic model at pH 8.5 while that prepared using MBAAm follow second-order kinetic equation at all the pHs studied. The swelling kinetics is also governed by Peppas model at all pHs. Release of gentamicin sulphate from both the composites are studied in phosphate buffers having pH 4.5, 7 and 8.5 at 37 ± 0.1 °C and observed that it is fastest in phosphate buffer having pH 7. On fitting drug release data into Peppas model, first and second-order kinetic equations, it is found that drug release is diffusion controlled and follows Fickian diffusion mechanism in case of the composite prepared by using PEG as cross-linker, whereas it is controlled by diffusion as well as chain relaxation in case of the composite prepared by using MBAAm. Mechanical testing using universal testing machine supports a higher mechanical strength of the hydrogel composite as compared to its film.

Keywords Hydrogel · Poly (acrylamide-*co*-acrylic acid) composite · Cross-linker · Swelling kinetics · Drug release

Introduction

Hydrogels have earned remarkable attention, especially in the past 30 years because of their considerable applications in many areas [1]. Hydrogels can imbibe a large quantity of water owing to their hydrophilic nature but can retain their existence due to three-dimensional network structure [2]. They can be used as absorbents in baby diapers [3], ion-exchange resins [4], biomedical substances like soft contact lenses [5] or carriers for the controlled release of drugs [6] and supports for catalysts [7, 8]. Preparation of any hydrogel involves three integral parts: Monomer, initiator and cross-linker. Their porosity, i.e. capacity to load a liquid is directly related to the amount and nature of cross-linker used. Different cross-linkers like formaldehyde [9] *N*,*N*′-methylene bisacrylamide (MBAAm) [10, 11], poly ethylene glycol diglycidal ether [12], tetraethylene glycol [13] and polyethylene glycol [14] had been used by different researchers.

Polyacrylamide hydrogel, a superabsorbent material, found applications in various fields such as agriculture, drug release, food packaging products, water treatment, ophthalmic, etc. [15]. But this hydrogel lacks hydrolytic stability. On the other hand, poly acrylic acid is a pH-sensitive polymer widely used for applications where pH sensitivity is required. But its gel also exhibits inferior gel strength, dispersion and elastic properties in aqueous media

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s13726-020-00897-3.

✉ Roli Purwar
roli.purwar@dtu.ac.in

[1] Discipline of Polymer Science and Chemical Technology, Department of Applied Chemistry, Delhi Technological University, Shahabad Daulatpur, Bawana Road, Delhi 110042, India

Published online: 23 January 2021

[16]. Presence of pendant –COOH groups and structural similarity with acrylamide makes acrylic acid a potent candidate for copolymerisation with acrylamide and making it pH sensitive. Bajpai et al. prepared a pH respondent hydrogel system using acrylamide, methacrylamide and acrylic acid [17]. Studies showed that copolymer of poly (acrylamide-*co*-acrylic acid) has high water absorption tendency [9]. However, weak mechanical strength of this copolymer limits its use for several applications like wound dressings, especially at load bearing sites such as knee and elbow. Concept of grafting was introduced to raise the mechanical strength of the hydrogel of the copolymer of poly (acrylamide-*co*-acrylic acid) [14]. Cross-linker plays a key role while converting polymer into hydrogel form. An optimum concentration of the cross-linker is required to convert the copolymer into a form having optimum pore size. Below this concentration hydrogel is not formed due to less crosslinking. But above this concentration, excessive cross-linking hinders the swelling tendency of the hydrogel due to reduction in pore volume. Besides quantity, nature of the cross-linker also influences their swelling tendency. A cross-linker which is hydrophilic and has somewhat longer chain may lead to the formation of the pore of optimum size as compared to that which is hydrophobic and has very short (2–4 carbon) chains. A more hydrophilic cross-linker along with longer chain results in more swelling in a shorter period. Therefore, pore size and hence, swelling ability and drug holding capacity completely depend upon the type and nature of cross-linker used.

For making a wound dressing, the main conditions are it must be safe and biocompatible along with controlled drug release properties. To make a smart wound dressing based on hydrogel it becomes necessary to understand what effect cross-linker have on the properties like swelling and drug release of the hydrogel. For developing poly (acrylamide-*co*-acrylic acid) hydrogel, MBAAm is widely and commercially used cross-linker but in some papers use of PEG [18] has also been reported (We did not involve glutaraldehyde as cross-linker for this study due to its highly toxic nature). PEG itself is a biocompatible material and has wide range applications. So in this paper, we studied the effect of these two different cross-linkers having difference in their nature and chain length, on the swelling and drug release properties. Although some previous efforts in this regard have also been observed, they do not give an insight into the difference in the kinetics of drug release which we elaborate in our work.

The novelty of this research stands in providing insight into the difference in the kinetics of swelling and drug release properties of poly (acrylamide-*co*-acrylic acid) hydrogel prepared by using PEG and MBAAm as cross-linker. This will provide a way of replacing MBAAm which is somewhat toxic by a biocompatible cross-linker or its derivatives to develop a safe and biocompatible wound dressing having controlled drug release properties.

In this study, two different type of cross-linkers, namely polyethylene glycol (hydrophilic and long chain) [18] and methylene bisacrylamide (short chain) were used for the development of poly (acrylamide-*co*-acrylic acid) hydrogel composite through graft copolymerisation on cotton fabric. The morphological, structural, swelling, drug release and mechanical properties of the obtained hydrogel composites were compared.

Materials and methods

Materials

Acrylic acid (AAc), methylene bis acrylamide (MBAAm) and ammonium per sulphate (APS) were procured from Central drug house (Delhi, India), acrylamide of reagent grade from Sisco Research Laboratories (Mumbai, India), polyethylene glycol (PEG 6000) from LOBA CHEMI (Mumbai, India), Potassium dihydrogen phosphate, Ortho phthalaldehyde, Methanol, Mercaptoethanol, 2-propanol, Borax, Sodium hydroxide, all from Sigma Aldrich, Gentamicin sulphate from HIMEDIA Laboratories Pvt (India) Ltd., Cotton fabric (139 g/m^2). Distilled water was used to carry out all the experiments.

Method of preparation of hydrogel composites

Hydrogel composite was prepared in the following steps: (i) Active sites' generation—washed, dried and weighed cotton fabric of (3 × 3) cm^2 was immersed in 5% (w/w) ammonium persulphate initiator solution at 50–55 °C for about 30 min to generate active sites over the cotton fabric surface. (ii) Grafting of polymer over the activated fabric—the activated fabric was graft copolymerized using acrylic acid and acrylamide monomers 15% (w/v) in 1:1 for about 55 min keeping the temperature at 45–50 °C. (iii) Insertion of cross-links—after graft copolymerization, PEG or MBAAm (0.1% w/w of monomer) was added to convert the graft copolymer into hydrogel form by inserting cross-links. The reaction was carried out for 15 min. The composite developed using PEG is referred to as sample A and that prepared using MBAAm as sample B in this paper.

At the optimized conditions, i.e. 5% APS, 15% monomer (acrylamide and acrylic acid), 0.1% (w/w) cross-linker concentration, the grafting yield was 102% and 184% for Sample A and B, respectively. Grafting yield is calculated by the same method as used by Purwar et al. [14] using the following formula:

$$\text{Grafting yield} = \frac{w_g - w_d}{w_d} \times 100,$$

where w_g is the weight of grafted material, and w_d is the weight of fabric before grafting.

The thickness of the grafted layer was measured using digital Vernier callipers at five different places and their mean was reported. Sufficiently thick, i.e. 0.23 ± 0.05 mm thick in case of sample A and 0.30 ± 0.04 mm in case of sample B, hydrogel layer was grafted over the cotton fabric. These layers have approximate resemblance with the thickness of the cotton fabric.

Characterization

The following techniques have been utilized to characterize the hydrogel grafted cotton fabric.

Fourier transformed infrared spectroscopy (FTIR)

Thermo-scientific Nicolet 380 spectrophotometer, USA, in ATIR mode was used to analyse the composites obtained. The Scanning was performed setting frequency range at 500–4000 cm^{-1} and resolution at 4 cm^{-1}.

Scanning electron microscopy (SEM)

Morphology of hydrogel composite samples was studied using SEM (model no. S-3700N, Hitachi, Germany). Before mounting in SEM, the samples were made conductor of electricity by gold-sputter coating and then fastened using adhesive tape in the sample stage.

Swelling studies

For swelling studies, hydrogel composites of size (1×1) cm^2 were immersed in the phosphate buffer solutions having pH 4.5, 7 and 8.5. At regular time intervals, the composite was kept out and weighed. The pHs selected for swelling test simulates with wound pH in fresh and curing conditions. Chronic wound has pH in the range of 7.15–8.9 and it becomes neutral and then acidic with the curing of the wound [19, 20]. Before weighing, the surface water was wiped out using tissue paper. The swelling ratio (α) was calculated using the following formula:

$$\text{Percent isothermal swelling}\left(\text{Sw}_t\right) = \frac{\left(W_w - W_i\right)}{W_i} \times 100\%. \tag{2}$$

Here, w_w is the weight of wet hydrogel composite and w_i is the initial weight of dry hydrogel composite.

$$\text{Swelling ratio}(\alpha) = \text{Sw}_t/\text{SD}_{eq}, \tag{3}$$

where SD_{eq} is the Equilibrium swelling degree at a certain temperature and pH.

The experiments were performed in triplicate.

Swelling behaviour of the hydrogel composites can also be explained by their number average molecular mass between crosslinks ($\overline{M}_c$), cross-link density (q) and mesh size (ξ). The following equation was used to predict the number average molecular mass between crosslinks ($\overline{M}_c$) [11, 21]:

$$\frac{1}{\overline{M}_c} = -\frac{\frac{\bar{v}}{V_1}\left[\ln\left(1 - v_{2,s}\right) + v_{2,s} + \chi v_{2,s}^2\right]}{\left(v_{2,s}^{\frac{1}{3}} - \frac{1}{2}v_{2,s}\right)}, \tag{4}$$

where V_1 is the molar volume of water $\simeq$ molar mass of water $= 18.1$ g mol^{-1} [22], χ is the Flory Huggins polymer –water interaction parameter $\simeq 0.5$[22, 23], $\bar{v}$ is the specific volume of the polymer = volume of the polymer/mass, $v_{2,s}$ is the polymer volume fraction in the swollen condition, it was predicted from equilibrium swelling results using the relation,

$$v_{2,s} = \frac{v_d}{v_s}. \tag{5}$$

v_d is the volume of dry polymer, v_s is the volume of the swelled hydrogel at equilibrium condition.

The cross-linking ratio, q, was obtained in terms of mole fraction of cross-linked units [24], i.e.

$$q = \frac{M_o}{\overline{M}_c}, \tag{6}$$

where M_o is the molar mass of the repeating unit; its value was obtained using the following relation [25]:

$$M_o = \frac{n_{AAm} \times M_{AAm} + n_{AAc} \times M_{AAc} + n_{CL} \times M_{CL}}{n_{AAm} + n_{AAc} + n_{CL}}, \tag{7}$$

where n_{AAm}, n_{AAc}, n_{CL} are the number of moles of acrylamide, acrylic acid, and cross-linker, respectively, and M_{AAm}, M_{AAc}, M_{CL} are the molar masses of acrylamide, acrylic acid and cross-linker, respectively.

The mesh size (ξ) was calculated using $\overline{M}_c$ and Flory characteristic ratio of the polymer C_n by the following relation:

$$\xi = v_{2,s}^{-1/3}\sqrt{\frac{2C_n\overline{M}_c}{M_o}}l, \tag{8}$$

where l is the length of C–C bond along the backbone of the polymer = 1.54 Å [22]

In the above equation, value of C_n was used as the weight average of C_n values of poly (acrylamide) and poly (acrylic acid) chains in accordance with their mole ratio used in the hydrogel matrix. The C_n values for poly (acrylamide) chain and poly (acrylic acid) chain are 2.74 and 6.7, respectively [11].

Swelling kinetics The swelling data of hydrogel composite was fitted in different models like Peppas model, Higuchi, zero, first and second-order kinetic equations. Equation having the highest value for correlation coefficient (R^2) was supposed to be the best model showing swelling kinetics.

The mathematical expression for the Peppas model is

$$\frac{M_t}{M_\infty} = kt^n, \tag{9}$$

where M_t/M_∞ is the normalized swelling ratio, n is the swelling exponent exhibiting mode of movement of penetrate, k is the constant for the hydrogel and t is the time taken in swelling.

The mathematical expression for the Higuchi model is

$$\frac{M_t}{M_\infty} = kt^{1}/2, \tag{10}$$

where M_t/M_∞ is the normalized swelling ratio, k is the constant for hydrogel and t is the swelling time.

The integrated equation for zero-order kinetic model is

$$M_t = M_\infty + kt. \tag{11}$$

The integrated equation for first-order kinetic model is

$$\alpha = (1 - Ae^{-kt}), \tag{12}$$

where A is the pre-exponential factor.

The integrated equation for second-order kinetic model is

$$\frac{t}{M_t} = \left(\frac{1}{kM_\infty^2}\right) + \left(\frac{1}{M_\infty}\right)t, \tag{13}$$

where M_∞ is the weight of composite at equilibrium and M_t is the weight of composite at time t.

After performing swelling experiment, the dry weight of the sample was measured to calculate total weight loss in wet conditions in 24 h.

Antimicrobial test

Antimicrobial activity was checked using the zone inhibition method [26]. Here, discs of drug-loaded hydrogel composite were sterilized in a laminar airflow under UV light for 24 h. The sterilized composites were mounted on agar plates cultured with E. *coli* (gram-negative bacteria) or S. *aureus* (gram-positive bacteria) as model bacteria. The plates were kept in an incubator at 37 °C for 24 h. Development of a clear zone below and around the discs on the plate medium recognizes the release of drug and antimicrobial behaviour of the composite. The experiments were performed in triplicate.

Drug release studies

Drug release kinetics was studied using Carry 300 UV-Visible spectrophotometer (Agilent Technologies). Drug-loaded sample was prepared by immersing 50 mg dried composite in 20 mL gentamicin sulphate solution (1 mg mL^{-1}) for 48 h. The composite material was then separated and dried. The amount of drug loaded was calculated by subtracting the final concentration of drug from its initial concentration in the solution.

To check release profile, the dried drug-loaded composites were placed separately in 10 mL phosphate buffers having pH 4.5 (acidic), 7(neutral) and 8.5(basic). At a regular time interval, 1 mL aliquot was taken out with subsequent mixing of the same amount (1 mL) of fresh buffer. The aliquot was derivatized using the method proposed by Pishbin et al. [27]. In this method a reactive solution was prepared by dissolving 130 mg ortho phthalaldehyde in 0.5 mL methanol and then mixing it with 3.8 mL borate buffer(pH 10.4) and 290 µL 2-mercaptoethanol. Final volume of the reactive mixture was made 5 mL using borate buffer. For derivatization, mixture of 1 mL aliquot, 0.4 mL reactive mixture and 1.2 mL 2-propanol was kept in a heating bath at 40 °C for 5 min.

The calibration curve was made using gentamicin sulphate derivative solutions having concentrations in the range 0.0–0.1 mg mL^{-1} at 333 nm. The aliquot derivatives were analysed spectrophotometrically at 333 nm using Cary 300 UV–VIS spectrophotometer. Here the homogeneous distribution of drug molecules in the composite was assumed. The experiments were carried out in triplicate and their mean value was used for analysis.

Drug release kinetics The drug release data were fitted in different models like Peppas, Higuchi, zero, first and second-order equations. Equations of Peppas, first-order, second-order models have been given in swelling studies.

The mathematical formulation for the Higuchi model is

$$M_t = M_\infty + k_H t^{1/2}, \tag{14}$$

where k_H is the Higuchi constant, an indicator of characteristics of a hydrogel system.

The model having the maximum value of correlation coefficient (R^2) was considered the best model for showing the drug release mechanism.

Mechanical testing

Universal Testing Machine (Instron-2700) in tension mode was used to study the mechanical properties of the hydrogel films and hydrogel composite in wet conditions. Dog bone shape (70 × 10 × 0.4 mm) wet sample was clamped between the two arms [28]. The gauze length of the sample was 50 mm. The ends were fixed using tape to avoid the slipperiness. The clamped sample was stretched at a rate of 5 mm min^{-1}. The test results were reported as stress–strain curves. Three samples of each were tested and their mean value is used for comparing mechanical properties.

Results and discussion

Characterisation of hydrogel composites

The cotton fabric got activated when treated with APS solution at 60 °C for 30 min. APS abstract cellulosic hydrogen to generate cellulose macro radicals. These radical initiate graft polymerization reaction in the presence of acrylamide. This leads to the generation of a new free radical which combines with acrylic acid to generate another free radical. In this way, a chain reaction starts. Subsequent monomer addition to the activated chain propagates the grafting reaction onto cotton fabric. The grafting reaction may be terminated by coupling or combination with the initiator, or by disproportionation with cross-linker.

After grafting reaction, the addition of PEG or MBAAm inserts cross-links among poly (acrylamide-*co*-acrylic acid) chains. Polyethylene glycol (PEG) inserts cross-links by forming an ester linkage with some of the –COOH group in between the chains. Contrary to this methylene bisacrylamide (MBAAm) insert an amide group between the polymeric chain by attacking the unsaturated chain and generating new free radicals. FTIR spectra of the cotton, Sample A and Sample B, are shown in Fig. 1. The characteristic bands of cotton at 3300 cm^{-1} (for –OH group) and 1027 cm^{-1} (C–O str.), are shown in Fig. 1a. The characteristic bands of polyacrylic acid, acrylamide and PEG are present in FTIR spectra of Sample A as shown in Fig. 1b. Bands in the range of 3100–3500 cm^{-1} (due to O–H and N–H stretching) confirm the formation of the copolymer of acrylamide and acrylic acid and that in the range of 2990–2700 cm^{-1} are due to asymmetric and symmetric stretching of C–H of acrylate unit. Band for carbonyl, C=O (C=O of ester group) stretching appears at 1732 cm^{-1} and for C = O (C = O of acrylamide group) stretching appears at 1652 cm^{-1}. A band at 1449 cm^{-1} is due to symmetrical stretching of COO [10]. Cross-linking by PEG in the hydrogel is revealed by a characteristic band of C–O–C stretching at 1168 cm^{-1} showing the presence of ester linkage in it. Figure 1c also reveals

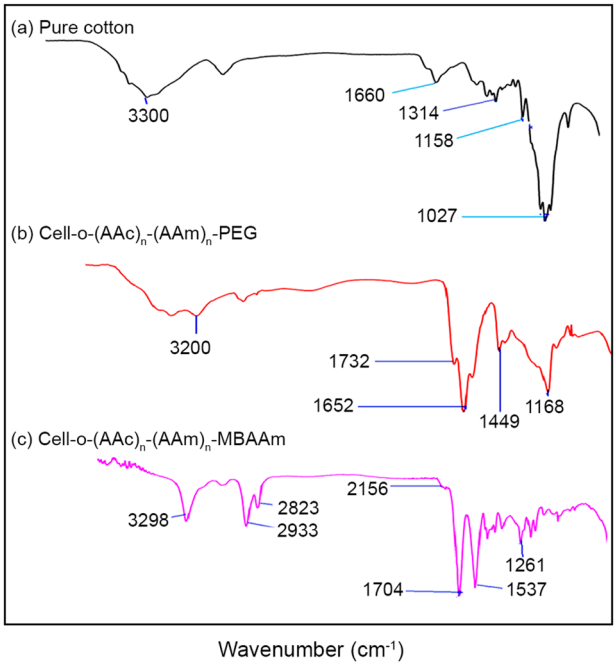


Fig. 1 FTIR spectra of **a** cotton, **b** sample A, **c** sample B

characteristic band in the range of 3100–3500 cm^{-1} (due to O–H and N–H stretching) showing the formation of the copolymer of acrylamide and acrylic acid and that in the range of 2990–2700 cm^{-1} showing asymmetric and symmetric stretching of C–H of acrylate unit [29]. Carbonyl band at 1537 cm^{-1} shows the presence of secondary (2°) amide group. It, indeed, shows N–H bending *trans* to the carbonyl oxygen. A weak band 2156 cm^{-1} is an indication of the presence of C–N group of crosslinking agent MBAAm [10].

Figure 2 shows the SEM images of pure cotton (a), sample A (b) and sample B (c), respectively. There appears no layer over the surface of cotton fabric in Fig. 2a, but uniform layers of hydrogel appear over the surface of cotton fabric in Fig. 2b, c, which is a clear indication of grafting of the hydrogel layer over cotton fabric.

Swelling behaviour

The hydrogel composites are pH responsive as revealed by Fig. 3a–c. There appears a gradual sweep in swelling up to a certain point and then it becomes constant. This constant value is taken as the equilibrium swelling of the composites. Equilibrium swelling is maximum (i.e.1466% and 720%, respectively, for sample A and sample B) at neutral (7) pH and minimum (i.e.736% and 490%, respectively, for sample A and sample B) in acidic medium below the pKa of acrylic acid (below pH 4.7). In alkaline medium, however, the equilibrium swelling is intermediate of these two (i.e.1050% and 690%, respectively, for sample A and sample B). This

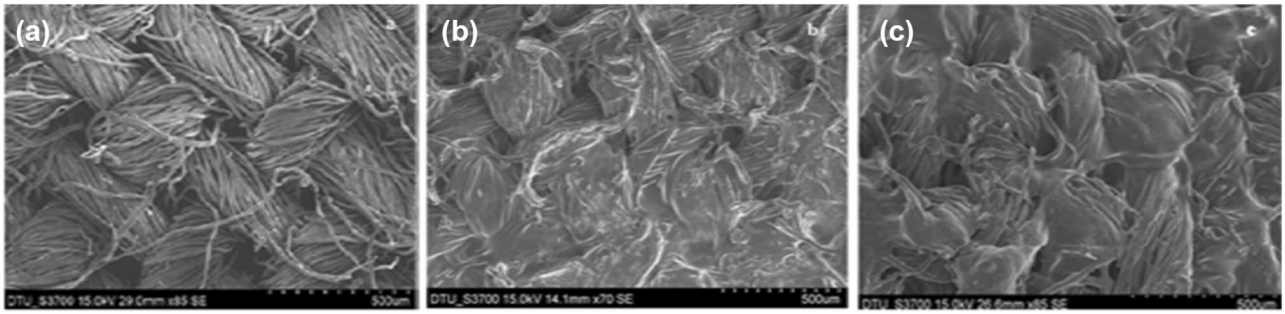


Fig. 2 SEM micrographs of **a** un-grafted cotton fabric, **b** sample A, **c** sample B

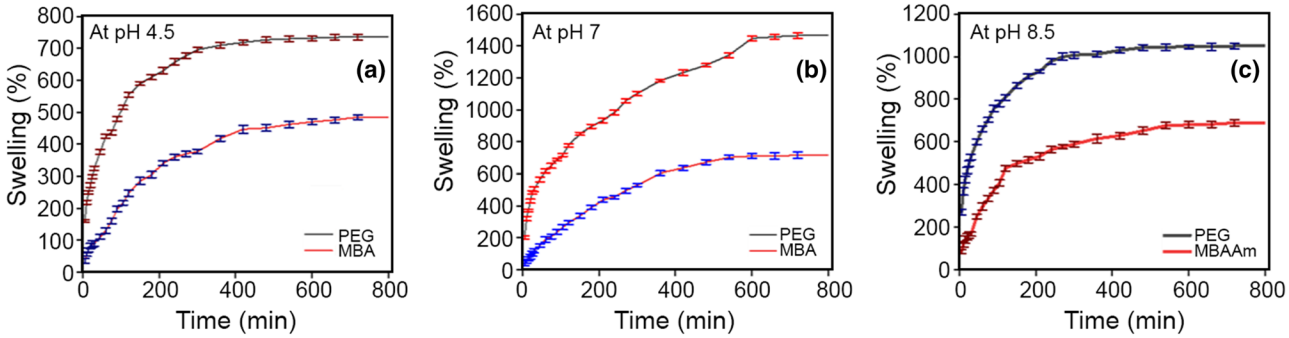


Fig. 3 Swelling behaviour of hydrogel composites at **a** pH 4.5, **b** pH 7, **c** pH 8.5

is because at lower pH (below the pK_a of acrylic acid) the carboxylic groups which were not involved in cross-linking with the hydroxyl group of PEG, remain unionised and form hydrogen bonds with the amine group of acrylamide. Thus, polymer–polymer interaction become more dominant over the polymer–solvent (water) interactions [30]. So the hydrogel shrank and aggregate together in absence of any repulsive forces causing less swelling. With rise in pH, the repulsion generated due to ionisation of carboxylic groups leads a large pore size and hence, results in more swelling. In alkaline media, the structure and physical forces get weak due to dissociation of the carboxylic group leading to smaller pore size and hence somewhat less swelling.

At all the pHs swelling is higher in the case of sample A than sample B. This may be attributed to the difference in the hydrophilicity and chain structure of the cross-linker used. In sample A, hydrophilicity and comparatively longer chain of PEG may lead to larger pores for holding more water. It can also be justified through the network parameter such as polymer volume fraction, average molar mass between cross-links $\overline{M}_c$, cross-link density, q and mesh size (ξ) for sample A and sample B. These parameters are shown in Table 1. This table reveals a higher $\overline{M}_c$ value in the case of sample A showing its more water absorbing tendency as compared to sample B. As the average molar mass between cross-links $\overline{M}_c$ increases, cross-link density decreases, which is also evident from the values obtained by calculations. This table also reveals an increase in mesh size (ξ) with decrease in polymer volume fraction showing a high water content which is seen in Sample A.

Table 1 Network parameter of hydrogel composite sample A and sample B

Network parameter	Hydrogel composite	
	Sample A	Sample B
Polymer volume fraction, $v_{2,s}$	0.28	0.67
Average molecular weight between crosslinks, $\overline{M}_c$	1.293×10^{6}	6.827×10^{4}
Cross-link density, q	4.53×10^{-5}	1.05×10^{-3}
Mesh size (ξ), (in Å)	1120	193

Swelling kinetics

To study swelling kinetics the swelling data were fitted in different models like Peppas model, first and second-order kinetic equations. The results are reported in Fig. S1 (supplementary data) for sample A and sample B. Related kinetic constants are summarised in Table 2. From Table 2 highest R^2 values for first-order kinetics at pHs 4.5 and 7 suggest a

Table 2 Kinetic constants for swelling at different pH for sample A and sample B

pH value		Model name						
		Peppas model			First-order		Second-order	
		R^2	n	k	R^2	k (s^{-1})	R^2	k (L mol^{-1} s^{-1})
4.5	For sample A	0.978	0.324	0.432	0.998	0.0036	0.997	3×10^{-5}
	For sample B	0.987	0.57	0.21	0.94	0.0022	0.97	1×10^{-5}
7	For sample A	0.997	0.327	0.392	0.989	0.0016	0.971	1×10^{-5}
	For sample B	0.997	0.65	0.173	0.62	0.56	0.96	4.2×10^{-6}
8.5	For sample A	0.976	0.35	0.416	0.994	0.004	0.998	2.3×10^{-5}
	For sample B	0.97	0.56	0.294	0.84	0.58	0.99	1.6×10^{-5}

first-order swelling kinetics in acidic and neutral medium. However, at pH 8.5 R^2 value is highest for second-order kinetics showing a second-order swelling kinetics in alkaline medium, which shows a direct relation of swelling with the square of water content present in the composite before achieving equilibrium.

Table 2 also reveals that value of n in all the cases is below 0.5 which is a case of Fickian diffusion, i.e., swelling takes place mainly by diffusion rather than chain relaxation [31]. A similar trend was reported by Purwar et al. for a similar type of hydrogel composite [32]. Contrary to this, swelling in case of sample B follows second-order kinetics at all the tested pHs (as revealed by R^2 values).

For sample B, n value is greater than 0.5 at 4.5 and 7 pH, showing anomalous (Non-Fickian) diffusion at these pHs. It shows the nearly equal contribution of diffusion and chain relaxation towards the swelling of hydrogel composites. Although the behaviour turned from anomalous to slight Fickian at higher, i.e., 8.5 pH as suggested by n value (0.46). Thakur et al. also observed a second-order swelling kinetics and non-Fickian mechanism for poly(acrylamide-*co*-acrylic acid) hydrogel system prepared by using MBAAm cross linker [11].

The weight loss after swelling experiment in the case of sample A is 40%, whereas in the case of sample B it is only 8%. It suggests that sample A disintegrates faster as compared to sample B because it holds more water due to its more hydrophilicity and large pores but the inter particles' entanglement is weaker as compared to that generated by MBAAm as cross-linker.

Antimicrobial action

Results of antimicrobial tests are revealed in Fig. 4. These Figures show a clear zone of inhibition of growth of bacterial colonies below and around the drug-loaded composite, whereas a dense population of bacterial colonies appear in the control set that contains hydrogel composite having no drug. The zone diameter for *E. coli* and *S. aureus* are 17.5 mm and 14 mm in case of sample A, whereas these are found to be 15 mm and 11 mm for sample B. This is because drug (gentamicin sulphate) destroy the bacterial cell when comes in its contact. Inhibiting action of drug-loaded composite material is a piece of evidence for drug release from both the composite samples. More diameter of zone of inhibition in sample A as compared to sample B supports the dependency of bactericidal action of composite material upon the amount of drug released, i.e., upon the swelling and drug-loading property of composite.

Drug release studies

UV–VIS spectrophotometer was used to check the amount of drug loaded in the composites. This amount was calculated from the difference of amount taken initially and the amount of drug remaining in the mother liquor after loading in the composites. The amount of drug loaded was found to be 76.45% and 55.89%, respectively, for samples A and B. Although a good amount of drug is loaded in both the cases, more drug is loaded in case of sample A due to more hydrophilicity and optimum pore size in this case.

Figure 5 reveals plots of drug release as a function of time. From these plots, it is evident that maximum drug release is observed at pH 7. Below and above this pH, hydrogen bonding and weakening of physical forces between polyacrylic acid and polyacrylamide, respectively, result in lesser release. Up to 68% drug is released in the first 12 h in the case of sample A while it is only 62% in the case of sample B.

Drug release kinetics

Drug release data were fitted in different kinetic models like Peppas model, Higuchi model, zero, first and second-order equations. These plots for sample A and B are shown in supplementary data Figure S2. The plots showing maximum linearity is supposed to be the best fit model in each case. The kinetic constants of different models for these two composites at different pHs are summarized in Table 3. From the table, it is evident that the value of correlation coefficient is

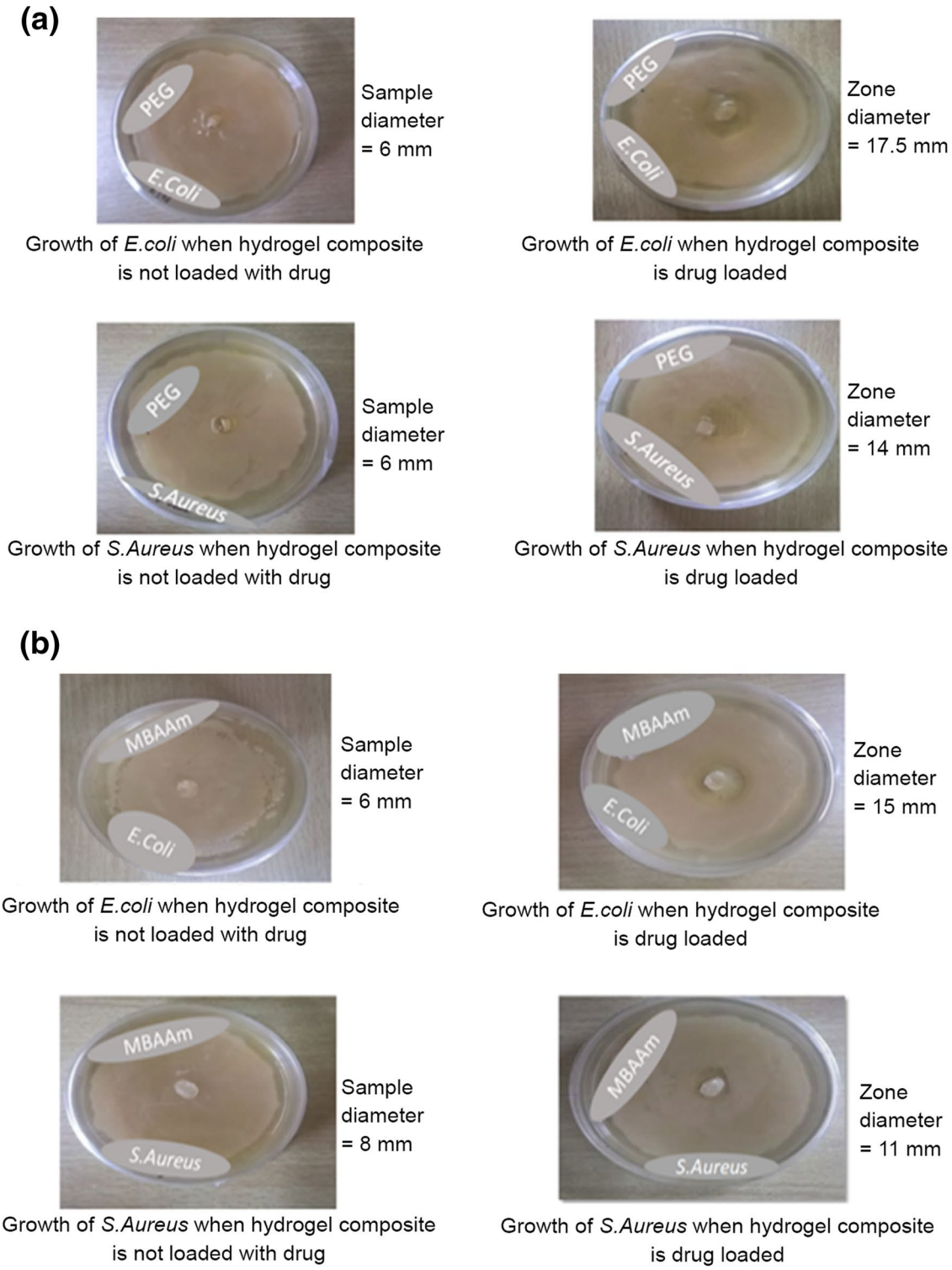


Fig. 4 Cumulative drug release from **a** sample A, **b** sample B

highest for Peppas model at all the studied pHs. It means drug release follow Peppas model in both cases. According to this model, for a slab like drug delivery system, value of release exponent provides information about the type of diffusion. A system having the value of $n \leq 0.5$ corresponds to Fickian diffusion, that having value $0.5 < n < 1$ corresponds to anomalous or non-Fickian transport and that having value ≥ 1 corresponds to case II transport [31, 33].

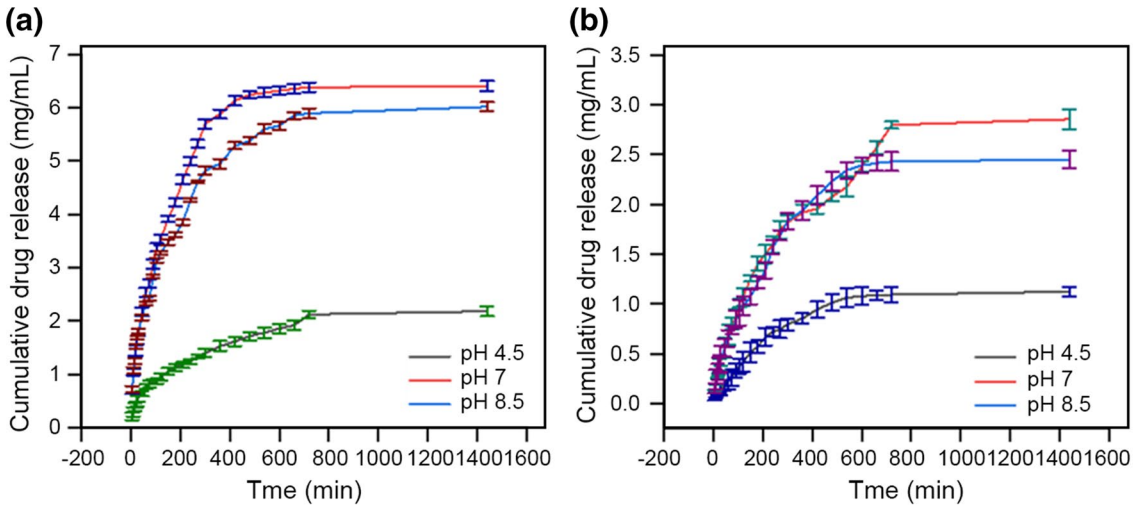


Fig. 5 **a** Antimicrobial behaviour of sample A. **b** Antimicrobial behaviour of sample B

Table 3 Kinetic constants for drug release at different pH for composite sample A and sample B

pH value		Model name						
		Peppas model			First-order		Second-order	
		R^2	n	k	R^2	k (s^{-1})	R^2	k (L mol^{-1} s^{-1})
4.5	For sample A	0.994	0.438	0.28	0.993	0.0012	0.97	5×10^{-3}
	For sample B	0.995	0.74	0.134	0.977	0.0036	0.96	1.8×10^{-3}
7	For sample A	0.99	0.43	0.29	0.99	0.0017	0.98	2.6×10^{-3}
	For sample B	0.991	0.62	0.174	0.984	0.0012	0.98	2×10^{-3}
8.5	For sample A	0.99	0.416	0.32	0.995	0.0019	0.98	1.8×10^{-3}
	For sample B	0.998	0.586	0.199	0.96	0.0023	0.93	1.7×10^{-3}

Table 3 also reveals the value of n below 0.5 showing Fickian type drug release, i.e. drug release is mainly controlled by diffusion rather than chain relaxation in case of sample A. Value of constant k also increases with a rise in pH of medium showing its dependency upon the structural constitution of the hydrogel composite. At lower pH, hydrogen bonds formed by unionised carboxylic group results in smaller pores. At a pH greater than pK_a of acrylic acid ($pK_a = 4.70$), repulsion among negatively charged carboxylate ions leads to the formation of larger pores. In alkaline medium diminishing of physical forces between polyacrylic acid and polyacrylamide further results in somewhat smaller pore size. These results are in agreement with that reported by Purwar et al. [32]. Table 3 shows value of n greater than 0.5 supporting a non-Fickian diffusion. It means drug release is governed by diffusion as well as chain relaxation in the case of sample B.

Mesh size calculation reveals a higher mesh size in case of sample A, i.e. it has more surface area so drug release is governed by mainly diffusion rather than chain relaxation, a case of Fickian diffusion. On the other hand, drug release from sample B is governed by diffusion as well as chain relaxation (a case of non-Fickian diffusion) because of its lesser mesh size (as represented in Fig. 6).

Mechanical properties

The mechanical properties of hydrogel film and composites were evaluated under wet conditions. The data are shown in Table 4. The tensile strength is lower (3.0 MPa) but extension at break is much higher (26.8 mm) in case of hydrogel film as compared to hydrogel composites, in case of which tensile strength is, respectively, 4.2 MPa and 6.23 MPa and extension at break is 7.5 mm and 9.7 mm, respectively, for samples A and B. Hydrogel composites show very high modulus as compared to hydrogel film. These results support the reinforcement of the strength of the hydrogel after grafting over the fabric.

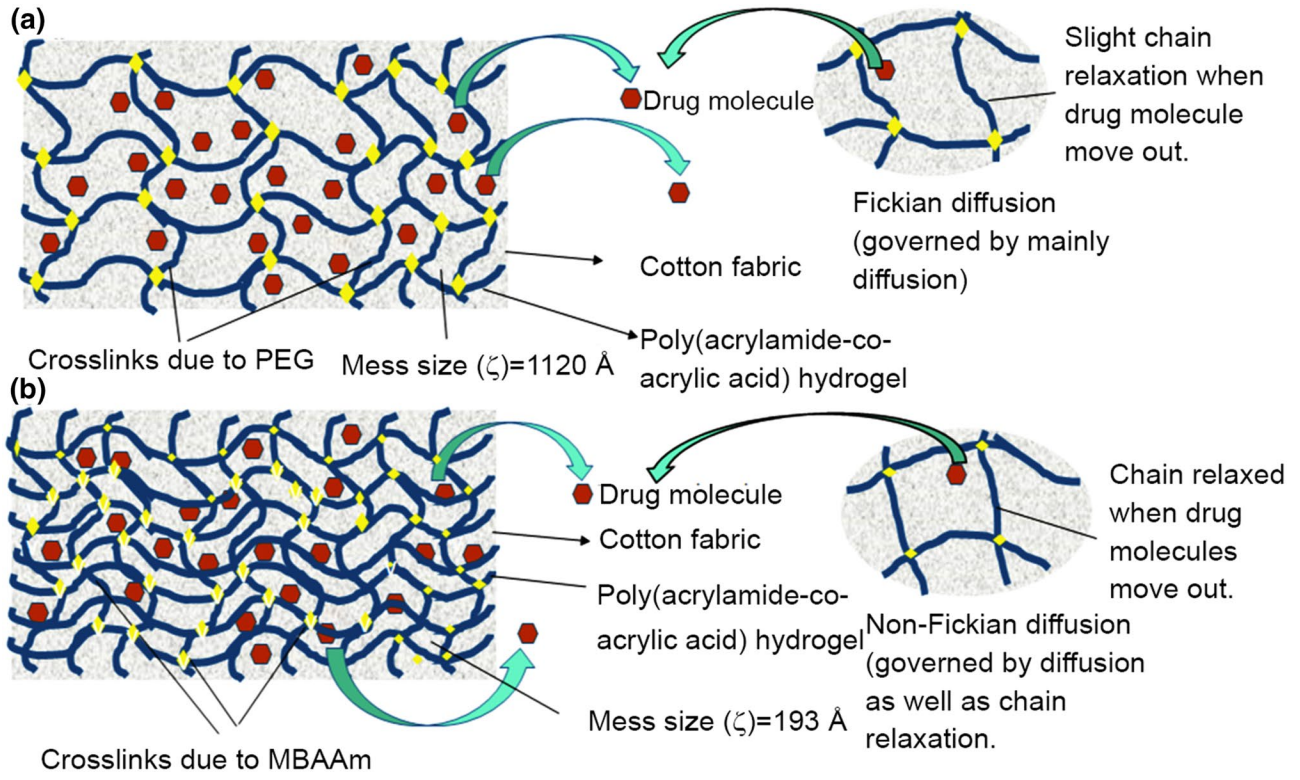


Fig. 6 Schematic diagram showing mechanism of drug release from **a** sample A, **b** sample B

Table 4 Mechanical properties of hydrogel film and composite wound dressings

Materials	Tensile stress at max load (MPa)	Tensile modulus (MPa)	Extension at break (mm)
Hydrogel film	3.07 ± 0.5 (0.03)	31.3 ± 1.4 (0.03)	26.8 ± 5 (0.02)
Sample A	4.21 ± 0.5 (0.01)	91.6 ± 5 (0.02)	7.51 ± 0.6 (0.01)
Sample B	6.23 ± 0.6 (0.03)	275.37 ± 4.9 (0.04)	9.73 ± 1.7 (0.05)

Values in the bracket show coefficient of variation (CV)

Conclusion

Hydrogel composites of poly (acrylamide-*co*-acrylic acid) hydrogel grafted over cotton fabric are prepared using cross-linkers polyethylene glycol (PEG) and *N*,*N*′-methylene bisacrylamide (MBAAm) separately. The composite prepared by using MBAAm as cross-linker has more condensed structure but the swelling ability of composite prepared with PEG is just the double as that prepared with MBAAm. Antimicrobial tests are performed to show the effectiveness of drug loaded within the hydrogel samples. Larger zone diameter also supports more drug release from the hydrogel composite prepared with PEG. Release kinetics of gentamicin sulphate from both the composites is studied in phosphate buffers of pH 4.5, 7 and 8.5 at 37 ± 0.1 °C and a faster release in phosphate buffer of pH 7 is observed. On fitting release data in Peppas model, Higuchi model, Zero, first and second-order kinetic equations, it is observed that drug release is diffusion controlled and follows Fickian diffusion mechanism in case of composite prepared with PEG as cross-linker, whereas it is controlled by diffusion as well as chain relaxation in case of the composite prepared with MBAAm. Mechanical testing using Universal Testing Machine (Instron-2700) shows a higher tensile strength (6.2 and 4.2 MPa, respectively, with MBAAm and PEG cross-linkers) of the composite as compared to the hydrogel film.

Acknowledgements This study was financially supported by UGC (University Grants Commission), under Joint CSIR-UGC (NET)-JRF scheme {F.11-3/2003(SA-I)}.

References

1. He G, Ke W, Chen X et al (2017) Preparation and properties of quaternary ammonium chitosan-*g*-poly(acrylic acid-*co*-acrylamide) superabsorbent hydrogels. React FunctPolym 111:14–21
2. Gharekhani H, Olad A, Mirmohseni A, Bybordi A (2017) Superabsorbent hydrogel made of NaAlg-*g*-poly(AA-*co*-AAm) and rice husk ash: synthesis, characterization, and swelling kinetic studies. CarbohydrPolym 168:1–13
3. Zohuriaan-Mehr MJ, Omidian H, Doroudiani S, Kabiri K (2010) Advances in non-hygienic applications of superabsorbent hydrogel materials. J Mater Sci 45:5711–5735
4. Conaghey OM, Corish J, Corrigan OI (1998) Iontophoretically assisted in vitro membrane transport of nicotine from a hydrogel containing ion exchange resins. Int J Pharm 170:225–237
5. Feng C, Shen Z, Yang D et al (2009) Synthesis of well-defined amphiphilic graft copolymer bearing poly (2-acryloyloxyethyl ferrocenecarboxylate) side chains via successive SET-LRP and ATRP. Polymer (Guildf) 47:4346–4357
6. Mohd Amin MCI, Ahmad N, Pandey M, Jue Xin C (2014) Stimuli-responsive bacterial cellulose-*g*-poly(acrylic acid-*co*-acrylamide) hydrogels for oral controlled release drug delivery. Drug Dev Ind Pharm 40:1340–1349
7. Reddy KR, Rajgopal K, Maheswari CU, Lakshmi Kantam M (2006) Chitosan hydrogel: a green and recyclable biopolymer catalyst for aldol and Knoevenagel reactions. New J Chem 30:1549–1552
8. Hapiot F, Menuel S, Monflier E (2013) Thermoresponsive hydrogels in catalysis. ACS Catal 3:1006–1010
9. Liu ZS, Rempel GL (1997) Preparation of superabsorbent polymers by crosslinking acrylic acid and acrylamide copolymers. J ApplPolymSci 64:1345–1353
10. Nesrinne S, Djamel A (2017) Synthesis, characterization and rheological behavior of pH sensitive poly(acrylamide-*co*-acrylic acid) hydrogels. Arab J Chem 10:539–547
11. Thakur A, Wanchoo RK, Singh P (2011) Structural parameters and swelling behavior of pH sensitive poly (acrylamide-*co*-acrylic acid) hydrogels. ChemBiochemEng Q 25:181–194
12. Ueshima T, Hosoda Y, Lieberman PE (1973) United States Patent (19) process for the production of highly water-absorbing but less water-soluble hydrogels
13. Parks LR (1981) Cross-linked sodium polyacrylate absorbent. US Pat, Doc
14. Purwar R, Rajput P, Srivastava CM (2014) Composite wound dressing for drug release. FibersPolym 15:1422–1428
15. Tangri A (2014) Polyacrylamide based hydrogels: synthesis, characterization and applications. Int J Pharm ChemBiolSci 4:951–959
16. Cheng WM, Hu XM, Zhao YY et al (2017) Preparation and swelling properties of poly(acrylic acid-*co*-acrylamide) composite hydrogels. E-Polymers 17:95–106
17. Bajpai SK, Dubey S (2004) Synthesis and swelling kinetics of a pH-sensitive terpolymeric hydrogel system. Iran Polym J Engl Ed 13:189–203
18. Hannuksela M, Pirilä V, Salo OP (1975) Skin reactions to propylene glycol. Contact Dermatitis 1:112–116
19. Schneider LA, Korber A, Grabbe S, Dissemond J (2007) Influence of pH on wound-healing: a new perspective for wound-therapy? Arch Dermatol Res 298:413–420
20. Gethin G (2007) The significance of surface pH in chronic wounds. Wounds UK 3:52–55
21. Isik B, Kis M (2004) Preparation and determination of swelling behavior of poly(acrylamide-*co*-acrylic acid) hydrogels in water. J ApplPolymSci 94:1526–1531
22. Gudeman LF, Peppas NA (1995) Preparation and characterization of pH-sensitive, interpenetrating networks of poly (vinyl alcohol) and poly (acrylic acid). J ApplPolymSci 55:919–928
23. Sen M, Yakar A, Güven O (1999) Determination of average molecular weight between cross-links (M̄(c)) from swelling behaviours of diprotic acid-containing hydrogels. Polymer (Guildf) 40:2969–2974
24. Singhal R, Tomar RS, Nagpal AK (2009) Effect of cross-linker and initiator concentration on the swelling behaviour and network parameters of superabsorbent hydrogels based on acrylamide and acrylic acid. Int J PlastTechnol 13:22–37
25. Tomar RS, Gupta I, Singhal R, Nagpal AK (2007) Synthesis of poly (acrylamide-*co*-acrylic acid) based superaborbent hydrogels: study of network parameters and swelling behaviour. PolymPlastTechnolEng 46:481–488
26. Majumder S, Dahiya UR, Yadav S et al (2020) Zinc oxide nanoparticles functionalized on hydrogel grafted silk fibroin fabrics as efficient composite dressing. Biomolecules 10:1–14
27. Pishbin F et al (2014) Electrophoretic deposition of gentamicin-loaded bioactive glass / chitosan composite coatings for orthopaedic implants. ACS Appl Mater Interfaces 6:8796–8806
28. Srivastava CM, Purwar R, Gupta AP (2019) Enhanced potential of biomimetic, silver nanoparticles functionalized *Antheraea mylitta* (tasar) silk fibroin nanofibrous mats for skin tissue engineering. Int J BiolMacromol 130:437–453
29. Tomar RS, Gupta I, Singhal R, Nagpal AK (2007) Synthesis of poly(acrylamide-*co*-acrylic acid)-based super-absorbent hydrogels by gamma radiation: study of swelling behaviour and network parameters. Des Monomers Polym 10:49–66
30. Turan E, Tuncer C (2007) Swelling and network parameters of pH-sensitive poly (acrylamide-*co*-acrylic acid) hydrogels. J ApplPolymSci 106:2000–2007
31. Ritger PL, Peppas NA (1987) A simple equation for description of solute release II. Fickian and anomalous release from swellable devices. J Control Release 5:37–42
32. Agrawal A, Purwar R (2018) Swelling and drug release kinetics of composite wound dressing. Indian J Fibre Text Res 43:104–111
33. Lamprecht A, Yamamoto H, Takeuchi H, Kawashima Y (2003) Microsphere design for the colonic delivery of 5-fluorouracil. J Control Release 90:313–322

2020 IEEE International Conference for Innovation in Technology (INOCON)
Bengaluru, India. Nov 6-8, 2020

Intrusion Detection System a Need

Raj Kishore
Information Technology
Delhi Technological University
New Delhi, India
raj.verma5454@gmail.com

Anamika Chauhan
Information Technology
Delhi Technological University
New Delhi, India
anamika@dce.ac.in

***Abstract*: - This review paper is concerned with intrusion detection systems and its several types. As we know that today's era is of computer networks or of internet of things which can lead to intrusion in our systems and can be devastated to it, so intrusion detection system can help computer administrators to curb such activities and prevent our systems. The main aim for this review paper is that to provide information about the need of intrusion detection systems in day to day life.**

Keywords—Intrusion detection, anomaly based, signature based intrusion detection, intrusion detection system.

I. INTRODUCTION

An Intrusion Detection System is a software to monitor and protect our network from any kind of intrusions. With the rapid growth of internet, fear of cyber intrusion increases [1]. The areas like financial, business, Industries, Health sector, security, WAN , LAN application have progressed. These sites have made the internet a favourable site to get abused and can lead to the threat of losing the personal information, data which is very big Susceptibility for the network community [1]. Malevolent users or we can say hackers' hacks the organizations private data to collect information and causes problems in the system or vulnerabilities such as software failure, temporary failure of administration like they set the systems on default settings which can be easily prune [2].

An IDS is known as burglar alarm. Like there is lock system in our house to protect it from theft. If somebody tries to enter into the house after breaking the lock system, then this burglar alarm detects that the locking system has been broken and raise the alarm to alert the owner. However Firewall is doing great job, it filters the incoming traffic from the network to bypass the firewall [4]. Like outside users can connect to internet by installed modem and dialling through modem which is installed in private network (VPN) of the company, this access control cannot be detected by firewall.

An (IPS) Intrusion Prevention system is used to guard network and ensures threat/security prevention technology which is used to audit and monitor network traffic flows to detect/notice and prevent vulnerability utilize. There are two types of Network Prevention system Host (HIPS) , Network (NIPS). These network systems are smart as they automatically watch the network traffics find out the necessary actions to be taken to protect system and networks. Issue with IPS is false negatives and positives. False positive defined as when there is no attack and still produces an alarm. False negative is define as when there is an attack and still no alarm to alert the owner.

II. TYPES OF IDS

1) *Host Based.*
2) *Network based.*
3) *Application based.*

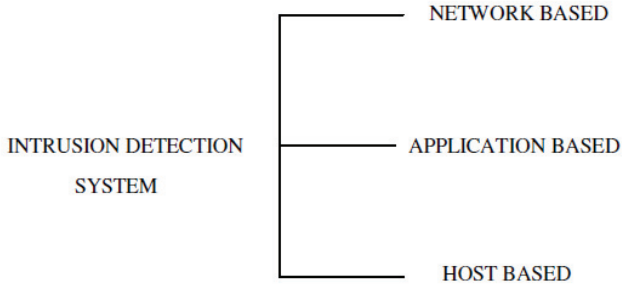


Fig. 1. Types of IDs

Host based: Host based intrusion detection systems (IDS) monitors the sign of malpractices in the local system. For the process of analyzing they use the other information and system's logging. Handler of Host based is basically referred as sensor [12] [13] [14]. Host based sensor collects the data from other sources which includes the log information of the system , log files created by the operating system (OS) processes which also contains content of objects which is not shown logging mechanism of operating system [5].

Host Based IDS blindly trust on the monitored trail. The data allows the IDS (intrusion detection system) to find out the several pattern of malpractices which would not be available at the at high level of abstraction [6] . The basic in IDS including NIDS starts from anomaly HIDS research which is based on Paramount work of Denning's [7]. Host based IDS compared to Network based IDS provides far better results. HIDS sometimes provide amazed results about intrusion like If there is attack then it can tell what type of attack was that, what type of commands were used by the attacker , which files they were accessing . It provides accurate results rather than false accusation and tells whether there was dangerous command which was executed [8]. It is less vulnerable or risky.

Advantages

- Verifies if the attack was successful or not.
- Always monitors system activity.
- It does not require any sort of external hardware.
- Near to real time protection.
- Lower entry cost.

978-1-7281-9744-9/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:03:32 UTC from IEEE Xplore. Restrictions apply.

Network based: Network based IDS systems gathers data from the internet by self rather than collecting data from individual host [9]. The NIDS go through the network malpractice while flows of packets are there in the network. The sensors of the network come with equipped signature of attacks that are rules through which an attack can be done. Most of the NIDS (network based intrusion detection system) allow the owner to define their own signature [9]. Malpractice on the network is based on sensors signature which are from their previous attacks and the process of being audited will also be transparent to the owner and this is very significant [10].

The clarity of the audit decreases the possibility of counselling that can locate it in future and void its potential without the efforts [6]. Network agents of the node are there at every host in the network coverage range which is being protected [11].

Advantages

- Low cost to own.
- Easy to deploy.
- Retaining evidence.
- Failed attacks can be detected.
- Detect network attacks.
- Quick response with real time detection.

Application based: Application Based intrusion detection System (APIDS) will monitor the successful actions and happening of the protocol [11]. The agent or the monitoring system is placed between sack of server and process that audit and analyzes the protocol between devices and the application [11]. Deliberate attacks are malevolent attacks which are done by aggrieved hired hand to cause distress to the organization and unintentional attacks causes' financial distress to the company by vandalizing the most important files of the organization [11]. There are many attacks which can be takes place in the OSI layer.

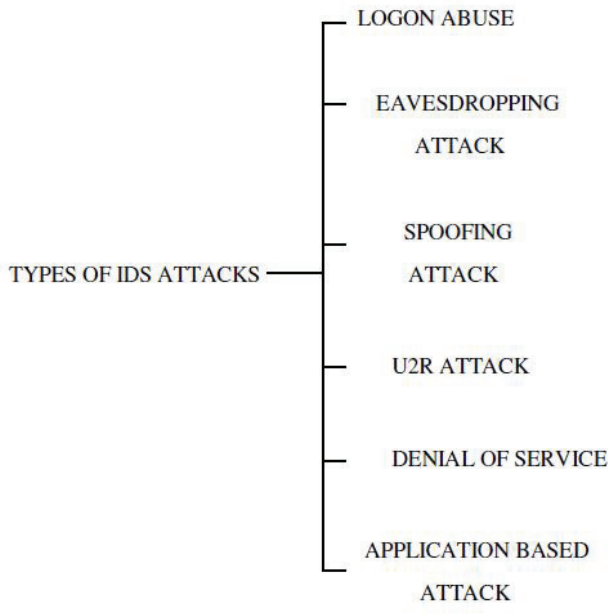


Fig. 2. Types of attacks

Denial of service: - A denial of service attack is a malpractice attempt which makes the server and network unavailable to the users. In this attack hacker sends plenty of requests to the server and makes it chock. For every server there is a limit to reply no of queries made by it, if the queries exceeds the limit server gets crashed [11].

Denial of service attack can we prevented by seeing plenty of queries are coming from the same IP address (Internet Protocol) and the owner can block that particular IP address. So to overcome this hackers use DDOS (Distributed denial of service attack). In this attack hacker attacks the system after giving requests from many individual systems. If request is coming from different sources, it is difficult for the user to identify the DDOS attacks, as there may be different users giving request which makes it hard to detect.

- Ping of death: In this the hacker sends targeted system a ping request which is way bigger than 65,536 bytes which leads to system crash. The normal size must be 56 or 84 bytes if we consider internet protocol header [11].
- Sync attacks: As cleared by its name SYNC attack means synchronous attack in which flood of request is being send from the hacker's side to server, which is beyond the capacity of the server to handle the entire request simultaneously and lead to system crash.

Eavesdropping Attacks: - In this attack hacker mainly affect the communication process. It is mainly done on telephone lines, electronic mails and several others communication channel. When two users are connected through some channel for communication, it may be internet, telephone line etc. then hacker manipulates the communication channel , one user is sending the data to legitimate user in the encrypted form and gives the user a private key to decrypt the file. Hacker uses that key to encrypt the data , and send the manipulated data to the user [11].

Spoofing Attacks:- In this attack hacker portrays itself as a legitimate user to steal , vandalize the data and sometimes takes advantage of the illegal events on the network protocol [11] [12]. It gain access to the system after changing its own IP address to the legitimate user . Systems get easily fooled by the trick , it always think the communication which has been established is done by the legitimate user but in reality it's the attacker. System despite of real fact gives the access to the attacker into the system.

User to root attacks (U2R):- In this attack attacks tries to gain access into the system through internet after data leak in the system. This data leak occurs when server receives more data than its capacity and not being programmed to handle that much amount of data, in this condition buffer overflow happens. Buffer overflow also happens if the program to handle the server is not good , like buffer overflow can occurs in the program of the system. Like string overflow and integer overflow if we try to store more data into the memory than also there is a leak in the memory which can be used by the hacker to gain access into the system through that memory leak, they try to inject malicious code into that leaked memory. The malicious code compiled itself and makes .exe file which automatically get executed and can provide the access to its owner it works like someone is serving for its owner [11] [12].

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:03:32 UTC from IEEE Xplore. Restrictions apply.

Logon abuse attack: - It is also called brute force attack. In this type of attack attacker does not care about the authentication, they directly access the control management of the system than the system grant an attacker with more advantages [11].

Application level attack: - In this attack hacker targets the weaker part of the application layer. For example there is weakness in the security of the server side or there are certain fault controls towards the server end [11].

III. Functions of the Intrusion Detection Systems (IDS)

- Collection of data
- Selection of features
- Analysis
- Action

Data collection: - This functionality of IDS passes the data to IDS as input. This mechanism is very useful factor for the IDS which can also affect the accuracy of the intrusion detection system (IDS). In data collection we don't apply hard sampling technique of statistics which can lead to bulkier system with more time to react, we simply use simple random sampling technique for the procedure of data collection which can provide new data collection data frame for IDS [16]. IDS records the data into a single file then analyze it. As network based IDS (NIDS) gather and adulterate the data packets where as host based IDS (HIDS) gathers the information of physical system like process of the system, usage of the disk. As proved the by experimental results data model can progress in the effectiveness of the data collection and improve the initial processing of IDS.

Feature selection: - This is a pre-processing feature of IDS which can lead to solve IDS queries smartly by selecting relevant data and removing redundancy with the irrelevant features [15]. To select the feature data which is available in the network , then this data will be use to evaluate any intrusion in the system. Like there are several ways which can be used as key for intrusion follow as IP address of targeted system, IP address of source, length of head, protocol type. Relevant data may have useful information of the classes, which can be essential for correct procedure of the classifier.

Analysis: - In this functionality of IDS, data which has been selected after the data selection and feature selection thoroughly analyzed to find whether it is correct or not. There is further division of IDS based on rule based and signature based. In rule based IDS it analyzed the data on the other hand predefined signature or patter will be checked against incoming traffic [15] [16]. There is another IDS known as anomaly based IDS in which Mathematical model are implemented for studying the system behaviour [16].

Action: - In this functionality it is defined about the reaction time of systems on attacks which are carried by hackers [17]. It can be of different type like it may alert the owner through mailing system or through alarm or it can play smartly in the system and start dropping the packets which are being sent by the intruder so that it cannot gain access through the system or even it can block the IP , if hackers id using DDOS then it may block the entire port and stop the communication.

IV. IDS Life Cycle

As there is vast increase in the networking areas which enhance the fear of getting intrude. Fear of intrusion increases the production of IDS , organizations frequently releases IDS with aggressive attitude to compete with the latest intrusion which is in the market [18]. Estimation about the new systems is not very efficient. Steps like hiring and training the workers with the administrator security , network security, benefits of IDS and even also about IDS is challenging task [18]. As the IT industries are growing faster and faster in which new technologies are coming frequently. As the new technologies are coming , hackers are also thinking differently which leads to new type of intrusion in the systems.

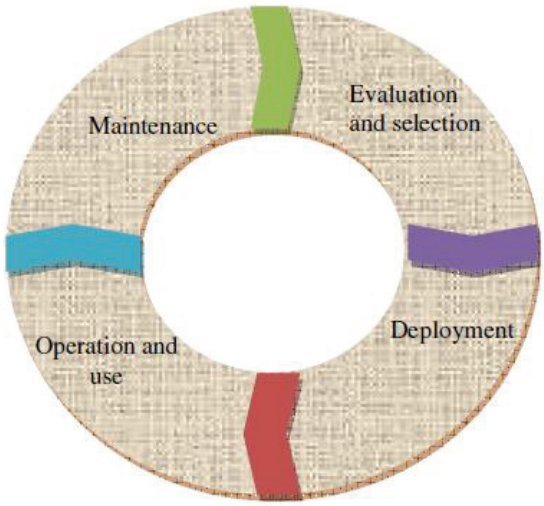


Fig. 3. IDS Life Cycle

Evaluation and selection: - If there is need of an IDS system into an organization and planning to get, it should definitely inspect the assets for the system working and maintenance [18]. Lifecycle of the product can be increased by productive IDS. There is also option of third party evaluation for the IDS and their outcomes are generally available openly [18]. This certain process can tell us about how to find an intruder and how much amount of work needed for maintaining the system. Selection process in the system defines as approach of IDS , Identification of the characters, effectiveness and the accuracy of the IDS.

Deployment: - In this phase it ensures working of sensor to increase the protection of the important files for the organization by configuring intrusion detection system with new attack policy after installing signatures [18]. Users of the IDS must design some new rules for monitoring the alerts and to synchronize with the other systems of the organization. The working group of Intrusion detection system (IETF) is creating a familiar attentive format that will use the IDS to alert different systems which will reportedly have a common display screen or console [18].

Operation and use: - Organization conducts the IDS to audit the host for responding on the report as alert. It builds roles and authority for auditing and monitoring results of manual and automotive responses [18]. Smart hackers have plenty of knowledge about the IDS systems, if they realize that in certain network there is IDS which has already been

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:03:32 UTC from IEEE Xplore. Restrictions apply.

deployed on internet attack done by them, then they force IDS to provide false report to the owner of the organization [18].

Maintenance :- Maintenance is very important factor of the IDS system because intrusion is increasing day by day; new attacks are there in the network field so we should install new signatures and upgrade the IDS for better security of the organization. We should check periodically the sensor of the system whether it ensures network changes [18]. For an organization it must attract qualified employees, train them and retain the qualified employees for better maintenance of IDS technologies [18].

V. IDS TECHNOLOGIES

A. Anomaly based IDS

Anomaly based intrusion detection system is a system used for detecting or identifying the intrusion in network as well as in computers[3], if any misuse is there by auditing the activity of the systems and classify it whether normal or anomalous with the help of predefined classical classifiers. The categorization is based on predefined rules or heuristics, instead of old pattern and signatures [19]. It will try to detect any type of malicious activity which does not fall in normal activity of system operations. This type of systems is good in respect of signature based intrusion detection because they can only detect the attack for which the signatures has already been defined [11].

In order to make our system feasible and identify the attacks with greater accuracy in network traffic attack, then the system must know about the normal activity of the systems and can able to distinguish between normal and abnormal activity. There are two major phases of anomaly based intrusion detection system which are training phase (where instructions are built for normal behaviour) and testing phase (where current network packets are comparing with the defined behaviour in the training phase) [20]. There are several ways for anomaly detection in the systems most common is Artificial Intelligent Techniques. Systems which are using neural networks have always given the best result. Techniques of deep learning using auto encoders [21], or generative models also give remarkable results. Another method is to define the normal usage activities of the system using strict mathematical models or Boolean flag. If any deviation occurs according to the defined model then it is an attack. This method known as strict anomaly detection [22]. It provides good results for finding attacks like buffer overflow or memory leak, Denial of service (DoS), Distributed Denial of service (DDoS), application overflow anomaly.

Techniques used in anomaly based IDS

1) Statistical model

- *Operational model*: - The action that happens over time which can regulate the alarming system. This can be evoked with WIN2K lock. Intruders after 'z' ineffectual access attempts can regulate the alarm. With the defined lower limit as 0 and upper limit as 'z'.
- *Markovian process*: - A stochastic process is collection of random variables with rest to time, so the system is inspected at fixed time and read the behaviour [24]. If the behaviour is anomaly then the probability of the current state is low as compared to the previously defined in the training set.

2) Cognition based detection Techniques

CBDT works on the audited data which uses the predefined rules of classification for classes and tributes [17].

- *Finite state machine*:-This is used to capture model of the behaviour of states, transition and actions. We define state as the past information. An action contains the elaboration of the activity which is to be performed lately [17] there are certain actions entry, exit transition action [17].
- *Description Scripts*: - Scripting languages like python can easily characterize the intrusion on computers networks. All scripting languages are self sufficient for examination of the specific event [17].

a) Cognition Based Detection Techniques

CBDT works on the audited data which uses the predefined rules of classification for classes and tributes [24].

- *Boosted Decision Tree*: - It uses ADA boost to create decision tree classifier based on the predefined data in the process of training in IDS [17].
- *Support Vector Machine (SVM)*:- - It is used for binary classification. If we merge decision tree and SVM then it will provide efficient results.

B. Signature based IDS

Signature based IDS also known as misuse detection. In this large dataset of instances is there and every instance of the dataset must be labelled with the normal or unwanted. Machine learning techniques have been used to train the data according to the labels. This method itself generates the signature to detect the intrusion. Misuse technique will get created automatically by the user, but work is harder and accurate compared to manually created [17].

Techniques used in misuse detection

- *Express matching*: - This technique is the simplest to implement in misuse detection. In this it will look a sack of like events ex- log entries for finding the exact pattern.
- *State transition analysis*: - This model intrudes the state or transition of network. Every occurrence of event in internet network will be applied to finite automata machine instances which will results in transition. It is always believed that an attack will be occurred in the final state of machine.

C. Target monitoring

In this technique IDS report the user if there is any modification happen in any location in the system. This is usually done through cryptographic algorithms, in which it calculates crypto checksum for each targeted file in the system [17]. If any changes occurred it will immediately informed to IDS. The name of checker which checks is any changes occurred or not is Tripwire checksum [24].

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:03:32 UTC from IEEE Xplore. Restrictions apply.

D. Stealth Probes

This technique is used to gather and associates the data. It can detect the intrusion which generally takes long time to react. Hackers waits for over months to see the change in the system and also wait for another several months to generate attack in the system and take wide area of samples.

VI. Tools for IDS

There are several tools available for intrusion detection for the organization to increase the security.

A. Snort

Snort is open source software which is light and easy to understand. It uses flexible rule-based language to give information about data packets. Snort can detect vulnerability in the system, worm's suspicious behaviour.

B. OSSEC-HIDS

This is also a open source software which uses client/server architecture. It is recommended to run this software on major operating systems. Provide authentication logs, ISP, HIDS etc.

C. Fragroute:

This is a fragmenting router. IP packets are sent to fragroute from attacker then fragment it and transformed to the party.

D. HoneyD:

It is a tool to create a virtual server, where attackers try to intrude it and our actual database is safe from the attacks. IDS analyze the type of attack and make our database even more invincible.

E. Kismet

It is basically a tour guide for the use of (wireless intrusion detection system). It finds the burglar point.

VII. Needs and challenges

There is many challenges for installing IPS devices in our systems. It is designed to work for network security to identify the intrusion and prevent the system. IPS continuously audits our network to find the possible malicious activity and also capture the information about them. Some attacks can be remained undetected if passive IDS fail and if inline device fails then this will affects the performance of the network [23]. Network intrusion prevention systems should perform like a switch of network. It also meets robustness and performance of the system in order to organise the challenges. There are no less than none customers are there to cost the reliability and performance of the network for the security [23].

The issue with NIPS (Network intrusion prevention systems) is that it drops the packets and also slows down the network which will be used for data stream. Most of the high usage vendors of IPS will certainly get these types of problems with hardware like FGSAs and ASICs [23]. It is important to build the product which can perform like intrusion detection and prevention device [23]. All the organizations IDS like a defence tool for their safety. There are numerous challenges organizations are facing while installing an intrusion detection systems.

1. IDS is a system which has all technologies related to current malicious activities , but as the time flows new malicious activities will evolve which can make our system out dated so we have to do lots of enhancement to system time to time, which is really important for the safety of the organizations. Ids doesn't need any kind of human intervention as it is becoming more automotive day by day like notify the administrator/owner if any kind of malicious activity happens, block the malicious connection for a period of time, for stopping malicious connection it dynamically modifies a access control list of router [2]. The logs should be monitored for every event in the system. Daily basis logs monitoring is required to analyze the activities which are detected by the IDS [2].
2. IDS instalment only depends on the success of the deployment phase. Planning plays a major role for design and implementation. As we have already discussed about network based and host based IDS, so in some cases we might have to implement hybrid solution of these two IDS [2], but this decision is purely based upon organizations whether to build hybrid solution or use some other technology. As we know the fragility of the Network based IDS is that it can monitor many systems simultaneously and it doesn't require a software to get installed on manufacturing system like Host based IDS which makes it swift choice of several organizations [2].
3. For designing new IDS it is really necessary to create baseline policy and also able to avoid false positive results. The ration of sensor manager must be made public. It is not necessary that IDS always gives right results there are some drawbacks also and can give false positive alarm to the sensor which can lead to inadequate ratio [2].
4. As time grows the enhancements in IDS also takes place but still it is reactive rather than proactive like it reacts when there is some malicious activity which purely works on signature of attack [2]. Signatures are basically predefined patterns of attacks. The database of predefined signatures must be updated and fixed in the database time to time whenever new attacks accounts in the system. The frequency of updating signature must be varies for different vendors [2].
5. The incoming traffic and outgoing traffic from the network port is not seen from the other host ports because of collisions in the network. But in HUB port there is a different scenario, we can see in and out traffic flow from any port. Network based intrusion detection systems needs sensor to monitor the in and out flow any port and sensors are also used to monitor the malicious traffic in the environment. So to accomplish this port mirroring or spanning is to be used [2].

VIII. IDS in various domains

IDS is used in various domains some of them are discussed below

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:03:32 UTC from IEEE Xplore. Restrictions apply.

A. IDS in MANET

Manet is a mobile adhoc network which is decentralized wireless network. Network never relies on pre-existing rules that is why network is adhoc. Each node in Manet participates in routing and forward the data for other nodes as well, so it is hard to determine which node will forward the data, the decision is purely dynamic in nature which is based upon network connectivity and the algorithm which is used for routing [25]. IDS are also used in Manet as it is mainly used in battlefield field area for the purpose of military actions and to get proper network coverage [25]. Normally the signal divided into no of packets and use modem and wire for communication. But Manet is wireless network in which two techniques are applied on protocol in Adhoc such as watchdog and path rater.

Watchdog identifies any intrusion in the network by the techniques of eavesdropping in the communication channel to the next hop [25]. Path rater efficiently finds the path which does not contain the nodes which are used in intrusion [25]. Intrusion detection systems are used in Manet to find the malicious activity in the network by sending lots of (series) packets to the destination through wireless mobile network.

B. IDS for cloud computing

Cloud computing is internet based server where virtual servers provides software infrastructures and other services and hosting user as pay and use the service [27]. The user of the cloud computing does not require to have physical hardware's instead they pay the third party. The payment is done only for the usage. IDS play a major role in network security and perseverance of the network which is active in any organization [26]. In this the applications of the cloud computing is accessed by the remote side and they have full control over usage of the data provided to the users. IDMEF standard is used to communicate in cloud application which is intrusion detection message exchange format.

Security issue in cloud computing

- Confidentiality of the data.
- Attack on remote side server.
- Security auditing of clouds.
- Interoperability of data is less.

C. IDS in data mining

Data mining is the techniques of extracting useful information from the pool of data. There are five main steps of data mining like identification of the source information, pick the points in the database to analyze it, extract relevant information, key data from the extracted dataset, giving results. IDS are really important in the data mining. It basically identifies the set of malicious activities which can affect the integrity and availability of the resources [28].

IDS in data mining have two different components which are anomaly detection and misuse detection. In misuse (Signature) detection the data is build using predicted model [28]. In the anomaly based intrusion detection system there is divergence among the models. First it must be transformed into featured data and then the data mining techniques are applied to use the data and then it will provide the results.

Challenges

- Generally contains large datasets.
- High dimensions of the datasets.
- Data pre-processing requires lots of work.

IX. CONCLUSION

As we have seen many things about intrusion detection systems in this paper which is also main objective to provide overview of IDS like necessity, Utility of IDS. This paper also gives complete idea of IDS life cycle, different types of attacks and tools, different domains. IDS are becoming more and more essential for the network users and the security in the organizations. Intrusion prevention systems give idea about the prevention for the security. In the life cycle of IDS the developed phases and stages are briefly illustrated, still there are many more challenges which we want to overcome. Signature based and anomaly based techniques are briefly discussed and many others techniques can be used. Future work will be done via comparing the analysis of the some popular classical machine learning or data mining algorithms applied of IDS which can help in enhancing the classification using feedback, feed forward, deep neural methods.

REFERENCES

[1] PeymanKabiri and Ali A.Ghorbani-"Research on Intrusion Detection and Response Survey"- International Journal of Network Security, Vol.1, No.2, PP.84–102, Sep. 2005.

[2] Christopher Low –"Understanding Wireless attacks &detection "- GIAC Security Essentials Certification (GSEC) Practical Assignment 13 April 2005 -SANS Institute InfoSec Reading Room.

[3] M. Latah and L. Toker, "Towards an efficient anomaly-based intrusion detection for software-defined networks," in *IET Networks*, vol. 7, no. 6, pp. 453-459, 11 2018.

[4] A. Borkar, A. Donode and A. Kumari, "A survey on Intrusion Detection System (IDS) and Internal Intrusion Detection and protection system (IIDPS)," *2017 International Conference on Inventive Computing and Informatics (ICICI)*, Coimbatore, 2017, pp. 949- 953.

[5] Bace, Rebecca-"An Introduction to Intrusion Detection &Assessment"- Infidel, Inc. for ICSA, Inc.

[6] Rebecca Gurley Bace-"Intrusion Detection"- Macmillan Technical Publishing, 2000.

[7] Denning, Dorothy E. – "An Intrusion Detection Model"- Proceedings of the Seventh IEEE Symposium on Security and Privacy May 1986.

[8] F. Sabahi and A. Movaghar, "Intrusion Detection: A Survey," *2008 Third International Conference on Systems and Networks Communications*, Sliema, 2008, pp. 23-26.

[9] "Global Information Assurance Certification Paper"- Copyright SANS Institute Copyright SANS Institute Author Retains Full Rights" 2014.

[10] "SANS penetration testing copyright by SANS"-Copyright SANS Institute Author Retains Full Rights 2014.

[11] Karthikeyan .K.R and A. Indra- "Intrusion Detection Tools and Techniques a Survey.

[12] F. Sabahi and A. Movaghar, "Intrusion Detection: A Survey," *2008 Third International Conference on Systems and Networks Communications*, Sliema, 2008, pp. 23-26.

[13] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat and S. Venkatraman, "Deep Learning Approach for Intelligent Intrusion Detection System," in *IEEE Access*, vol. 7, pp. 41525-41550, 2019.

[14] A. K. Saxena, S. Sinha and P. Shukla, "General study of intrusion detection system and survey of agent based intrusion detection system," *2017 International Conference on Computing, Communication and Automation (ICCCA)*, Greater Noida, 2017, pp. 471-421.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:03:32 UTC from IEEE Xplore. Restrictions apply.

[15] A. Gül and E. Adalı, "A feature selection algorithm for IDS," *2017 International Conference on Computer Science and Engineering (UBMK)*, Antalya, 2017, pp. 816-820, doi: 10.1109/UBMK.2017.8093538.

[16] K. Zhao, M. Zhang, K. Yang and L. Hu, "Data Collection for Intrusion Detection System Based on Stratified Random Sampling," *2007 IEEE International Conference on Networking, Sensing and Control*, London, 2007, pp. 852-855, doi: 10.1109/ICNSC.2007.372892.

[17] Sriram Sundar Rajan, Vijaya Krishna Cherukuri-"An Overview of Intrusion Detection Systems 2012.

[18] John McHugh, Alan Christie, and Julia Allen- "The Role of Intrusion Detection Systems"- Software Engineering Institute, CERT Coordination Center, pp. 29041-29053, 2018.

[19] Wang, Ke (2004). "Anomalous Payload-Based Network Intrusion Detection" . *Recent Advances in Intrusion Detection*. Lecture Notes in Computer Science. Springer Berlin. 3224: 203–222. doi:10.1007/978-3-540-30143-1_11. ISBN 978-3-540-23123-3. Archived from the original on 2010-06-22. Retrieved 2011-04-22.

[20] Khalkhali, I; Azmi, R; Azimpour-Kivi, M; Khansari, M. "Host-based web anomaly intrusion detection system, an artificial immune system approach". *ProQuest* .2011.

[21] S. Zavrak and M. İskefiyeli, "Anomaly-Based Intrusion Detection From Network Flow Features Using Variational Autoencoder," in *IEEE Access*, vol. 8, pp. 108346-108358, 2020, doi: 10.1109/ACCESS.2020.3001350.

[22] A. Taylor, N. Japkowicz and S. Leblanc, "Frequency-based anomaly detection for the automotive CAN bus," *2015 World Congress on Industrial Control Systems Security (WCICSS)*, London, 2015, pp. 45-49, doi: 10.1109/WCICSS.2015.7420322.

[23] "Intrusion Detection and Intrusion Prevention"-Ed Sale VP of Security Pivot Group, LLC.

[24] Carl Endorf, Eugene Schultz, Jim Mellander "Intrusion detection & prevention" by Written-published by McGraw-Hill.

[25] Shankar Sharan Tripathi, Sonu Agrawal- "A Survey on Enhanced Intrusion Detection System in Mobile Ad hoc Network"-International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 1, Issue 7, September 2012.

[26] Ms Asiya Jaleel -"Security Challenge in Cloud Computing"- Provided by International Journal of Engineering Sciences & Research Technology (IJESRT), Feb 2014.

[27] Ms. Parag K. Shelke, Ms. Sneha Sontakke, Dr. A. D. Gawande – "Intrusion Detection System for Cloud Computing". International Journal of Scientific & Technology Research Volume 1, Issue 4, May 2012 ISSN 2277-8616 67 IJSTR©2012.

[28] Aleksandar Lazarevic,Jaideep Srivastava, Vipin Kumar-"Data Mining for intrusion detection"- Tutorial on the Pacific-Asia Conference on Knowledge Discovery in Databases 2003.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:03:32 UTC from IEEE Xplore. Restrictions apply.

CEEW
THE COUNCIL

SHAKTI
SUSTAINABLE ENERGY
FOUNDATION

Is Ex-situ Crop Residue Management a Scalable Solution to Stubble Burning?

A Punjab Case Study

L. S. Kurinji and Sankalp Kumar

Report | January 2021





In 2018, 65 per cent of paddy residue (nearly 13 million tonnes) was set on fire in the fields of Punjab, choking the air in the entire Indo-Gangetic plains (Ministry of Agriculture & Farmers Welfare 2018).

Image: iStock

CEEW
THE COUNCIL

SHAKTI
SUSTAINABLE ENERGY
FOUNDATION

Is Ex-situ Crop Residue Management a Scalable Solution to Stubble Burning?

A Punjab Case Study

L. S. Kurinji and Sankalp Kumar

Report
January 2021
ceew.in

Copyright © 2021 Council on Energy, Environment and Water (CEEW).

Open access. Some rights reserved. This report is licensed under the Creative Commons Attribution-Noncommercial 4.0. International (CC BY-NC 4.0) license. To view the full license, visit: www.creativecommons.org/licenses/ by-nc/4.0/legalcode.

Disclaimer:

The views expressed in this report are those of the authors and do not reflect the views and policies of the Council on Energy, Environment and Water or Bloomberg Philanthropies.

The views/analysis expressed in this report do not necessarily reflect the views of Shakti Sustainable Energy Foundation. The Foundation also does not guarantee the accuracy of any data included in this publication nor does it accept any responsibility for the consequences of its use. For private circulation only.

Suggested citation :

Kurinji, L. S, and Sankalp Kumar. 2021. *Is Ex-situ Crop Residue Management a Scalable Solution to Stubble Burning? A Punjab Case Study.* New Delhi: Council on Energy, Environment and Water.

Cover image:

iStock.

Peer reviewers:

Prof. Vish Krishnan, Professor & Chair, University of California; Priya Shyamsundar, Lead Economist, The Nature Conservancy, Mohit Sharma, Associate Counsellor, CII-ITC Centre of Excellence for Sustainable Development, Udhaya Kumar V, Senior Research Analyst, Center for Study of Science, Technology and Policy Rishabh Jain, Manager – Market Intelligence, CEEW Centre for Energy Finance, and Niti Gupta, Programme Associate, CEEW.

Publication team:

Alina Sen (CEEW), Venkatesh Krishnamoorthy, Priyanka Adhikari, and Friends Digital.

Organisations:

The **Council on Energy, Environment and Water** (ceew.in) is one of Asia's leading not-for-profit policy research institutions. The Council uses data, integrated analysis, and strategic outreach to explain – and change – the use, reuse, and misuse of resources. It prides itself on the independence of its high-quality research, develops partnerships with public and private institutions, and engages with wider public. In 2020, CEEW once again featured extensively across nine categories in the 2019 Global Go To Think Tank Index Report. The Council has also been consistently ranked among the world's top climate change think tanks. Follow us on Twitter @CEEWIndia for the latest updates.

Shakti Sustainable Energy Foundation seeks to facilitate India's transition to a sustainable energy future by aiding the design and implementation of policies in the following areas: clean power, energy efficiency, sustainable urban transport, climate change mitigation and clean energy finance.

Council on Energy, Environment and Water
Sanskrit Bhawan, A-10, Qutab Institutional Area,
Aruna Asaf Ali Marg, New Delhi - 110067, India

About CEEW

The Council on Energy, Environment and Water (CEEW) is one of Asia's leading not-for-profit policy research institutions. **The Council uses data, integrated analysis, and strategic outreach to explain — and change — the use, reuse, and misuse of resources.** The Council addresses pressing global challenges through an integrated and internationally focused approach. It prides itself on the independence of its high-quality research, develops partnerships with public and private institutions, and engages with the wider public.

The Council's illustrious Board comprises Mr Jamshyd Godrej (Chairperson), Mr Tarun Das, Dr Anil Kakodkar, Mr S. Ramadorai, Mr Montek Singh Ahluwalia, Dr Naushad Forbes, Ambassador Nengcha Lhouvum Mukhopadhaya, and Dr Janmejaya Sinha. The 100 plus executive team is led by Dr Arunabha Ghosh. CEEW is certified as a **Great Place To Work®**.

In 2020, CEEW once again featured extensively across nine categories in the 2019 Global Go To Think Tank Index Report, including being ranked as **South Asia's top think tank (15th globally) in our category for the seventh year in a row**. CEEW has also been ranked as South Asia's top energy and resource policy think tank for the second year running. It has consistently featured among the world's best managed and independent think tanks, and twice among the world's 20 best climate think tanks.

In ten years of operations, The Council has engaged in 278 research projects, published 212 peer-reviewed books, policy reports and papers, created 100+ new databases or improved access to data, advised governments around the world nearly 700 times, promoted bilateral and multilateral initiatives on 80+ occasions, and organised 350+ seminars and conferences. In July 2019, Minister Dharmendra Pradhan and Dr Fatih Birol (IEA) launched the CEEW Centre for Energy Finance. In August 2020, Powering Livelihoods — a CEEW and Villgro initiative for rural start-ups — was launched by Minister Mr Piyush Goyal, Dr Rajiv Kumar (NITI Aayog), and H.E. Ms Damilola Ogunbiyi (SEforAll).

The Council's major contributions include: The 584-page National Water Resources Framework Study for India's 12th Five Year Plan; the first independent evaluation of the National Solar Mission; India's first report on global governance, submitted to the National Security Adviser; irrigation reform for Bihar; the birth of the Clean Energy Access Network; work for the PMO on accelerated targets for renewables, power sector reforms, environmental clearances, Swachh Bharat; pathbreaking work for the Paris Agreement, the HFC deal, the aviation emissions agreement, and international climate technology cooperation; the concept and strategy for the International Solar Alliance (ISA); the Common Risk Mitigation Mechanism (CRMM); critical minerals for Make in India; modelling uncertainties across 200+ scenarios for India's low-carbon pathways; India's largest multidimensional energy access survey (ACCESS); climate geoengineering governance; circular economy of water and waste; and the flagship event, Energy Horizons. It recently published *Jobs, Growth and Sustainability: A New Social Contract for India's Recovery.*

The Council's current initiatives include: A go-to-market programme for decentralised renewable energy-powered livelihood appliances; examining country-wide residential energy consumption patterns; raising consumer engagement on power issues; piloting business models for solar rooftop adoption; developing a renewable energy project performance dashboard; green hydrogen for industry decarbonisation; state-level modelling for energy and climate policy; reallocating water for faster economic growth; creating a democratic demand for clean air; raising consumer awareness on sustainable cooling; and supporting India's electric vehicle and battery ambitions. It also analyses the energy transition in emerging economies, including Indonesia, South Africa, Sri Lanka and Viet Nam.

The Council has a footprint in 21 Indian states, working extensively with state governments and grassroots NGOs. It is supporting power sector reforms in Uttar Pradesh and Tamil Nadu, scaling up solar-powered irrigation in Chhattisgarh, supporting climate action plans in Gujarat and Madhya Pradesh, evaluating community-based natural farming in Andhra Pradesh, examining crop residue burning in Punjab, and promoting solar rooftops in Delhi and Bihar.

Acknowledgments

The authors of this study would like to thank Bloomberg Philanthropies and Shakti Sustainable Energy Foundation (SSEF) for their support to carry out this study.

We would like to thank our reviewers—Prof. Vish Krishnan, Professor & Chair, University of California, Priya Shyamsundar, Lead Economist, The Nature Conservancy, Mohit Sharma, Associate Counsellor, CII-ITC Centre of Excellence for Sustainable Development, Udhaya Kumar V, Senior Research Analyst, Center for Study of Science, Technology and Policy Rishabh Jain, Manager, Market Intelligence, CEEW Centre for Energy Finance and Niti Gupta, Programme Associate, CEEW—for providing critical feedback and comments to refine this report. We extend out thanks to Karthik Ganesan, Research Fellow, CEEW, and Tanushree Ganguly, Programme Associate, CEEW, for providing us with valuable guidance throughout the study.

We sought out entrepreneurs in the supply chain— Monish Ahuja, Lt. Col. (Retd.), Promoter and Managing Director, Punjab Renewable Energy Systems Private Limited; Sukhmeet Singh, Co-Founder, A2P Energy Solution Private Limited; Nitin Gupta, Director, Farm2Energy; Aayushi Aggarwal, Co-Founder, RY Energies; and Naresh Kumar, Co-Founder, Shri Rishabh Travels—for information on supply chain management of crop residue in Punjab and they were gracious enough to share data and details we needed, for which we express our sincere gratitude.

Finally, we thank our Outreach team, particularly Alina Sen, Communications Specialist, for her guidance and support at the publication stage and her efforts to ensure that the report meets the CEEW's quality standards.

The authors



L. S. Kurinji
kurinji.selvaraj@ceew.in

A policy researcher at The Council, Kurinji's research focuses on the use of alternative methods of monitoring air quality and activities that cause pollution. She also plays an active role in providing tangible solutions to address regulatory challenges in implementing clean air policies. She holds a bachelor's degree in energy and environmental engineering from Tamil Nadu Agricultural University, Coimbatore. She is an Indian Green Building Council (IGBC) accredited professional.

"Though ex-situ crop residue management appears like an attractive proposition for Punjab's stubble burning crisis, the supply chain of the system is still an issue because of its economics. The sector needs optimisation. Through this report, we identify tangible solutions to support the supply chain, thereby to scale up ex-situ management in the state."



Sankalp Kumar
sankalp.vibhu@gmail.com

A research intern at The Council, Sankalp researches on ways to redress air pollution. He believes even a small step to solving the crisis of air pollution is, a giant leap towards creating sustainable societies. He is pursuing environmental engineering from the Delhi Technological University and prior to CEEW, he was a research intern at the Centre for Fire, Explosive and Environmental Safety (CFEES), a part of Defence Research and Development Organisation (DRDO).

"Crop residue burning has been an age-old problem vitiating air quality in the Indo-Gangetic plains; it is high time we look towards solutions to solve this crisis. Ex-situ management of residue offers a practical solution to the problem. This report will assist the stakeholders in scaling up the ex-situ management of crop residue in Punjab."



Since manual harvesting is labour- intensive and costs time and money, farmers opt for mechanised harvesting using the combine harvester. The machine cuts only the top portion of the crops leaving standing stubble in the field.

Contents

Executive summary xv

1. Motivation for the study 1

2. Status of crop residue management in Punjab 4

3. Approach and methodology 7

3.1 Cost computation 8

4. Results and discussion 10

4.1 Cost of biomass aggregation from farm to the straw bank 11
4.2 Delivered cost of residue from farm to end-user 13
4.3 Concerns regarding the use of paddy residue for ex-situ applications 19

5. Policy recommendations 21

5.1 Establish a dense network of straw banks and ex-situ supply chain ecosystem 22
5.2 Regulation for minimum crop residue usage 22
5.3 Price regulation of crop residue and its products 22
5.4 Mapping of end-users 23
5.5 Digital platform for farmers to trade their residue 23

6. Conclusion 24

References 27

Annexures 33



Image: iStock

Figures

Figure ES1: Raking and Baling cost is the highest in the aggregation of crop residue from farm to the straw bank xvii

Figure 1: Ludhiana and Sangrur districts generate the highest amount of paddy residue in Punjab (2019) 5

Figure 2: Raking and Baling cost is the highest in the transportation of crop residue from farm to the straw bank 11

Figure 3: Cutting and baling entail a high cost in the shifting of bale from farm to the straw bank 12

Figure 4: Cost of aggregation of biomass takes the largest slice of the delivered cost of per tonne bale from farm to end-user 14

Figure 5: Briquetting cost adds up significantly to the delivered cost of briquettes from farm to end-user 15

Figure 6: Delivered cost of per tonne of pellets from farm to end-user is highest among all processing options for crop residue 16

Figure 7: Delivered cost of biomass-based products increases with distance 17

Figure 8: Delivered cost of briquetted and pelleted biomass is lower than that of coal up to a certain distance between the farm and the end-user 19

Tables

Table ES1: Supply chain of crop residue for ex-situ management xvi

Table ES2: Delivered cost of biomass per tonne to the enduser located at a distance of 50 km xviii

Table 1: A few prominent ex-situ crop residue management companies in Punjab 6

Table 2: A few prominent ex-situ crop residue management companies in Punjab 8

Table 3: Data considered for computing cost to transport crop residue from farm to end-user 9

Table 4: Pathways for different versions of biomass 13

Table 5: Delivered cost of biomass to the end-user located at a distance of 50 km 14

Table 6: Timeline of an ex-situ crop residue operation 17

Table A1: Data used to compute the delivered cost of biomass 33

Table A2: Data on farm implements involved in cutting and baling the residue 34

Table A3: Freight rates for final transportation 34

Table A4: Economics of briquette production 34

Table A5: Economics of pellet production 35

Table A6: Formulae used in the study 36

Acronyms

A2P	Agri2Power
CEA	Central Electricity Authority of India
CHCs	custom hiring centres
CIL	Coal India Limited
CNG	compressed natural gas
COVID-19	disease caused by novel coronavirus, first detected in 2019
CSERC	Chhattisgarh State Electricity Regulatory Commission
eNAM	National Agriculture Market
EOI	expression of interest
INR	Indian rupee
IREDA	Indian Renewable Energy Development Agency Ltd
LCA	life-cycle assessment
MoES	Ministry of Earth Sciences
MoU	memorandum of understanding
NAAQS	National Ambient Air Quality Standards
NABARD	National Bank for Agriculture And Rural Development
NREBPL	Neway Renewable Energy Bathinda Private Limited
NRSE	New and Renewable Sources of Energy Policy
NTPC	National Thermal Power Corporation
PAU	Punjab Agricultural University
PEDA	Punjab Energy Development Agency
PM	particulate matter
PRESPL	Punjab Renewable Energy Systems Pvt. Ltd
PSCST	Punjab State Council on Science and Technology
PSPCL	Punjab State Power Corporation Limited
SAFAR	System of Air Quality and Weather Forecasting And Research
SCM	supply chain management



Collection of crop residue is a labour intensive task and is a major cost in the crop residue supply chain.

Image: A2P Energy

Executive summary

Agriculture and allied sectors contribute around 16.5 per cent to India's GDP and employs nearly half of the country's workforce (PRSIndia 2020). A massive amount of crop residue (~683 million tonnes) is generated during crop production in the net sown area of 140 million hectares across the country. While farmers use crop residue as animal fodder and for roof thatching, a significant portion (178 million tonnes) is left unused ever year (TIFAC and IARI 2018). Further, the unhealthy practice of on-farm burning of agricultural residue to clear land for the next crop, primarily in the north-western states of India, contributes to alarming levels of air pollution in the Indo-Gangetic plains.



While farmers use crop residue as animal fodder and for roof thatching, a significant portion (178 million tonnes) is left unused every year in India

Farmers in Punjab, where 20 million tonnes of paddy residue is generated every year during the Kharif season (Ministry of Agriculture & Farmers Welfare 2018), face an unenviable task of clearing this residue in a short window of 15–20 days. This reduced timeframe is an offshoot of the *Punjab Preservation of Subsoil Act (2009)*, implemented to save groundwater by mandatorily postponing the transplanting of paddy from April–May to beyond 10 June (Jain 2019). In 2018, 65 per cent of paddy residue (nearly 13 million tonnes) was set on fire in the fields of Punjab, choking the air in the entire Indo-Gangetic plains (Ministry of Agriculture & Farmers Welfare 2018). The System of Air Quality and Weather Forecasting And Research (SAFAR) under the Ministry of Earth Sciences (MoES) estimated that paddy stubble burning in Punjab and Haryana contributed 40–45 per cent to Delhi's air pollution during peak burning days in 2019 (Press Trust of India 2019).

The courts have come down heavily on stubble burning, forcing the state and central governments to initiate measures to clamp down this practice in Punjab, Haryana and Uttar Pradesh. One such effort was through the New and Renewable Sources of Energy (NRSE) policy 2019, wherein the Punjab government encourages setting up of biomass power generation units and production of biofuels (bio-compressed natural gas [CNG], bio-ethanol, and bio-diesel) using biomass (mainly rice straw) as feedstock. As of September 2020, Punjab has 11 operational biomass power plants, with an aggregate capacity of 97.5 MW, in which 0.88 million tonnes of paddy straw are consumed annually (Chaba 2020b). In 2018, the central government reported that 1.10 million tonnes of paddy residue (5.5 per cent of total residue generated) were used in various ex-situ methods such as in paper/cardboard mill and biomass power projects (Ministry of Agriculture & Farmers Welfare 2018).

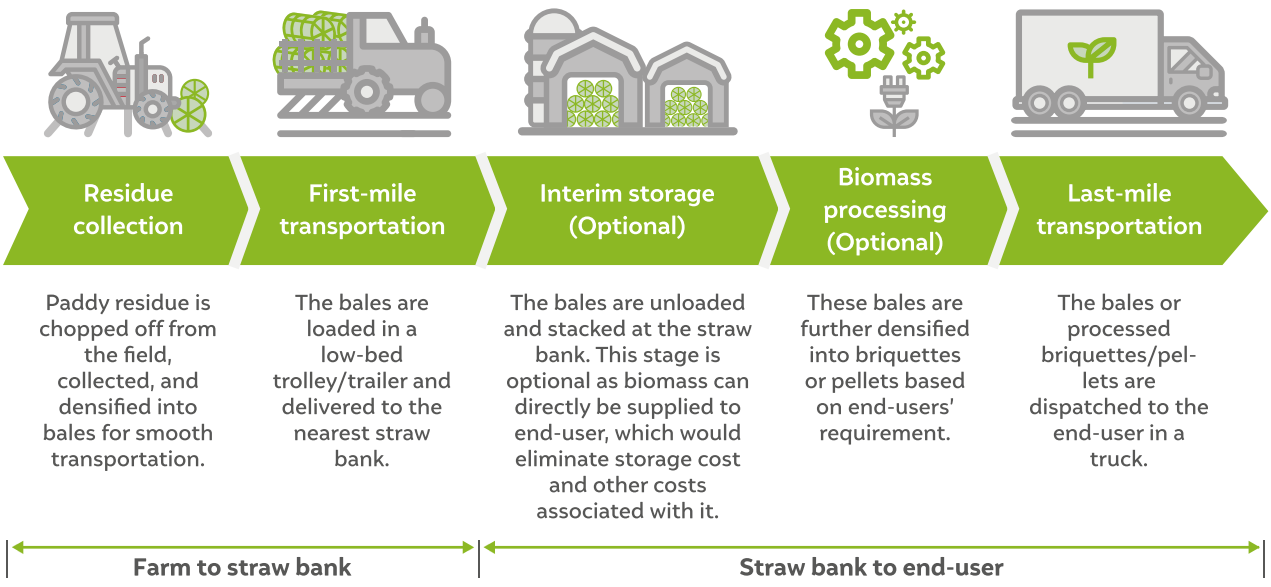
Ex-situ residue management methods such as biomass power plants and biofuel projects offer an attractive option of managing the excess crop residue generated. Therefore, to

ensure sustainable use of crop residue, Punjab government is inclined to sign memoranda of understanding (MoUs) with private players to set up more biomass-based projects in the state. An investment of INR 4.45 to 6 crore per MW is needed for biomass power plants (Central Electricity Regulatory Commission 2019). But due to the lack of assured biomass supply, the private sector is not finding it economically viable to invest in additional biomass plants. To ensure a reliable supply of biomass to the end-user, the Punjab government needs to create a dense network of collection centres for effective supply chain management (SCM).

Entities involved in SCM facilitate the fast clearing of residue from the field to reach the end-user. Given the voluminous nature and seasonal availability of the crop residue, its handling and on-time delivery becomes a central issue as it requires a vast workforce, heavy vehicles for logistics and extensive storage infrastructure. In addition, as crop residue has lower energy density,[1] it is subjected to densification processes such as baling, briquetting, and pelletisation, which adds to the feedstock supply cost (J. Singh, Panesar, and Sharma 2010). Therefore, entrepreneurs in the supply chain find the economics of handling crop residue unattractive.

Pushed by an interest to understand the economics involved in the supply of biomass to end-users in Punjab, we undertook this study. We interviewed prominent firms involved in ex-situ SCM namely Punjab Renewable Energy Systems Private Limited, A2P Energy Solution Private Limited, Farm2Energy, and RY Energies in Punjab to chart the various steps (Table ES 1) involved and computed the delivered cost of paddy residue from the farm to the end-user. From these interviews, we have gathered insights on the delivered cost of various types of biomass products like bales, briquettes, and pellets to end-users. We further investigate the viability of the use of paddy residue in coal-fired power plants to supplement the use of coal in the state.

Table ES1 Supply chain of crop residue for ex-situ management



Source: Authors' compilation

1. The density of loose paddy straw, collected directly from the field, can range from 13 to 18 kg m^{-3} in dry matter (dm). The density of round paddy straw bale with a 70-cm length and 50-cm diameter range from 60–90 kg m^{-3} dm. The density of rice straw briquettes with a 90-mm diameter and 7- to 15-mm thickness is 350–450 kg m^{-3} dm. The density of rice straw pellets with an 8-mm diameter and from 30 to 50 mm in height is 600–700 kg m^{-3} dm (Van Hung et al. 2020).

Key findings

We summarise our findings on the delivered cost of paddy residue, optimum distance, and different forms of crop residue used for transit in the following. To calculate the delivered cost of biomass for the end-user, we divided the supply chain into two parts: farm to straw bank and straw bank to end-user (Table ES1).

Farm to straw bank

For computing expenses incurred for residue transportation from farm to the straw bank, we consider the rental cost of machinery (such as chopper/cutter-cum-spreader, baler, and raker), labour, and fuel required for cutting and baling of paddy residue, loading and unloading cost of bales at farm and straw bank, and transportation charges for carrying bales from farm to straw bank (Table ES1).

- The transit of paddy residue in the form of bales for a distance of 15 km from farm to straw bank typically costs INR 1,330 per tonne of bale (Figure ES1). Out of this, the cost associated with the cutting and baling of residue entails 50 per cent of the procurement cost.
- In most cases, SCM entities offer support by providing customised farm implements like raker and baler, needed for residue collection, as individual farmers do not maintain an inventory of such implements needed for ex-situ management. A few entities make use of an 'on-demand' model to locally source the farm implements from custom hiring centres (CHCs) or large farmers and provide those implements 'on demand' to individual farmers.

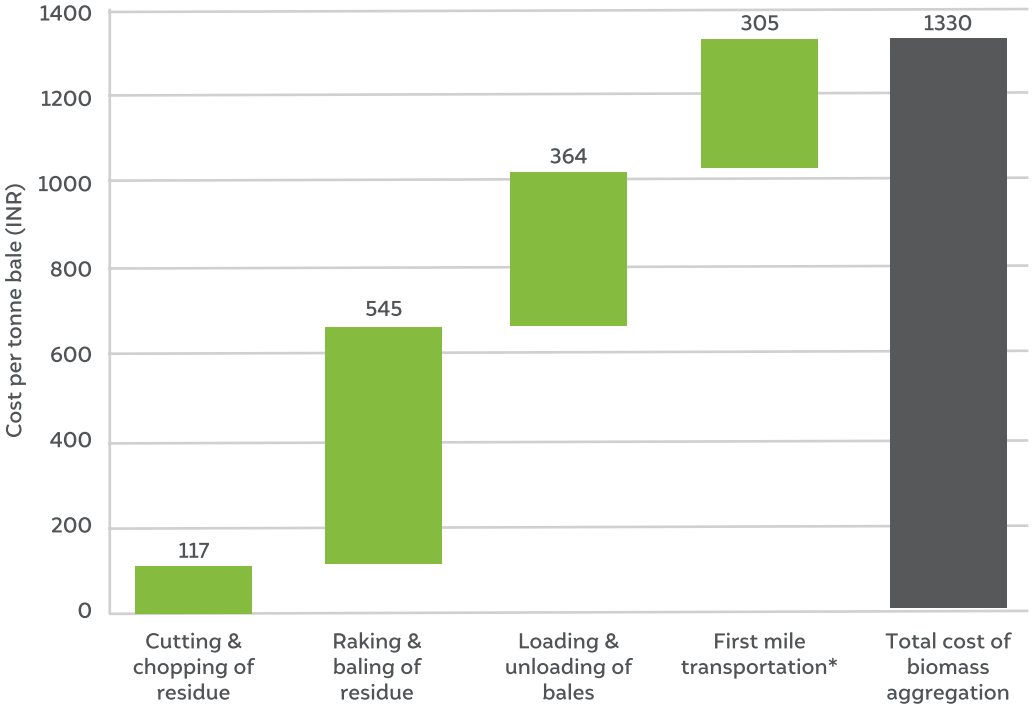


Figure ES1
Raking and baling cost is the highest in the aggregation of crop residue from farm to the straw bank

Source: Authors' analysis ;

*An optimum distance of 15 km between farm and straw bank is considered.

Straw bank to end-user

For costs incurred to transport farm residue from straw bank to end-user, we consider biomass conversion cost (briquetting and pelleting cost), loading and unloading cost at the straw bank and end-user, respectively, and transportation charges. We take into account three forms of biomass transported to the end-user — bales, briquettes and pellets. Both briquettes and pellets can be used in boilers and stoves. But considering their different sizes, briquettes are preferred in large and medium-scale boilers, while pellets are widely used in small-sized devices, such as pellet stove, furnace, and cooking range (The AGICO 2020). As briquette and pellets are preferred over bales because of their increased energy content in industries and power plants, we also estimate and compare the delivered cost per 5,000 Mcal of bales with briquettes and pellets.

Fragments	Components	Bale (INR)	Briquette (INR)*	Pellet (INR)*
Residue collection + first-mile transportation	Residue cost	1,351	1,757	1,757
Interim storage (optional)	Storage cost	880	1,144	1,144
Biomass processing (optional)	Biomass processing cost (briquette/pellet)	NA	1,034	1,437
Final transportation	Transport charges	347	250	250
	Loading and unloading cost	364	135	135
Total cost (INR/tonne)		**2,942**	**4,319**	**4,719**
Total cost (INR/5,000 Mcal)		**5,979**	**6,013**	**6,626**

Table ES2
Delivered cost of biomass per tonne to the end-user located at a distance of 50 km

Source: Authors' analysis

*About 1.3 tonne of residue is required to produce 1 tonne of briquette/ pellet. Additional cost of collection, transportation and storage of 0.3 tonnes of residue for every tonne of briquette and pellets has been added to briquettes and pellets resulting in the variation of collection, first mile transportation and storage cost between the three forms of biomass.

- Interim storage of the crop residue in straw banks assumes importance, given the limited span of 15–20 days available for sowing the next crop. A typical straw bank is spread over an area of 10–15 acres and can hold up to 10,000–15,000 tonnes of paddy residue. The storing cost constitutes 20–30 per cent of the total delivered cost of biomass.

- Transporting baled residue to the end-user beyond 50 km entails a high delivered cost, which may not be viable for companies engaged in the supply chain. Dispatching bales over a distance of 50 km incurs a total delivered cost of INR 2,950 per tonne of bale (Table ES2).

- Several studies have identified that densifying biomass in the form of briquettes or pellets would decrease the transportation cost (Balingbing et al. 2020; Possidônio et al. 2016). However, our analysis indicates that the high cost paid for briquetting and pelletising significantly eats into the benefits of reduced transportation cost of the densified residue.

- Dispatching briquettes over a distance of 50 km costs INR 4,320 per tonne delivered. But dispatching pellets over the same distance works out to a total delivered cost of INR 4,720 per tonne (Table ES2).

2. The two thermal power plants operating on coal in Punjab require 11.34 million tonnes of coal per annum (Government of India and Government of Punjab 2016)

- The high cost of pelleting makes transporting pelletised residue the costliest among all the options available. Supply chain entities concur with this observation. In most cases, SCM entities prefer to transport residue in the form of bales to avoid the high investment needed for biomass processing. However, biomass bales have limited end-use applications. They are used as packaging material, as raw material in cardboard and paper industries, in animal bedding, and in biochar and biogas production. For other end uses that require densified biomass, the end-user needs to invest in processing bales to briquettes or pellets.

- Co-firing 10 per cent of biomass pellets with coal in coal-fired power plants in Punjab[2] would require 1.47 million tonnes of paddy residue annually, which would account for nearly 7.4 per cent of paddy residue generated in Punjab. The delivered cost of biomass pellets per 5,000 Mcal within a radius of 50 km of the power plant is economical compared to coal. Beyond 50 km, the cost increases significantly.

- The delivered cost of briquettes per 5,000 Mcal is economical up to a distance of 145 km. Industries that consume coal for heating applications should explore the prospect of locally sourcing biomass briquettes within the state, as it will reduce their operating cost.

Key recommendations to support supply chain management of crop residue

Based on our assessment and consultations with prominent stakeholders, we recommend the following strategies for better management and scaling up the ex-situ use of crop residue.

- **Establish a dense network of straw banks and ex-situ supply chain ecosystem**

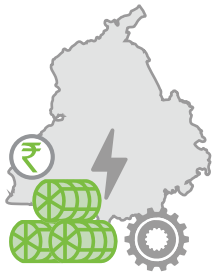
  Sourcing of crop residue for ex-situ use, in our estimates, is found to be financially viable for SCM entities when the crop residue collection point is located within 15 km and end-user is situated within 50 km. Hence, a dense network of straw banks needs to be created for streamlining the supply chain. Private sector should get adequate support for the high cost of investment in SCM. Agencies such as National Bank for Agriculture And Rural Development (NABARD) and Indian Renewable Energy Development Agency Ltd (IREDA) can be roped in to offer loans and insurance for entities involved in SCM. Punjab Energy Development Agency (PEDA), in partnership with private entities, should also implement pilot programmes to scale up participation in SCM.

- **Boost demand for biomass**

  Participation of the private sector on a larger scale in ex-situ supply chain operations of crop residue is the need of the hour. The Central Electricity Authority of India (CEA) recommends a blend of 5–10 per cent of biomass pellets along with coal in fluidised bed and pulverised coal units. Mandating a similar blending ratio and technical guidelines for residue usage in other industries would create significant demand for crop residue.

- **Regulate prices of crop residue and its products**

  In December 2018, National Thermal Power Corporation (NTPC) invited expression of interest (EOI) for supplying non-torrefied biomass-based pellets for 21 power plants, none



The Punjab State Electricity Regulatory Commission (PSERC) should work with PEDA to establish price regulation for biomass-based products

of which are in Punjab (NTPC Limited 2018). In our analysis, the delivery of pellets over a distance of 200 km costs INR 5,500 per tonne. Delivering biomass from Punjab to these NTPC plants would push up the delivered cost of biomass pellets even higher, making it unprofitable for SCM entities to enter into a long-term fuel supply agreement with those power plants. Every year, the Chhattisgarh State Electricity Regulatory Commission (CSERC) fixes the minimum support price for rice husk, which is one of the largest biomasses produced in the state. Unprofitability of supplying residue is one of the reasons SCM entities are not willing to enter into a long-term fuel supply agreement with power plants. Similarly, the Punjab State Electricity Regulatory Commission (PSERC) should work with PEDA to establish price regulation for biomass-based products, making adequate allowance for compensating farmers for costs incurred in managing the waste.



Punjab Energy Development Agency (PEDA) needs to create a database of end-users mapped to their annual crop residue demand

- **Map end-users**

  From the analysis, it is evident that the delivered cost rises with an increase in distance and is very high for a longer distance between the farm and the end-user. If a provision for decentralised sourcing of crop residue within a radius of 5–10 km exists for the end-user, interim storage is not needed and transportation cost comes down significantly. We recommend that the Punjab Energy Development Agency (PEDA) create a database of end-users mapped to their annual crop residue demand. This database would help the SCM entities to optimally plan for storage and logistics. In addition, the areas where there is a deficit in demand or potential end-users the state government can focus their in-situ management resources in that particular areas. As with increase in distance ex-situ becomes less economical and the policymakers should prioritize deployment of in-situ farm implements like Happy seeders in these locations.

- **Create a digital platform to trade crop residue**

  A digital platform, similar to National Agriculture Market (eNAM), connecting farmers, straw banks, end-users, and various stakeholders involved would go a long way in streamlining the supply chain of crop residue. The digital platform should facilitate farmers to raise a request to collect crop residue, and the nearest straw bank could collect the residue from the farm within the stipulated time. For streamlining demand, we recommend that the database mapping individual annual biomass demand to each end-user can be fed into the platform. This would allow the SCM entities to arrange for decentralised sourcing which condense the interim storage and transportation charge significantly.

1. Motivation for the study


Image: iStock

Air pollution was the fourth leading risk for premature death globally and responsible for 12.5 per cent of the total deaths in India in 2017 (India State-Level Disease Burden Initiative Air Pollution Collaborators 2020). More than three-fourths of Indians were exposed to particulate matter ($PM_{2.5}$) levels exceeding National Ambient Air Quality Standards (NAAQS) in 2017 (India State-Level Disease Burden Initiative Air Pollution Collaborators 2020). Several factors contribute to air pollution in the country such as emissions from vehicles, industries, and thermal power plants, agricultural residue and waste burning, construction dust, and use of cheap fuels in the households (Bernard and Kazim 2018).

Stubble burning on farm fields, a common practice in the north-western states of India, is one of the prime contributors to the build-up of aerosols in the Indo-Gangetic plains (Jethva et al. 2019). Residents living in areas around intense stubble burning were found to have a three-fold higher risk of developing an acute respiratory infection (Chakrabarti et al. 2019). The System of Air Quality and Weather Forecasting And Research (SAFAR) under the Ministry of Earth Sciences (MoES) estimates that stubble burning in Punjab and Haryana contributes 40–45 per cent to Delhi's air pollution during peak burning days (Press Trust of India 2019).

Every year, paddy is sown in 31 lakh hectares in Punjab, which generates 20 million tonnes of paddy residue. In 2018, 13 million tonnes (65 per cent of the generated residue) of crop residue were set on fire in the fields of Punjab, which caused a worsening of air quality in the Indo-Gangetic plains (Ministry of Agriculture & Farmers Welfare 2018).



It was estimated that Punjab would require 35,000 Happy Seeders to manage the 31 lakh hectares of land under paddy cultivation in 2019

Farmers are forced to burn crop residue in the north-western states of India as massive amount of paddy comes under cultivation and cropping cycle is shortened to less than 20 days (Gupta 2019; Kurinji 2019). The reduced cropping cycle is an offshoot of the *Punjab Preservation of Subsoil Act (2009)*, implemented to conserve groundwater by mandatorily delaying the transplanting of paddy to beyond 10 June (Jain 2019). Before the Act was conceived, farmers cultivated/sowed paddy during April–May every year. A majority of the farmers burn stubble after the kharif harvesting period as they have to clear the field for rabi sowing period, which starts just days after the harvest.

The courts have taken a serious note of stubble burning and have directed the central and state governments to initiate suitable measures to clamp down this practice in Punjab, Haryana and Uttar Pradesh. Happy Seeder, a tractor-operated machine developed by Punjab Agricultural University (PAU) in 2002, is one of the commonly recommended solutions for in-situ management[3] of paddy residue (Goyal 2019). In response to this very ticklish issue, in 2018, the central government launched a scheme titled Promotion of Agricultural Mechanization for In-situ Management of Crop Residue in the States of Punjab, Haryana, Uttar Pradesh and NCT of Delhi with an outlay of INR 1,151.80 crore (INR 591.65 crore for 2018–19 and INR 560.15 crore for 2019–20) (Ministry of Agriculture & Farmers Welfare 2018). The major thrust of this scheme is to provide subsidies for the purchase of several farm equipment, such as Happy Seeders, to ensure adoption of such equipment for crop residue management across Punjab, Haryana and (western) Uttar Pradesh.

It was estimated that Punjab would require 35,000 Happy Seeders to manage the 31 lakh hectares of land under paddy cultivation in 2019 (N. Gupta 2019). But only 13,700 Happy Seeders were deployed in the state in 2019 (Nirmal 2019; Bhattacharya 2018), well short of the estimated requirement. Surveys indicated that Happy Seeders can save farmers about INR 1,000 per hectare on average as opposed to using conventional methods (Gupta and Somanathan 2016; Gupta 2019). Despite the benefits and government subsidies, farmers state high upfront cost needed to purchase the Happy Seeder[4] and limited role of the implement for just 20 days and idle for the rest of the year as a hindrance for the limited uptake of in-situ methods (A. Kumar 2019). Apprehensions about the yields of wheat and input costs, difficulty in operating the equipment under excess soil moisture due to monsoon rains, and possibility of pest attack further result in poor acceptance of the Happy Seeder by farmers (Kurinji 2019; Chaba 2020a).

Ploughing back farm waste in situ yields many benefits such as enhancing soil health and improving water retention capacity. But the sheer scale of the residue generated makes it impossible to plough back all of it back into the soil (CII and NITI Aayog 2018). Globally, removing excess residue and using as feedstock for energy purposes proved to increase farmer income (Suardi et al. 2019) . Various existing and emerging technologies such as pyrolysis (biochar), bio-methanation (biogas), and conversion to biofuels (such as briquettes,

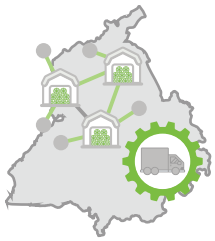
3. In-situ management of crop residue involves managing residue in the same field where it gets generated through process such as mulching and incorporation.

4. The average maximum price of a Happy Seeder is INR 1,51,200. The average maximum permissible subsidy per Happy Seeder per beneficiary is INR 75,600, which is 50 per cent of the total cost (Ministry of Agriculture & Farmers Welfare 2018).

pellets, bio-compressed natural gas [CNG], and bio-diesel) have been recommended for ex-situ use of paddy crop residue (CII and NITI Aayog 2018). The National Thermal Power Corporation Limited (NTPC) has demonstrated the use of crop residue in power generation by co-firing seven per cent blend of biomass pellets with coal in Dadri Power Plant, Uttar Pradesh (Central Electricity Authority 2017). However, lack of an assured supply of biomass in adequate quantities proves to be a dampener for private firms to set up biomass plants as they find it economically unviable to invest INR 4.45 to 6 crore per MW (Central Electricity Regulatory Commission 2019).

Given the voluminous nature and seasonal availability of the crop residue, its handling and on-time delivery become a central issue, as it requires a vast workforce, heavy vehicles for logistics and extensive storage infrastructure. The consumption of 13 million tonnes (that was burnt in 2018) of crop residue and a reliable supply of biomass to the end-user requires a dense network of collection centres and supply chain management (SCM) facilities in Punjab. But the high cost of collection and transportation of residue from the field to end-user proves to be the prime impediment for scaling up its ex-situ management (J. Singh, Panesar, and Sharma 2010).

We carried out this study to improve the current understanding of the economics involved in the supply of biomass to end-user in Punjab. In the course of our work, we have gathered insights on the delivered cost of various types of biomass products such as bales, briquettes, and pellets to the end-users. We further probe the delivered cost of crop residue or its derivatives, which can be used as a supplement to coal in power generation.



A reliable supply of biomass to the end-user requires a dense network of collection centres and supply chain management (SCM) facilities in Punjab



Farmers use balers to densify biomass into bales for better handling and smoother transportation.

Image: A2P Energy

2. Status of crop residue management in Punjab

In 2018, about 31.03 lakh hectares of land in Punjab was under paddy cultivation, producing 13 million tonnes of rice as well as generating 20 million tonnes of paddy straw (Government of Punjab 2019). Districts such as Amritsar, Bhatinda, Ferozepur, Ludhiana, Moga, Patiala, and Tarn Taran generated more than 1 million tonnes of residue (Figure 1).

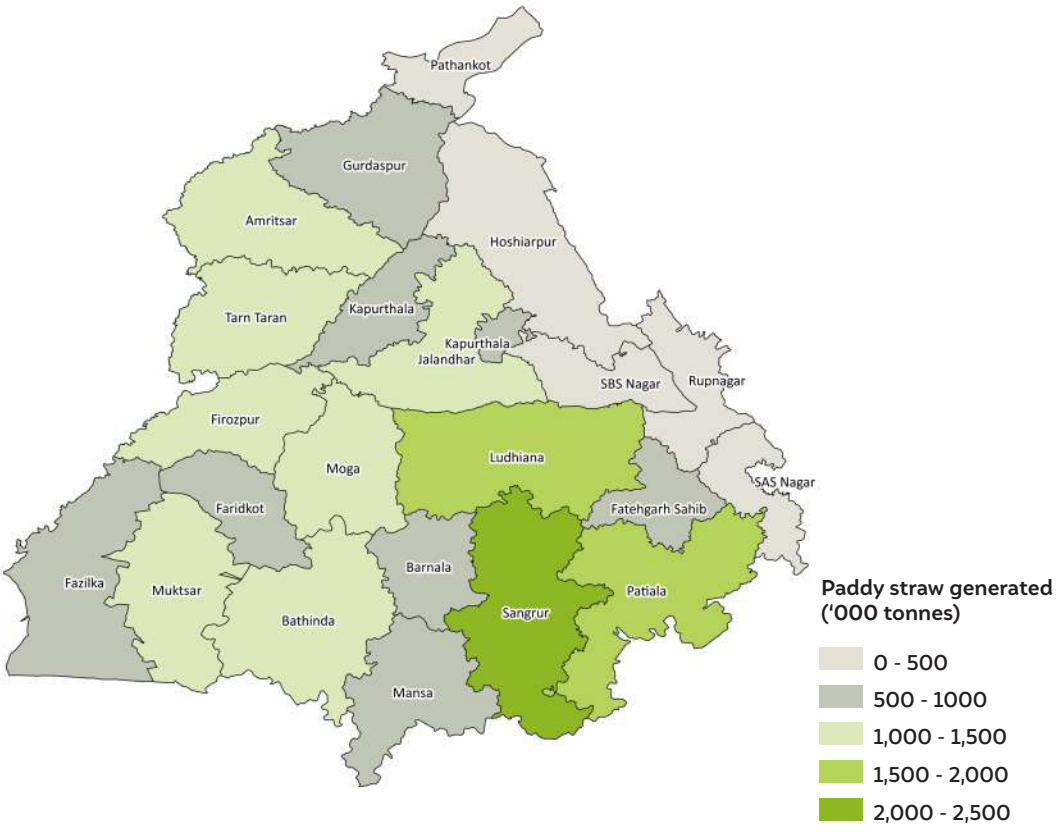


Figure 1
Ludhiana and Sangrur districts generate the highest amount of paddy residue in Punjab (2019)

Source: Authors' compilation

The government's review report claims that 3.52 million tonnes of residue were managed using in-situ methods in the state in 2018 (Ministry of Agriculture & Farmers Welfare 2018). But about 13 million tonnes were burnt on the field in 2018, as a result of which 55,000 incidents of fire were recorded (Kurinji 2019). About 2.7 million tonnes of paddy residue produced in 5 lakh hectares of basmati cultivation was utilised as animal fodder (Ministry

of Agriculture & Farmers Welfare 2018). Information gathered from prominent stakeholders involved in ex-situ management shows that 1.1 million tonnes of paddy residue (Ministry of Agriculture & Farmers Welfare 2018) were utilised for ex-situ applications such as in biomass power plants and paper and cardboard industries. Table 1 provides a list of few prominent companies involved in the SCM of crop residue in Punjab, some of which are in construction stage.

As of 2020, Punjab has 11 biomass power plants, with an aggregate capacity of 97.50 MW, consuming 8.80 lakh million tonnes of paddy residue annually. Two more biomass power projects with 14 MW capacity (consuming 1.26 lakh metric tonnes per annum) are slated to be commissioned in June 2021. Other than biomass projects, eight bio-CNG projects are being executed in the state. These can use around 3 lakh metric tonnes of paddy residue annually. A bio-ethanol plant located in Bhatinda with a capacity of 100 kilolitres per day, which is under execution, will utilise 2 lakh metric tonnes of paddy stubble annually. When all these projects are commissioned, 1.5 million tonnes of paddy residue would be consumed by them annually ex-situ, which constitutes 7 per cent of the total residue generated in the state (Chaba 2020b).

Company	Products	Residue managed
Agri2Power (A2P) and Energy Harvest, Chandigarh	- Biomass supply chain management - Pellet	2,450 tonne/annum
Punjab State Council on Science and Technology (PSCST), Chandigarh	- Briquettes	4,000 tonne/annum
Punjab Renewable Energy Systems Pvt. Ltd. (PRESPL)	- Pellets - Briquettes - Biomass supply chain management	62,500 tonne/annum
Neway Renewable Energy Bathinda Private Limited (NREBPL), Mohali	- Bio coal	7,35,000 tonne/annum
CH4 Biotech, Fazilka, Punjab	- Biogas	9,800 tonne/annum
Bioendev AB, Mohali, Punjab	- Torrified biomass pellets	700 tonnes/annum
Farm2energy, Bija, Punjab	- Biomass supply chain management - Bio coal	61,250 tonne/annum
RY Energies	- Biomass supply chain management	1,500 tonne/annum
Sukhbir Agro Energy Ltd	- Biomass energy generation	5,80,000 tonne/annum
Verbio India Pvt Ltd*	- Bio-CNG - Organic manure	90,000 tonne/annum

Table 1
A few prominent ex-situ management companies in Punjab

Source: Authors' compilation

Note: Plant in the development stage

3. Approach and methodology

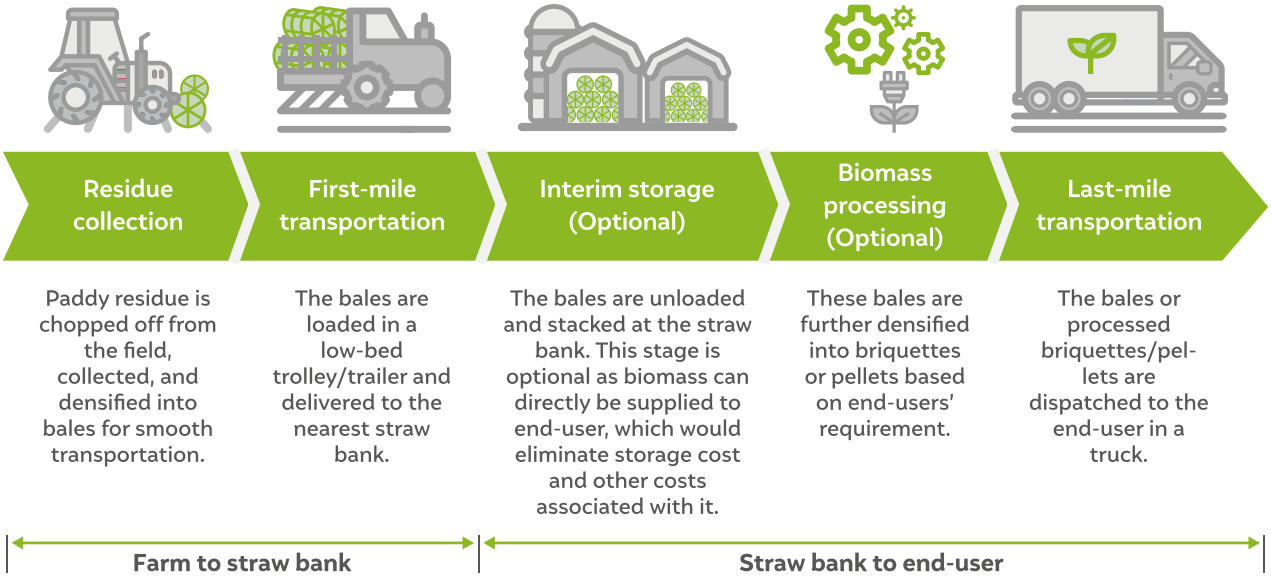

Image: iStock

We estimate the cost incurred in each stage of biomass supply to study the economics of mobilising paddy crop residue for energy conversion. We collected primary data through direct interaction with prominent stakeholders involved in crop residue supply chain management (SCM) in Punjab. Besides primary data from stakeholders, we also used the secondary data from various sources such as government reports, newspaper articles, official government websites, and peer-reviewed journals to estimate the cost involved in the processing of crop residue.

3.1 Cost computation

The supply chain of crop residue involves several steps such as residue collection from the field, first-mile transportation to collection centre/straw bank, interim storage at the straw bank, processing of biomass, and final transportation to the end-user. These steps are illustrated in Table 2.

Table 2 Supply chain of crop residue for ex-situ management



Source: Authors' compilation

To calculate the delivered cost of biomass for the end-user, we divided the supply chain into two parts: farm to straw bank and straw bank to end-user (Table 3).

Farm to straw bank

For computing expenses incurred for residue transportation from farm to the straw bank, we consider the rental cost of machinery (such as chopper/cutter-cum-spreader, baler, and raker), labour, and fuel required for cutting and baling of paddy residue, loading and unloading cost of bales at farm and straw bank, and transportation charges for carrying bales from farm to straw bank (Table 3). Our discussion with stakeholders indicates that the collection radius of bales should not exceed 15 km from the straw bank (optimum distance) as the transport charges become uneconomical beyond 15km. Hence, we consider the average cost of sourcing bales from the farm to the straw bank in the 5–15 km radius for the final delivered cost of biomass from straw bank to the end-user.

Straw bank to end-user

For costs incurred to transport farm residue from straw bank to end-user, we consider biomass conversion cost (briquetting and pelleting cost), loading and unloading cost at the

straw bank and end-user, respectively, and transportation charges. Truck rental charges include the labour and fuel charge, and it varies with the distance travelled (Table 3). Annexure A1 provides the details on the data used and their source.

Fragments	Steps involved	Farm implements/ facilities needed	Truck
Farm to straw bank	Residue collection	Chopper cum spreader	Rental cost: INR 1,000–1,200/8 hours in a day Field capacity: 3–4 acre/hour
		Raker + baler	Rental cost: INR 1,500–2,000/acre (rental cost includes fuel and labour)
	First-mile transportation	Tractor-trailer	Rental cost: INR 245/hour
Straw bank to end-user	Interim storage (optional)	Unloading machinery (optional)	Rental cost: INR 750–800/hour
		Storage Cost	Total capital and operational cost: INR 880/tonne
	Biomass processing (optional)	Cutter and grinder	Total operational cost: INR 587/tonne
		Briquette/pellet press	Total operational cost: INR 418/tonne (briquette) INR 818/tonne (pellet)
	Final transportation	Truck	Transport cost (Includes fuel cost and driver charges) (INR/km): Transport cost (0–50 km) 40 Transport cost (51–150 km) 39 Transport cost (151–300 km) 38 Transport cost (301–500 km) 37 Transport cost (501–1,000 km) 36

Table 3
Data considered for computing cost to transport crop residue from farm to end-user

Source: Authors' compilation; Based on the discussion with prominent ex-situ crop residue management entities in Punjab; (M. Sharma et al. 2019)



Heavy vehicles for logistics and a vast workforce are needed for handling and on-time delivery of crop residue to end-users.

Image: iStock

4. Results and discussion

In this chapter, we pay attention to metrics involved in the supply chain of residue, including the delivered cost of paddy residue, optimum distance, and form for residue preferred for transit, and the challenges plaguing the adoption of ex-situ option.

4.1 Cost of biomass aggregation from farm to the straw bank

The collection cost of crop residue varies depending on the distance between the straw bank and the farm. A straw bank should be ideally located within 15 km from the farm. Crop residue is generally transported in the form of bales as it makes handling and transportation easier and facilitates safe storage (A. Sharma and Chandel 2016). As per our calculations, the transit of paddy residue in the form of bales over a distance of 5–15 km typically costs INR 1,150-1,330 per tonne (Figure 2). The cost of cutting and baling residue accounts for 55 per cent of the sourcing cost of residue from the farm to a straw bank located at a distance of 15 km (Figure 3). The major component of this processing cost is the rental cost of several farm machinery involved such as chopper, raker, and baler (details shown in Annexure A2).

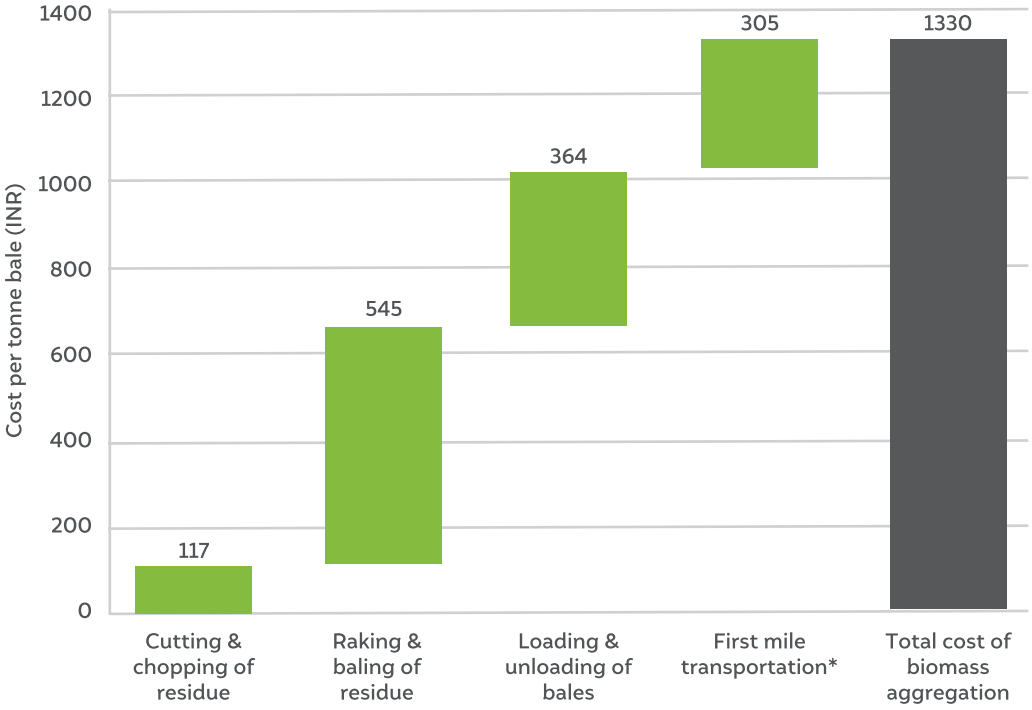


Figure 2
Raking and baling cost is the highest in the transportation of crop residue from farm to the straw bank

Source: Authors' analysis;

*An optimum distance of 15 km between farm and straw bank is considered.

In most cases, supply chain entities provide customised farm implements needed for residue collection, as evidently a majority of individual farmers do not maintain a repository of ex-situ farm implements, as heavy investment is needed for buying them (Confederation of Indian Industry; German Agribusiness Alliance 2020)(Rao 2019; Moudgil 2020). A few entities use 'on-demand' model to locally source the farm implements from custom hiring centres (CHCs) or large farmers and provide those implements 'on-demand' to individual farmers. SCM entities usually bear all the cost incurred in residue collection and transportation of residue.

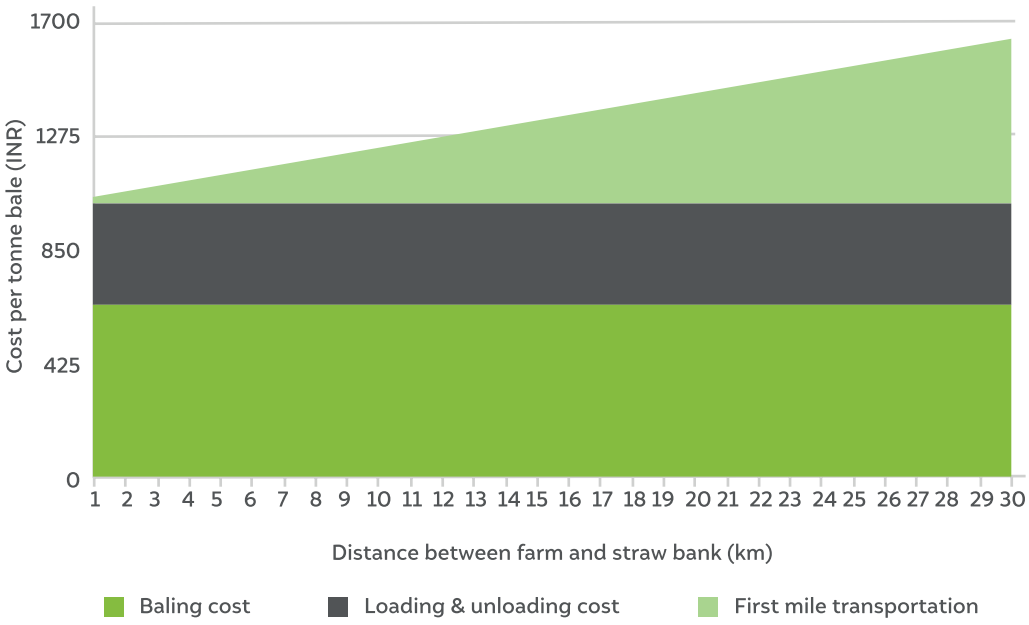


Figure 3
Cutting and baling entail a high cost in the shifting of bale from farm to the straw bank

Source: Authors' analysis

Role of straw banks

Farmers in Punjab have a limited span of 15-20 days available for preparing the land for the next crop. This makes the collection process of paddy residue onerous, as it gives a small window for chopping, balling, and transfer of paddy residue from farm to the straw bank. So interim storage in the form of straw banks/ collection centres allows more time for the processing and transport of residue to the end-user. A straw bank, spanning over an area of 10–15 acres, can hold up to 10,000–15,000 tonnes of residue.[5] This means 850–1,300 straw banks are needed, spread over an area of 13,000 acres (5,260 hectares) in Punjab, for preparing 13 million tonnes[6] of residue for ex-situ applications.

SCM entities usually lease land, typically for three to six months, to set up a straw bank for a rental cost of INR 50,000–1,00,000 per hectare, depending on the type of land. Generally, straw banks should be located on common panchayat land or other fallow lands available, as rental charges would be lower for them compared to farmlands. Straw banks should be equipped with safeguards against fire hazards as stored residue often catches fire due to various factors such as increased ambient temperature, chafing of bales with each other, and human-induced factors, resulting in considerable losses. Maintaining a safe distance between piles may reduce the risks of fire. Further, prolonged storage can also result in loss of

5. Small SCM entities may also set up straw banks spaning over an area of less than 10 acres.

6. We use the example of 13 million tonnes, as it is the amount of paddy residue that was burnt on the farm in 2018.

dry and organic matter, substantial reductions in heating value, and volatile matter (Blunk et al. 2003).

4.2 Delivered cost of residue from farm to end-user

SCM entities generally densify the biomass to increase the bulk and energy density of raw biomass for easy storage and handling and improve transport efficiency and cost (Ray et al. 2017). Baling, briquetting, and pelleting are the methods used to densify biomass based on the end-users' need.[7] To compute the delivered cost of residue from farm to end-user, we take into account three forms of biomass transported to the end-user as shown in Table 4.

Pathway I	Pathway II	Pathway III
Transit of residue in the form of bales from straw bank to end-user	The residue is converted into briquettes in the straw bank and then dispatched to the end-user	The residue is converted into pellets in the straw bank and then dispatched to the end-user
End-users of bales: Biomass power plants, packaging industry, and paper and cardboard industry	End-users of briquettes: Large and medium-scale industrial boilers and other heating requirements	End-users of pellets: coal power plants, smaller devices such as pellet stove and furnace

Table 4
Pathways for different versions of biomass

Source: Authors' analysis

For scenarios II and III, we assume that decentralised briquetting and pelletisation plants are available at the straw bank. We calculate the total operational cost of these plants at INR 1,034 per tonne of briquettes and INR 1,433 per tonne of pellets, respectively. Tables A4 and A5 respectively in the Annexure provide detailed estimates on the economics of briquetting and pelletisation plants. The cost break-up of per tonne of biomass supplied to the end-user located at a distance of 50 km is shown in Table 5.

Biomass pellets occupy a larger contact area with air compared to briquettes. Hence, in well-ventilated conditions, biomass pellets are easier to burn, faster to transfer heat, and burn more sufficiently. But biomass briquettes are superior in terms of burning time. Both briquettes and pellets can be used in boilers and stoves. But considering their different sizes, briquettes are preferred in large and medium-scale boilers, while pellets are widely used in small-sized devices, such as pellet stove, furnace, and cooking range (The AGICO 2020). As briquette and pellets are preferred over bales because of their increased energy content in industries and power plants, we also estimate and compare the delivered cost per 5,000 Mcal of bales with briquettes and pellets.

7. The density of loose paddy straw, collected directly from the field, can range from 13 to 18 kg m^{-3} in dry matter (dm). The density of round paddy straw bale with a length of 70 cm and a diameter of 50 cm range from 60 to 90 kg m^{-3} dm. The density of paddy straw briquettes with a 90-mm diameter and 7- to 15-mm thickness is 350–450 kg m^{-3} dm. The density of rice straw pellets with an 8-mm diameter and from 30 to 50 mm in height is 600–700 kg m^{-3} dm. (Van Hung et al. 2020).

Fragments	Components	Bale (INR)	Briquette* (INR)	Pellet* (INR)
Residue collection + first-mile transportation	Residue cost	1,351	1,757	1,757
Interim storage (optional)	Storage cost	880	1,144	1,144
Biomass processing (optional)	Biomass processing cost (briquette/pellet)	NA	1,034	1,434
Final transportation	Transport charges	347	250	250
	Loading and unloading cost	364	135	135
Total cost (INR/tonne)		**2,942**	**4,319**	**4,720**
Total cost (INR/5,000 Mcal)		**5,979**	**6,012**	**6,626**

Table 5
Delivered cost of biomass to the end-user located at a distance of 50 km

Source: Authors' analysis

*About 1.3 tonne of residue is required to produce 1 tonne of briquette/ pellet. Additional cost of collection, transportation and storage of 0.3 tonnes of residue for every tonne of briquette and pellets has been added to briquettes and pellets resulting in the variation of collection, first mile transportation and storage cost between the three forms of biomass.

Scenario I: Transit in the form of bales

The delivered cost of bales for a distance of 50 km between the straw bank and the end-user is INR 2,940 per tonne of bale. Storage charges account for 30 per cent of the total delivered cost per tonne. Transportation charge per tonne is INR 350 for a distance of 50 km between the straw bank and the end-user, which will go up as the distance increases (Figure 4). Conversations with crop residue SCM entities reveal that transport of biomass is profitable up to a distance of 50 km, beyond which it gets expensive. However, in the case of end-users located far away from the farms, the residue can also be dispatched beyond 100 km.

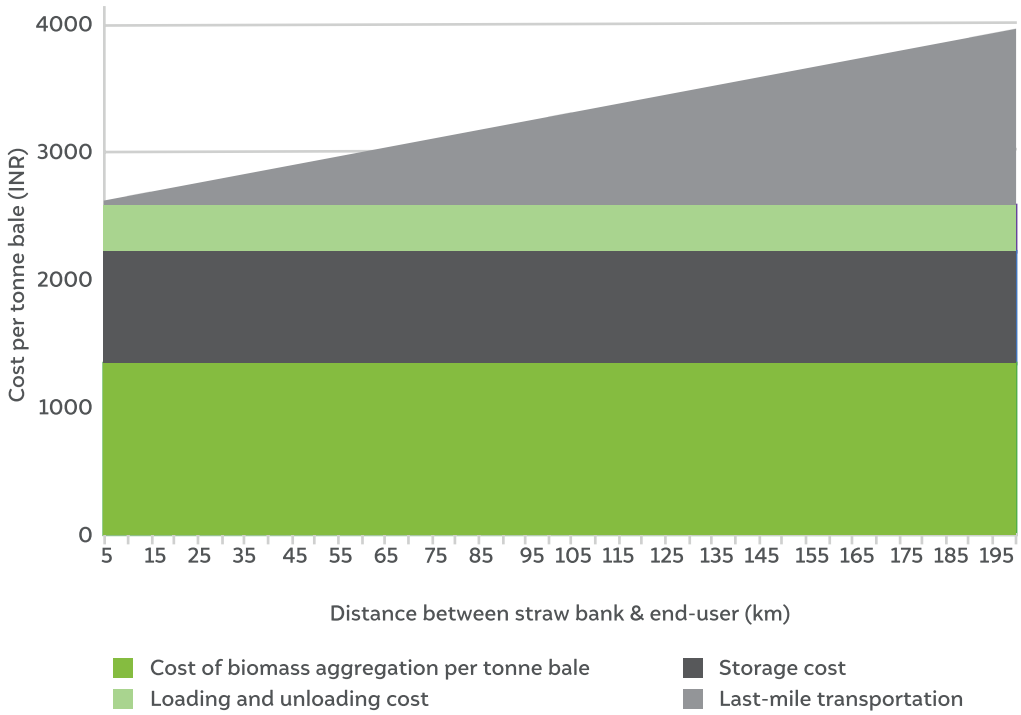


Figure 4
Cost of aggregation of biomass takes the largest slice of the delivered cost of per tonne bale from farm to end-user

Source: Authors' analysis

Scenario II: Transit in the form of briquettes

Several field studies demonstrate that densifying biomass reduces transportation cost by 60 per cent and the required quantity of trucks by 63 per cent (Balingbing et al. 2020; Possidônio et al. 2016). The other benefits are that densifying also increases energy density and significantly reduces handling difficulties (Clarke and Preto 2011). But, in our analysis, the cost paid for the operation of briquette plants largely offsets the cost benefits of the densification process. But the higher processing cost incurred is subsequently passed on to the end-user.

We estimate a delivered cost of INR 4,320 and INR 4,570 per tonne briquettes for a distance of 50 km and 100 km, respectively (Figure 5). For a distance of 150 km, the delivered cost of briquettes is INR 4,820 per tonne, 1.3 times higher than the delivered cost per tonne of bales for the same distance. However, the delivered cost of briquettes per 5,000 Mcal of energy for a distance of 150 is INR 6,340, 4 per cent lower than that of bales.

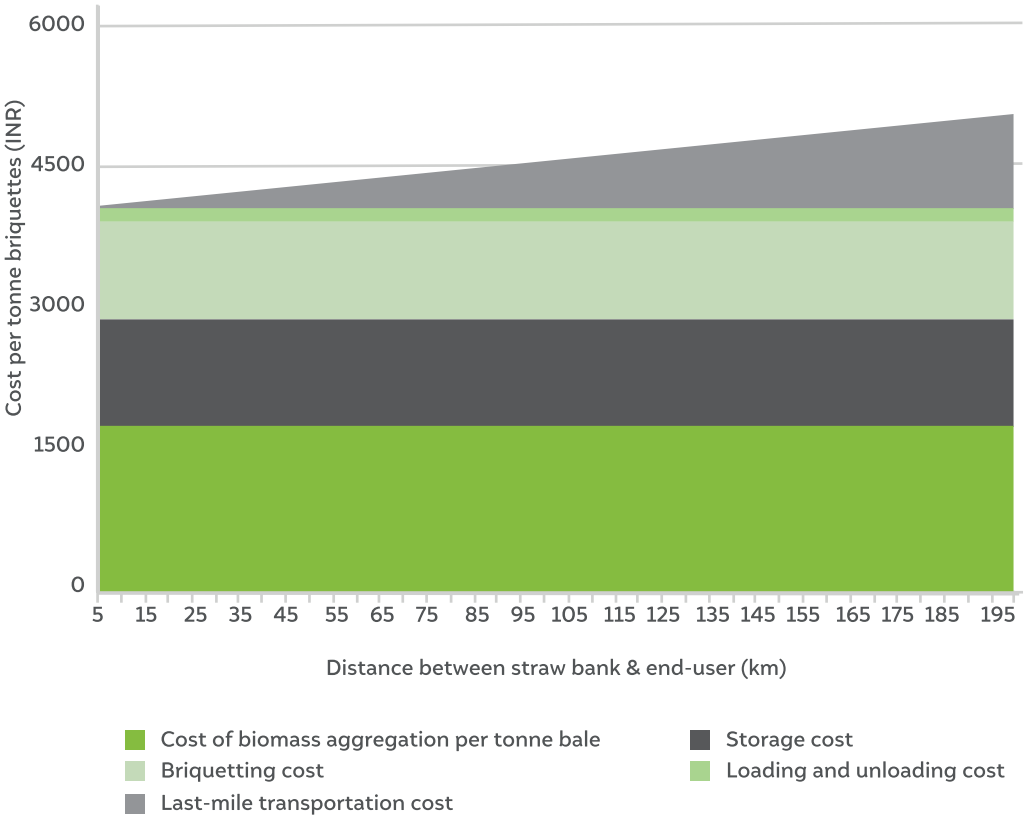


Figure 5
Briquetting cost adds up to significantly to the delivered cost of briquettes from farm to end-user

Source: Authors' analysis

Scenario III: Transit in the form of pellets

Compared with other densification processes, such as baling and briquetting, pelleting makes biomass most durable as the crop residue is subjected to the highest amount of pressure (Whittaker and Shield 2017). Pelletisation can increase the bulk density of the biomass from an initial value of 40–200 kg m^{-3} to a final bulk density of 600–700 kg m^{-3} (Balingbing et al. 2020; Van Hung et al. 2020).

We find a delivered cost of INR 4,720 and INR 4,970 per tonne of pellets for a distance of 50 km and 100 km, respectively (Figure 6). For a distance of 150 km, the delivered cost of pellets is INR 5,220 per tonne, 1.44 times higher than the delivered cost per tonne of bales for the same distance. Further, the delivered cost of pellets per 5,000 Mcal of energy for a distance of 150 km shoots up to INR 7,000, a 5.3 per cent higher cost than bales.

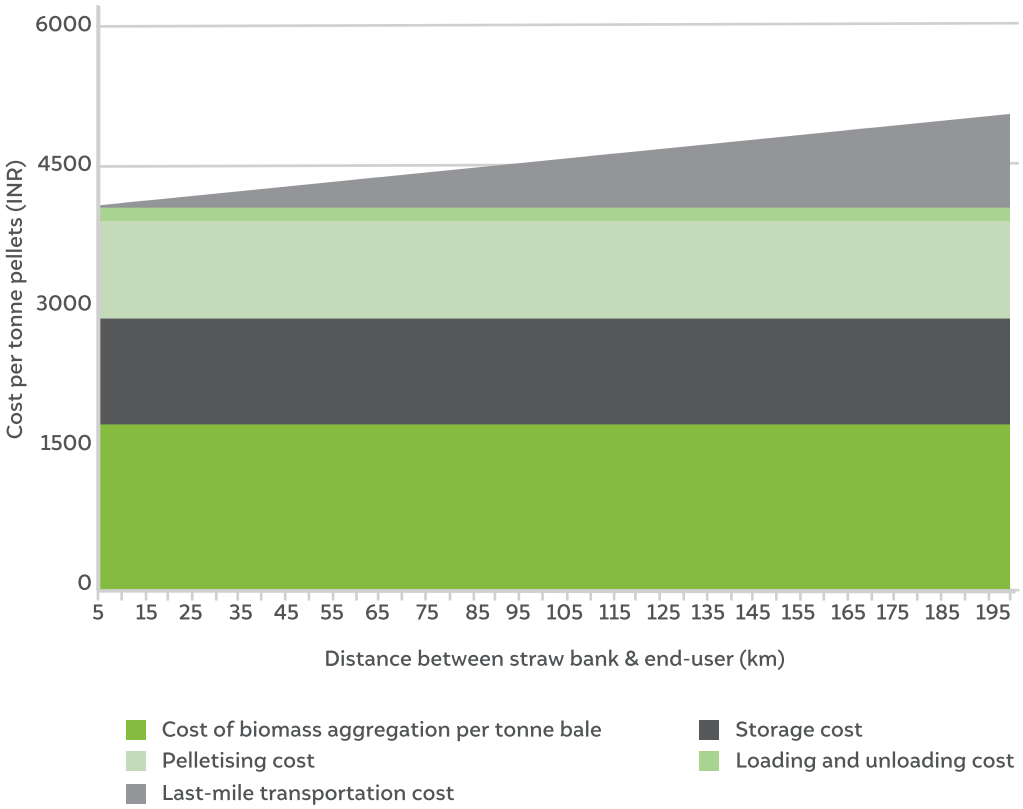


Figure 6
Delivered cost of per tonne of pellets from farm to end-user is highest among all processing options for crop residue

Source: Authors' analysis

Different end-users require biomass in different forms. Industries such as biomass power plants, packaging, and paper and cardboard industries require biomass in the bale form, whereas briquettes are used in large and medium-scale industry boilers as it is conducive to heating. But coal-fired power plants and smaller heating devices such as pellet stove and furnace prefer the pellet form of the biomass. Transporting residue in the form of pellets entails the highest cost compared to bales or briquettes owing to the high processing cost involved (Figure 7). In most cases, SCM entities transport residue in the form of bales to avoid the high investment needed for biomass processing. In such instances, the end-user has to invest in processing bales to briquettes or biomass if required.

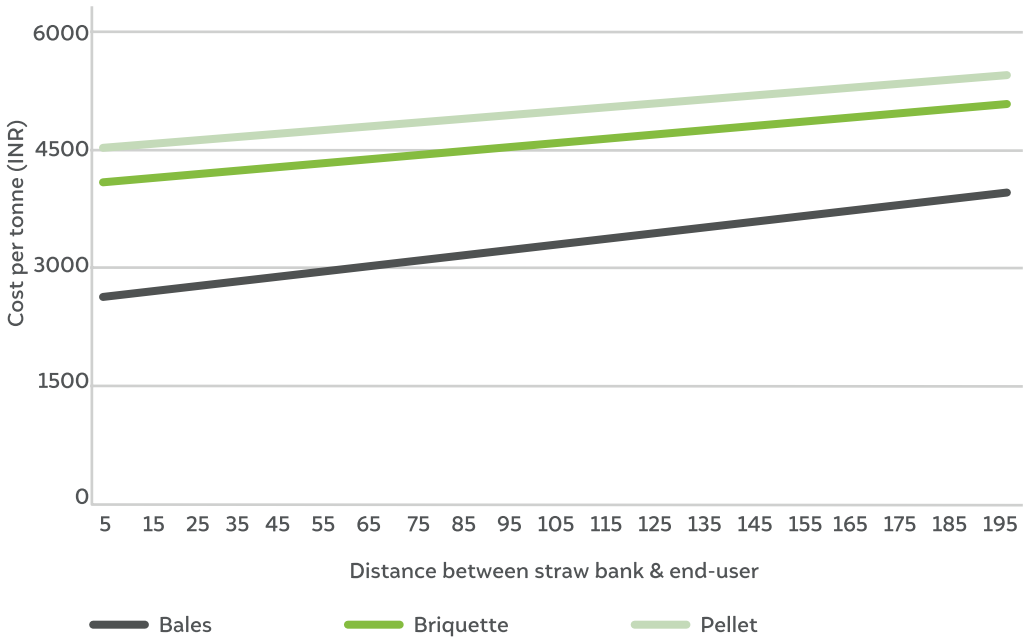


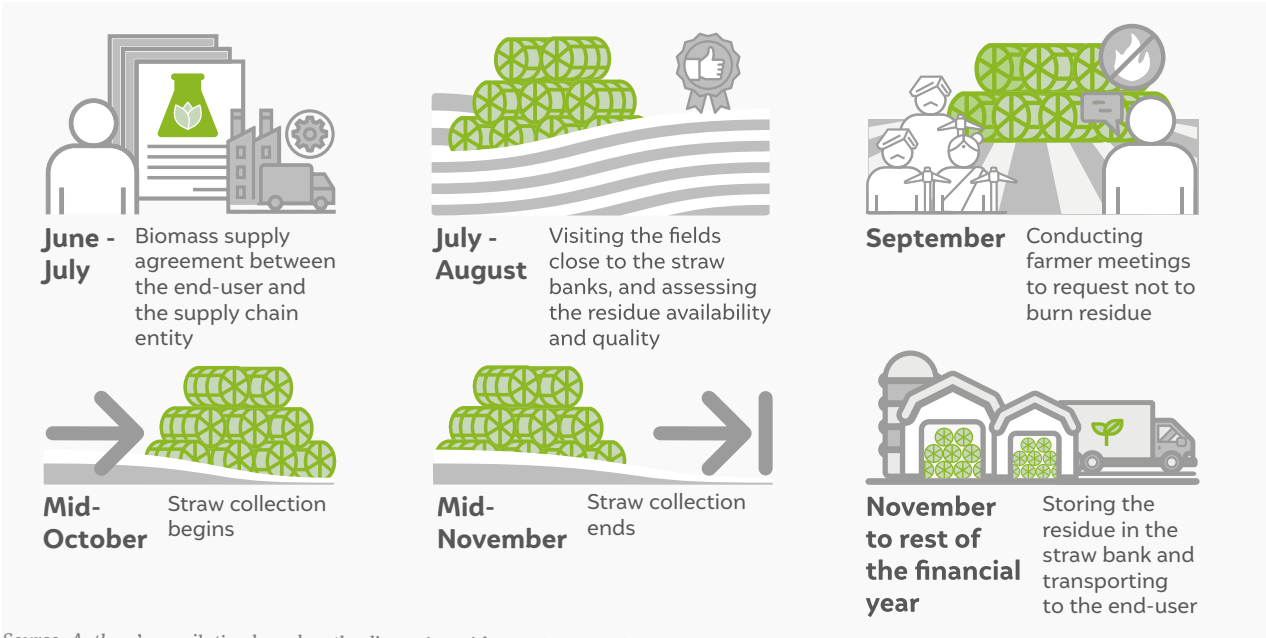
Figure 7 Delivered cost of biomass-based products increases with distance

Source: Authors' analysis

Mapping the end-users

The entire process of supplying crop residue for ex-situ uses takes between 6 and 12 months (Table 6). The delivered cost rises with an increase in distance and is very high for a longer distance between the farm and the end-user (Figure 7). If a provision for decentralised sourcing of crop residue within a radius of 5–10 km exists for the end-user, interim storage is not needed and transportation cost comes down significantly. We recommend that the Punjab Energy Development Agency (PEDA) create a database of end-users mapped to their annual crop residue demand. This database would help the SCM entities to optimally plan for storage and logistics. In addition, the areas where there is a deficit in demand or potential end-users the state government can focus their in-situ management resources in that particular areas. In those regions, more CHCs need to be strategically established to increase the accessibility of in-situ farm implements such as Happy seeder.

Table 6 Timeline of an ex-situ crop residue operation



Source: Authors' compilation based on the discussion with prominent ex-situ entities in Punjab

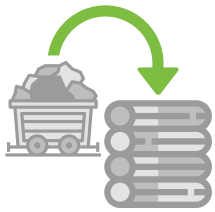
Co-firing biomass in thermal power plants

Several countries see co-firing biomass with coal for electricity generation and other heating requirements in power plants and industries as an attractive option because of the derived social and environmental benefits (Demirbas 2003; Hansson et al. 2009; Sullivan and Meijer 2010). The limitations of this application of biomass include low heating values, variable chemical and physical properties, high investment and operating cost, and unreliable biomass feedstock supply (Dai et al. 2008).

Coal-fired thermal power plants contribute 58.26 per cent to the total installed capacity of 14,205.42 MW in Punjab (Central Electricity Authority 2020). The two thermal power plants operating on coal (Punjab State Power Corporation Limited 2020) in Punjab require 11.34 million tonnes of coal per annum (Government of India and Government of Punjab 2016).

Power demand in Punjab generally peaks during the kharif season mainly for irrigation needs. In early 2020, due to the COVID-19 pandemic-induced lockdown, energy demand dropped as most industries were not functioning in the state. By procuring cheaper coal from Coal India Limited (CIL) rather than relying upon coal imports (Tribune News 2020), Punjab State Power Corporation Limited (PSPCL) intends to save INR 100 crore for the Punjab government. But the use of cheaper coal is not conducive to the environment due to increased NOx emissions (Agraniotis et al. 2017). One possible option to reduce coal consumption is to co-fire thermal plants with 10 per cent of biomass, which offers twin benefits of reduced pollution and consumption of crop waste. For instance, if 10 per cent of paddy residue pellets are blended with coal, it would lead to a consumption of 1.47 million tonnes of paddy residue annually. We recommend that the Punjab government actively look at the option of co-firing coal-fired power plants with biomass. We also offer the same recommendation to states such as Haryana and Uttar Pradesh that resort to stubble burning on a large scale.

We now see how the economics of biomass as feedstock supply works out against coal. We take the delivered price of coal from the tariff order of Guru Gobind Singh Super Thermal Power Plant, Ropar, Punjab, for the financial year 2018–19 for our comparison. We note that the delivered cost of biomass pellets per 5,000 Mcal within a radius of 50 km of the power plant is economical compared to coal (Figure 8). On the other hand, the delivered cost of briquettes per 5,000 Mcal is economical up to a distance of 145 km. Therefore, we recommend that industries that consume coal for its heating needs should explore the prospects of locally sourcing the biomass briquettes within the state.



Industries that consume coal for their heating needs should explore the prospects of locally sourcing the biomass briquettes within the state

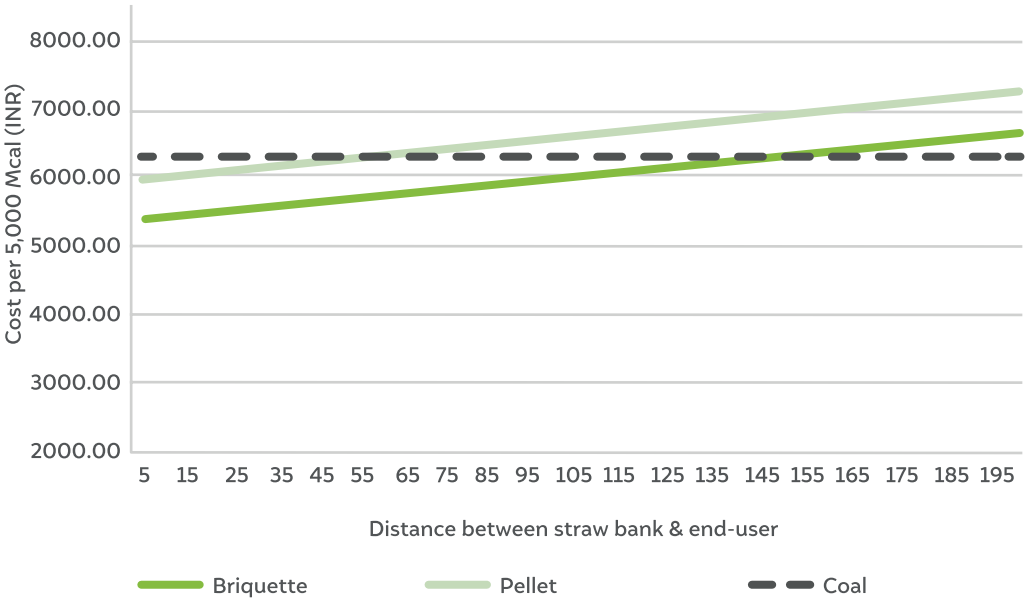


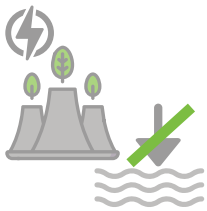
Figure 8
Delivered cost of briquetted and pelleted biomass is lower than that of coal up to a certain distance between the farm and the end-user

Source: Authors' analysis

4.3 Concerns regarding the use of paddy residue for ex-situ applications

The Punjab government reported that 12.85 million tonnes of paddy were burnt in the state in 2018 (Ministry of Agriculture & Farmers Welfare 2018). For collecting all of this paddy straw, transporting, and storing them at straw banks, 37,000 tractor-trailers would be needed every day on the village roads of Punjab for 20 days during the kharif harvesting season. This is not viable as village roads may get congested when handling a high volume of traffic due to poor infrastructure.[8] Also, residue collection at this scale is estimated to lead to an emission of 29 Gg CO2.[9] A detailed assessment of traffic congestion on village roads and vehicular emissions involved in logistics of a huge scale of biomass use is needed. Further, a rigorous life-cycle assessment (LCA) of SCM would help project managers in deciding the ideal fuel for logistics and biomass processing.

During the harvesting season in Punjab, the high dew content leaves a larger moisture content in paddy residue. The residue is generally sun-dried to bring down the moisture content to 10 per cent. If moisture content doesn't drop to 10 per cent or below, drying equipment should be used, which entails additional expenditure. SCM entities also need to find a way to optimise logistics cost as trucks return empty to the farm after delivering biomass to straw banks.



Setting up new biomass plants for power generation would strain the state's already depleting water reserves

Thermal power plants need water in the range of 3.5–9 m3/hour per MW depending upon the technology deployed for power generation (Chaturvedi et al. 2018). Comparatively, an 18 MW biomass power plant in Punjab needs 225 m3/hour of water (Joshi et al. 2019). Punjab has 11 operational biomass power plants with an aggregate capacity of 92.60 MW. These plants consume 8.80 lakh tonnes of paddy straw, which is 2.75 per cent of the total paddy residue generated in Punjab annually. The Punjab government is also planning to set up several new biomass-based refineries and power plants in the state to tackle the problem of stubble

8. To collect 13 million tonne of paddy residue, 37,143 tractor-trailer trips would be required for 20 days (assuming 7 tonnes of paddy residue is collected in each trip and round trips made by a tractor in one day).

9. Assuming an average distance of 15 km between farm and straw bank, the emission factor for tractor-trailer = 515.2 g CO2/km (Ramachandra and Shwetmala 2009).

burning (Chaba 2019). A dystopian future for Punjab is predicted due to its depleting water reserves on account of a dominant paddy cultivation system in the state (Tur 2018; Krishan et al. 2014; Baweja et al. 2017). It is also known that out of all power generation technologies, biomass-based projects need a very high amount of water (Zhu, et al. 2019). Setting up new biomass plants for power generation surely would strain the state's already depleting water reserves. Hence, detailed studies are needed to quantify water footprint for setting up and operating new plants and also to determine the source of required water.

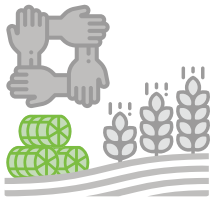
5. Policy recommendations



For this study, we interacted with several stakeholders involved in the SCM of crop residue in Punjab. We specifically looked at the ex-situ crop residue management in the state and its economics. Based on our observations and consultations, we recommend the following strategies for better management and scaling up the ex-situ applications of crop residue in Punjab.

5.1 Establish a dense network of straw banks and ex-situ supply chain ecosystem

From our estimates and stakeholder dialogues, ex-situ SCM is found to be financially viable when crop residue collection is done within 15 km and supplied to the end-user situated within 50 km. A dense network of straw banks and an efficient SCM can help achieve these targets. Also, a huge opportunity for green jobs is also created in the rural areas of Punjab. Interim storage at straw banks approximately cost INR 800 per tonne. Storing 13 million tonnes of paddy residue (burnt in 2018) in the straw banks would cost INR 1,040 crores. Needless to say, this investment needs to be mobilised from the private sector. The Punjab government should make way for SCM entities to lease common panchayat land or other fallow lands as it reduces storage cost. Agencies such as National Bank for Agriculture And Rural Development (NABARD) and Indian Renewable Energy Development Agency Ltd (IREDA) could be roped in to provide loans and insurance for entities involved in SCM. Government agencies such as PEDA in partnership with private entities should also implement pilot programmes to scale up participation of the private sector in ex-situ crop residue management.



Government agencies such as PEDA in partnership with private entities should also implement pilot programmes to scale up participation of the private sector in ex-situ crop residue management.

5.2 Regulation for minimum crop residue usage

Boosting the demand for biomass would encourage participation of the private sector in ex-situ supply chain operations. The Central Electricity Authority of India (CEA) recommends a blend of 5–10 per cent of biomass pellets along with coal in the fluidised bed and pulverised coal units (Central Electricity Authority 2017). If state's coal-fired thermal power plants are co-fired with 10 per cent of blended biomass pellets, it would take up 1.47 million tonnes of paddy residue annually. We also urge the Punjab government to introduce a regulation mandating the minimum usage ratio of crop residue biomass in industrial boilers, coal-fired thermal plants, and biomass electricity generation projects. Stipulating a blending ratio and issue of technical guidelines for residue usage in industries will certainly boost its uptake.

5.3 Price regulation of crop residue and its products

In December 2018, NTPC invited expression of interest (EOI) for supplying non-torrefied biomass-based pellets for 21 power plants, none of which are in Punjab (NTPC Limited 2018). We arrived at a delivered cost of biomass at INR 5,500 for a distance of 200 km. Delivering biomass from Punjab to these NTPC plants would push up the delivered cost of biomass pellets even higher, making it unprofitable for SCM entities to enter into a long-term fuel supply agreement with those power plants. Every year, the Chhattisgarh State Electricity Regulatory Commission (CSERC) fixes the minimum support price for rice husk, which is one of the largest biomasses produced in the state. We also suggest to PSERC (along with PEDA) to establish a minimum price for biomass-based products every year by making an allowance for adequate compensation to farmers for the costs they incur in getting the crop residue out of their farm. This would also reduce the monopoly of SCM entities over prices and provide them with satisfactory profit and also reasonable prices for end-users.

5.4 Mapping of end-users

The entire process of supplying crop residue for ex-situ uses takes between 6 and 12 months. From the analysis, it is evident that the delivered cost rises with an increase in distance and is very high for a longer distance between the farm and the end-user. If a provision for decentralised sourcing of crop residue within a radius of 5–10 km exists for the end-user, interim storage is not needed and transportation cost comes down significantly. We recommend that the PEDA create a database of end-users mapped to their annual crop residue demand. This database would help the SCM entities to optimally plan for storage and logistics. In addition, the areas where there is a deficit in demand or potential end-users the state government can focus their in-situ management resources in that particular areas. As with increase in distance ex-situ becomes less economical, the policymakers should prioritize deployment of in-situ farm implements like Happy seeders in these locations.



If a provision for decentralised sourcing of crop residue within a radius of 5–10 km exists for the end-user, interim storage is not needed and transportation cost comes down significantly.

5.5 Digital platform for farmers to trade their residue

For an efficient supply chain management, a digital platform, similar to National Agriculture Market (eNAM), connecting farmers, straw banks, end-users, and various stakeholders needs to be created. We also suggest a process for the functioning of this platform. Farmers should be able to raise a request to collect crop residue from their farm through the platform. The nearest straw bank (ideally within 15 km) would accept the request and collect the residue from the farm within the stipulated time frame. The compensation for the crop residue could be deposited to the farmers' bank account directly. This would ease the burden of farmers in collecting and transporting the residue from their farms to straw banks and increase the profitability of both farmers and straw banks. Also, if a provision exists for end-users to raise a demand request for biomass, SCM entities should be able to arrange for residue collection within a radius of 5–10 km of the end-user. This improves the optimisation of SCM and reduces transport and storage cost.



After harvesting paddy, farmers in Punjab have a limited span of 15-20 days to prepare the land for the next crop. This reduced timeframe is an offshoot of the *Punjab Preservation of Subsoil Act (2009)*, implemented to save groundwater by mandatorily postponing the transplanting of paddy from April–May to beyond 10 June.

Image: iStock

6. Conclusion

Punjab has been experiencing the problem of stubble burning for several decades, causing extensive air pollution. Needless to say, a multi-pronged strategy is needed to curb the practice. Several approaches, including in-situ treatments such as the use of Happy Seeder, mulching, composting, and ex-situ options, have been recommended to dispose of crop residue beneficially. The farmer's choice of crop residue management depends upon the economics of the selected option, availability of implements, and the time needed for its implementation. Burning the stubble on the open field is always the easiest option for farmers. In 2018, 13 million tonnes of paddy residue (65 per cent of the total 20 million tonnes generated) were set on fire in the fields of Punjab, choking the Indo-Gangetic plains. Several experts have expressed the view that biomass-based power generation in Punjab can result in a useful consumption of crop residue as well as reduce pollution from coal-fired thermal power plants. Lack of a dense network of biomass supply chain facilities is proving to be a hurdle for scaling up ex-situ management of crop waste in the state.



As with increase in distance ex-situ becomes less economical, the policymakers should prioritize deployment of in-situ farm implements like Happy seeders in these locations

Supply chain entities play a prominent role in reaching the biomass from the farmer and the end-user. Only 10 well-known SCM entities are currently involved in the supply chain business in Punjab. Developing a dense network of crop residue managing facilities could significantly reduce cost of transport. Our estimates show that transit of paddy residue in the form of bales for a distance of 5–15 km from farm to straw bank typically costs INR 1,150–1,350 per tonne. Cutting and baling of residue constitutes 55 per cent of the delivered cost of residue from farm to the straw bank. The role of supply chain entities role in providing customised farm implements needed for residue collection assumes greater importance, as individual farmers do not maintain an inventory of ex-situ farm implements.

If the state government decides to use 10 per cent of blended biomass pellets for co-firing state's coal-fired thermal power plants, 1.47 million tonnes of paddy residue will be consumed annually in this process. Introducing a regulation mandating the minimum usage ratio of biomass derived from crop residue and stipulating a blending ratio and technical guidelines for residue usage in appropriate industries will significantly boost the demand for crop residue.

Though biomass pellets are the most demanded form of paddy residue in power plants, our estimates indicate that the delivered cost of pelleted crop residue is the highest among available options. Decentralising the sourcing of crop residue within a radius of 5–10 km of the end-user's location would curtail the need for interim storage and drastically bring down transportation costs. Further, a clear data that maps all the end-users with their annual biomass demand would help supply chain entities to develop clear storage and logistics plans. In addition, the areas where there is a deficit in demand or potential end-users the state government can focus their in-situ management resources in that particular areas. As with increase in distance ex-situ becomes less economical, the policymakers should prioritize deployment of in-situ farm implements like Happy seeders in these locations.

References

Agraniotis, Michalis, Christian Bergins, Malgorzata Stein-Cichoszewska, and Emmanouil Kakaras. 2017. "5 - High-Efficiency Pulverized Coal Power Generation Using Low-Rank Coals." In *Low-Rank Coals for Power Generation, Fuel and Chemical Production*, edited by Zhongyang Luo and Michalis Agraniotis, 95–124. Woodhead Publishing. doi:10.1016/B978-0-08-100895-9.00005-X.

Balingbing, Carlito, Nguyen Van Hung, Nguyen Thanh Nghi, Nguyen Van Hieu, Ampy Paulo Roxas, Caesar Joventino Tado, Elmer Bautista, and Martin Gummert. 2020. "Mechanized Collection and Densification of Rice Straw." In *Sustainable Rice Straw Management*, edited by Martin Gummert, Nguyen Van Hung, Pauline Chivenge, and Boru Douthwaite, 15–32. Cham: Springer International Publishing. doi:10.1007/978-3-030-32373-8_2.

Baweja, Samanpreet, Rajan Aggarwal, Mandeep Brar, Rajan Aggarwal, and Mandeep Brar. 2017. "Groundwater Depletion in Punjab, India." *Encyclopedia of Soil Science*. CRC Press. January 11. doi:10.1081/E-ESS3-120052901.

Bernard, Steven, and Amy Kazim. 2018. "Dirty Air: How India Became the Most Polluted Country on Earth," December 11. https://ig.ft.com/india-pollution/.

Bhattacharya, Amit. 2018. "8-Fold Rise in Happy Seeder Sales Raises Hope | Delhi News - Times of India." *The Times of India*. October 5. https://timesofindia.indiatimes.com/city/delhi/8-fold-rise-in-happy-seeder-sales-raises-hope/articleshow/66077515.cms.

Blunk, S. L., M. W. Yore, M. D. Summers, K. Lau, S.T. Tang, and B. M. Jenkins. 2003. "Quality and Property Changes in Rice Straw During Long Term Storage." In . ASABE Paper No. 036091. St. Joseph, MI: ASABE. doi:10.13031/2013.15406.

Central Electricity Authority. 2017. "Biomass Utilization for Power Generation through Co-Firing in Pulverized Coal Fired Boilers- Advisory." CEA/TETD-TT/2017/M-25/1137-1251. R K Puram, New Delhi-66: Ministry of Power. http://www.cea.nic.in/reports/others/thermal/tetd/Biomass%20Utilization%20Advisory.pdf.

Central Electricity Authority. 2020. "All India Installed Capacity (in MW) of Power Stations (As on 30.06.2020)." CEA. http://cea.nic.in/reports/monthly/installedcapacity/2020/installed_capacity-06.pdf.

Central Electricity Regulatory Commission. 2019. "CERC RE Tariff Order 2019-20." Tariff Order Petition No. 1/SM/2019 (Suo-Motu). New Delhi. http://www.cercind.gov.in/2019/orders/Draft%20RE%20Tariff%20Order%20for%20FY%202019-20.pdf.

Chaba, Anju Agnihotri. 2019. "Punjab Looks to Biomass Power Plants to Tackle Stubble Problem." *The Indian Express*. November 25. https://indianexpress.com/article/india/punjab-looks-to-biomass-power-plants-to-tackle-stubble-problem-6135111/.

Chaba. 2020a. "Farm Pollution: Happy Seeder Produces Not-so-Happy Results on Ground." *The Indian Express*. January 16. https://indianexpress.com/article/india/farm-pollution-happy-seeder-produces-not-so-happy-results-on-ground-6218741/.

Chaba. 2020b. “Explained: How Punjab Plans to Spin Paddy Stubble into Gold.” *The Indian Express*, September 13. https://indianexpress.com/article/explained/biomass-power-plants-bio-cng-bioethanol-how-peda-plans-to-spin-paddy-stubble-into-gold-6590158/.

Chakrabarti, Suman, Mohammed Tajuddin Khan, Avinash Kishore, Devesh Roy, and Samuel P. Scott. 2019. “Risk of Acute Respiratory Infection from Crop Burning in India: Estimating Disease Burden and Economic Welfare from Satellite and National Health Survey Data for 250 000 Persons.” *International Journal of Epidemiology* 48 (4). Oxford Academic: 1113–24. doi:10.1093/ije/dyz022.

Chaturvedi, Vaibhav, Rudresh Sugam, Poonam Nagar Koti, and Kangkanika Neog. 2018. “Energy-Water Nexus and Efficient Water-Cooling Technologies for Thermal Power Plants in India: An Analysis within an Integrated Assessment Modelling Framework.” New Delhi, India: Council on Energy, Environment and Water. https://niti.gov.in/sites/default/files/2019-08/Research%20Study%20on%20Energy%20Water%20Nexus%20%26%20Efficient%20Water%20Cooling%20Technologies%20for%20thermal%20Power%20Plants%20in%20India_CEEW_New%20Delhi.pdf.

Clarke, Steve, and Fernando Preto. 2011. “Biomass Densification for Energy Production.” 11–035. Ontario: MINISTRY OF AGRICULTURE, FOOD AND RURAL AFFAIRS. http://www.omafra.gov.on.ca/english/engineer/facts/11-035.htm.

Confederation of Indian Industry, and National Institution for Transforming India Aayog. 2018. “ACTION PLAN FOR BIOMASS MANAGEMENT.” Cleaner Air BetterLife Initiative. Sansad Marg, Sansad Marg Area, New Delhi, Delhi 110001. https://niti.gov.in/writereaddata/files/document_publication/ReportOfTheTaskForceOnBiomassManagement.pdf.

Dai, Jianjun, Shahab Sokhansanj, John R. Grace, Xiaotao Bi, C. Jim Lim, and Staffan Melin. 2008. “Overview and Some Issues Related to Co-Firing Biomass and Coal.” *The Canadian Journal of Chemical Engineering* 86 (3): 367–86. doi:10.1002/cjce.20052.

Demirbas, Ayhan. 2003. “Sustainable Cofiring of Biomass with Coal.” *Energy Conversion and Management* 44 (9): 1465–79. doi:10.1016/S0196-8904(02)00144-9.

Gill, Nishant, Ritu Dogra, and Baldev Dogra. 2018. “Influence of Moisture Content, Particle Size, and Binder Ratio on Quality and Economics of Rice Straw Briquettes.” *BioEnergy Research* 11 (1): 54–68. doi:10.1007/s12155-017-9877-9.

Government of India, and Government of Punjab. 2016. “24X7 Power For All, Punjab.” https://powermin.nic.in/sites/default/files/uploads/joint_initiative_of_govt_of_india_and_Punjab.pdf\.

Government of Punjab. 2019. “Statistical Abstract of Punjab.” https://www.esopb.gov.in/static/PDF/Abstract2019.pdf.

Goyal, Divya. 2019. “Explained: Using Happy Seeder and How It Affects Wheat Yield.” The Indian Express. September 22. https://indianexpress.com/article/explained/explained-using-happy-seeder-and-how-it-affects-wheat-yield-6017640/.

Gupta, Niti. 2019. “Paddy Residue Burning in Punjab.” https://www.ceew.in/publications/paddy-residue-burning-punjab.

Gupta, Ridhima, and E. Somanathan. 2016. “Happy Seeder: A Solution to Agricultural Fires in North India.” *Ideas For India*. November 12. http://www.ideasforindia.in/topics/environment/happy-seeder-a-solution-to-agricultural-fires-in-north-india.html.

Hansson, Julia, Göran Berndes, Filip Johnsson, and Jan Kjärstad. 2009. “Co-Firing Biomass with Coal for Electricity Generation—An Assessment of the Potential in EU27.” *Energy Policy* 37 (4): 1444–55. doi:10.1016/j.enpol.2008.12.007.

India State-Level Disease Burden Initiative Air Pollution Collaborators. 2020. “The Impact of Air Pollution on Deaths, Disease Burden, and Life Expectancy across the States of India: The Global Burden of Disease Study 2017 - The Lancet Planetary Health.” Accessed August 18. https://www.thelancet.com/journals/lanplh/article/PIIS2542-5196(18)30261-4/fulltext.

Ishii, Kazuei, Toru Furuichi, Atsushi Fujiyama, and Shintaro Watanabe. 2016. “Logistics Cost Analysis of Rice Straw Pellets for Feasible Production Capacity and Spatial Scale in Heat Utilization Systems: A Case Study in Nanporo Town, Hokkaido, Japan.” *Biomass and Bioenergy* 94 (November): 155–66. doi:10.1016/j.biombioe.2016.08.007.

Jain, Akshat. 2019. “The Curious Case of Punjab Preservation of Subsoil Water Act, 2009 and Air Pollution - *LexQuest* Foundation.” Lexquest.In/. November 21. https://lexquest. in/the-curious-case-of-punjab-preservation-of-subsoil-water-act-2009-and-air-pollution/.

Jethva, Hiren, Omar Torres, Robert D. Field, Alexei Lyapustin, Ritesh Gautam, and Vinay Kayetha. 2019. “Connecting Crop Productivity, Residue Fires, and Air Quality over Northern India | Scientific Reports.” *Scientific Reports* 9 (November). doi:https://doi.org/10.1038/s41598-019-52799-x.

Joshi, Milind, Nilesh Deshmukh, Krishna Kumar Sinha, Pawan Kuma Verma, D.C. Meshram, Manoj Kumar Mishra, Pavanrekha Mishra, et al. 2019. “Final Environmental Impact Assessment Reprt & Environmental Management Plan for Proposed 18 (1X18) MW Biomass Based Thermal Power Plant.” SMSEL/ EIA/ SAEL/ 0022-R1. SMS Envocare Limited. http://environmentclearance.nic.in/writereaddata/FormB/EC/EIA_EMP/11012019M8Q8Q8Q8EIA REPORT.pdf.

Krishan, Gopal, M. Someshwar Rao, R.s Loyal, Anil Kumar Lohani, NK Tuli, KS Takshi, C. P. Kumar, Prabhat Semwal, and Sandeep kumar. 2014. “Groundwater Level Analyses of Punjab, India: A Quantitative Approach.” *Octa Journal of Environmental Research* 2 (September): 221–26.

Kumar, Akanksha. 2019. “Why Farmers Are Forced to Choose Stubble Burning Over Happy Seeder.” *TheQuint*. November 5. https://www.thequint.com/news/india/air-pollution-in-delhi-ncr-stubble-burning-in-punjab.

Kumar, Naresh. 2020. Rate chart by Sri Rishabh TravelsTelephonic.

Kurinji, L.S. 2019. “Alternative Methods to Monitor Air Pollution.” *CEEW*. March 29. https://www.ceew.in/publications/alternative-methods-monitor-air-pollution.

Ministry of Agriculture & Farmers Welfare. 2018. “Review of the Scheme on Promotion of Agricultural Mechanization for In-Situ Management of Crop Residue in the States of Punjab, Haryana, Uttar Pradesh and NCT of Delhi.” OPERATIONAL GUIDELINES. KrishiBhavan, New Delhi-110001: Department of Agriculture, Cooperation& Farmers Welfare. https://agrimachinery.nic.in/Files/Guidelines/CRM.pdf.

Moudgil, Manu. 2020. "Why Punjab Failed to Douse Farm Fires." *Mongabay-India*. January 5. https://india.mongabay.com/2020/01/why-punjabs-farm-fires-continue-to-burn/.

Nirmal, Rajalakshmi. 2019. "Where Are the Happy Seeders That Punjab's Farmers Were Promised?" @ *businessline*, November 17, sec. Agri Business. https://www.thehindubusinessline.com/economy/agri-business/where-are-the-happy-seeders-that-punjabs-farmers-were-promised/article30000119.ece.

NTPC Limited. 2018. "Expression Of Interest (EOI) for Production and Supply of Paddy Straw and Other Agro Residue Based Pellets / Torrefied Pellets to NTPC Power Plants for Power Generation Through Biomass Co-Firing." Invitation for Expression of Interest EOI No.: PEM-9598-999-3. NTPC Limited. http://www.ntpctender.com/morqueDocs/Revised_EOI_Document.pdf.

Office of the Labour Commissioner. 2020. "Adjustment of Minimum Rates of Wages W.E.F. 01.03.2020 in the State of Punjab." ST/9408. Mohali, Punjab. https://labourlawreporter.com/wp-content/uploads/2019/11/Minimum_1_5_2020_WagesEng-Punjab-2.pdf.

Possidônio, Elisangela Fernandes da Silva Campana, João Eduardo Azevedo Ramos da Silva, Eli Angela Vitor Toso, Elisangela Fernandes da Silva Campana Possidônio, João Eduardo Azevedo Ramos da Silva, and Eli Angela Vitor Toso. 2016. "Biomass Supply Chain Analysis with Densified Waste." Revista Árvore 40 (2). Sociedade de Investigações Florestais: 355–62. doi:10.1590/0100-67622016000200018.

Press Trust of India. 2019. "Stubble Burning Caused 44% Of Delhi Pollution This Year: Pollution Board." *NDTV.Com*. November 14. https://www.ndtv.com/india-news/stubble-burning-caused-44-of-delhi-pollution-this-year-pollution-board-2132125.

Punjab Renewable Energy System Pvt. Ltd. 2020. "Biomass Briquetting & Pelleting." Accessed July22. https://prespl.com/product-and-services/biomass-briquetting--pelleting.

Punjab State Power Corporation Limited. 2020. "Thermal Projects – PSPCL." Accessed August 27. https://www.pspcl.in/thermal-projects/.

Ramachandra, T.V., and Shwetmala. 2009. "Emissions from India's Transport Sector: Statewise Synthesis." *Atmospheric Environmen*t 43 (34): 5510–17. doi:10.1016/j.atmosenv.2009.07.015.

Rao, G. v r Subba. 2019. "Use of Paddy Balers Picks Up." *The Hindu*, May 13, sec. Andhra Pradesh. https://www.thehindu.com/news/national/andhra-pradesh/use-of-paddy-balers-picks-up/article27110782.ece.

Ray, Allison E., Chenlin Li, Vicki S. Thompson, Dayna L. Daubaras, Nicholas J. Nagle, and Damon S. Hartley. 2017. "Biomass Blending and Densification: Impacts on Feedstock Supply and Biochemical Conversion Performance." *Biomass Volume Estimation and Valorization for Energy*, February. IntechOpen. doi:10.5772/67207.

Sharma, Ankit, and Rupinder Chandel. 2016. "Comparative Field and Economic Evaluation of Baler for Baling Paddy Straw." *Scientific Journal AGRICULTURAL ENGINEERING* 3: 69–76.

Sharma, Mohit, Ishan Sahajpal, Aditya Bhuyan, and Akansha Chauhan. 2019. "Impact Assessment Report for Crop Residue Management Project." Cleaner Air Better Life' Initiative. Lodi Road, New Delhi 110003, India: Confederation of Indian Industry (CII).

Singh, Jagtar, B. S. Panesar, and S. K. Sharma. 2010. “A Mathematical Model for Transporting the Biomass to Biomass Based Power Plant.” *Biomass and Bioenergy* 34 (4): 483–88. doi:10.1016/j.biombioe.2009.12.012.

Singh, Sukhpal. 2018. “Farm Machinery Rental Services: Case Studies from Punjab.” In *Institutional Innovations in the Delivery of Farm Services in India: A Smallholder Perspective*, edited by Sukhpal Singh, 33–70. New Delhi: Springer India. doi:10.1007/978-81-322-3753-2_3.

Suardi, Alessandro, Simone Bergonzoli, Vincenzo Alfano, Antonio Scarfone, and Luigi Pari. 2019. “Economic Distance to Gather Agricultural Residues from the Field to the Integrated Biomass Logistic Centre: A Spanish Case-Study.” *Energies* 12 (16). Multidisciplinary Digital Publishing Institute: 3086. doi:10.3390/en12163086.

Sullivan, Kevin, and Ronald Meijer. 2010. “Co-Firing Biomass with Coal.” *Renewable Energy World*. May 19. https://www.renewableenergyworld.com/2010/05/19/co-firing-biomass-with-coal/.

The Economic Times. 2020. “Diesel Price in Punjab | Check Latest Diesel Rates in Punjab.” *The Economic Times*. Accessed July 22. https://economictimes.indiatimes.com/wealth/fuel-price.

TIFAC, and IARI. 2018. “Estimation of Surplus Crop Residues in India for Biofuel Production.” Joint Report. New Delhi. https://tifac.org.in/images/pdf/pub/TIFACReports/newreports/biomass_w(1).pdf.

Tribune News. 2020. “Punjab Not to Import Coal for Power Plants, to Save Rs100 Cr.” *Tribuneindia News Service*. June 16. https://www.tribuneindia.com/news/punjab/state-not-to-import-coal-for-power-plants-to-save-rs100-cr-99724.

Tur, Jatinder Kaur. 2018. “Rapid Fall in Groundwater Level in Punjab, Haryana.” *Hindustan Times*. May 25. https://www.hindustantimes.com/punjab/rapid-fall-in-groundwater-level-in-punjab-haryana/story-fIscV9WfuEo65g7xnKHohI.html.

Van Hung, Nguyen, Monet Concepcion Maguyon-Detras, Maria Victoria Migo, Reianne Quilloy, Carlito Balingbing, Pauline Chivenge, and Martin Gummert. 2020. “Rice Straw Overview: Availability, Properties, and Management Practices.” In *Sustainable Rice Straw Management*, edited by Martin Gummert, Nguyen Van Hung, Pauline Chivenge, and Boru Douthwaite, 1–13. Cham: Springer International Publishing. doi:10.1007/978-3-030-32373-8_1.

Whittaker, Carly, and Ian Shield. 2017. “Factors Affecting Wood, Energy Grass and Straw Pellet Durability – A Review.” *Renewable and Sustainable Energy Reviews* 71 (C). Elsevier: 1–11.

Annexure I

Table A1: Data used to compute the delivered cost of biomass

Parameter	Data	Source
Farm to straw bank		
Diesel price	INR 72.21	Diesel prices for the state of Punjab w.e.f. 10-07-2020 (The Economic Times 2020)
Rice yield of Punjab	4132 kg/ha	Statistical Abstract of Punjab, 2019 (Government of Punjab 2019)
Tractor rent charges	INR 217.14/hour	S. Singh (2018)
Trailer rent charges	INR 27.92/hour	S. Singh (2018)
Working hours of skilled labour required for loading, unloading and stacking	3.4 hour/tonne	Gill, Dogra, and Dogra (2018)
Trailer capacity	7 tonnes	Stakeholder consultation
Straw bank to end-user		
Storage cost	INR 0.80 per kg	Stakeholder consultation
The payload capacity of the truck	8,000 kg	Specifications provided by truck rental service
Truck rental cost per km	INR 40	Per km rate obtained from private truck rental service (includes fuel and other logistic charges) Can vary according to diesel prices and border toll charges (N. Kumar 2020)
Minimum wage (skilled worker)	INR 48.79/hour	Minimum wages for labour for the state of Punjab w.e.f. 01-05-2020 from the Office of The Labour Commissioner, Punjab (Office of the Labour Commissioner 2020)
Energy content of paddy residue	2,750 kcal/kg	NITI Aayog's action Plan for biomass management (CII and NITI Aayog 2018)
Energy content of paddy briquette	3,800 kcal/kg	Technical specification of PRESPL products (Punjab Renewable Energy System Pvt. Ltd 2020)
Energy content of paddy pellet	3,750 kcal/kg	Ishii et al. (2016)

Source: Authors' compilation

Table A2: Data on farm implements involved in cutting and baling the residue

Farm machinery	Fuel (diesel) consumption(litre/acre)	Rental cost (INR/acre)	Note
Chopper/ cutter-cum-spreader	3	157	- Rental cost on time basis: INR 1,000–1,200/ day as charged to farmers by farmer co-op societies - Field capacity: 3–4 acre/hour (15–20 min/ acre)
Baler + Raker	6–7 (Baler) 3–4 (Raker)	1,500–2,000	- Rental cost on acreage basis as charged to farmers by a service provider is INR 1,500–2,000/acre - This rental cost also includes the fuel and labour costs for combine harvester operation - Field capacity: 2 acre/hour (30 min./acre)

Source: (M. Sharma et al. 2019)

Table A3: Freight rates for final transportation

Distance (km)	Transport cost (INR/km)
0 - 50	40
51 - 150	39
151 - 300	38
301 - 500	37
501 - 1000	36

Source: Authors' compilation based on stakeholder consultation
**Note: Cost is relevant for trucks with a payload capacity of 8 tonnes and is subject to change with fluctuation in diesel price*

Table A4: Economics of briquette production

Title	Chaff cutter	Chipper grinder	Briquette press	Operational shed
Depreciation + interest + tax + insurance (INR/hour)	6.26	53.51	66.34	16.62
Electricity required (kWh)	13.51	37.84	53.16	0.96
Electricity cost (INR/hour)	65.80	184.30	258.87	4.65
Labour charges (INR/hour)	104.58	104.58	104.58	
Repair and maintenance (INR/hour)	1.89	16.12	19.98	8.84
Cost of lubrication (INR/hour)	13.16	36.86	51.77	0.93
Miscellaneous cost (INR/hour)	NA	NA	NA	3.54
Total cost (INR/hour)	**191.69**	**395.37**	**501.54**	**34.58**
Total cost (INR/tonne)	**191.69**	**395.37**	**417.95**	**28.82**

Production cost of briquette (INR/tonne)	1,033.83			
Capital cost (principal)	80,000	4,90,000	10,95,000	4,84,376
Average use/yr (hours)	2,120	1,520	2,740	2,740
Salvage value	10% of principal			
Electricity tariff	INR 4.87/kWh			
Interest rate	12%			
Taxes and insurance	1% of principal			
Repair and maintenance	5% of principal			
Miscellaneous cost (INR/hr)	2% of principal			

Source: Authors' analysis; (Gill, Dogra, and Dogra 2018; Punjab State Electricity Regulatory Commission 2019)

Table A5: Economics of pellet production

Title	Chaff cutter	Chipper grinder	Pellet press	Operational shed
Depreciation + interest + tax + insurance (INR/hour)	6.26	53.51	55.33	16.62
Electricity required (kWh)	13.51	37.84	150.00	0.96
Electricity cost (INR/hour)	65.80	184.30	730.50	4.65
Labour charges (INR/hour)	104.58	104.58	261.45	
Repair and maintenance (INR/hour)	1.89	16.12	33.33	8.84
Cost of lubrication (INR/hour)	13.16	36.86	146.10	0.93
Miscellaneous cost (INR/hour)				3.54
Total cost (INR/hour)	**191.69**	**395.37**	**1,226.72**	**34.58**
Total cost (INR/tonne)	**191.69**	**395.37**	**817.81**	**28.82**

Production cost of briquette (INR/tonne)	1.433.69			
Capital cost (principal)	80,000	4,90,000	19,99,970	4,84,376
Average use/yr (hours)	2,120	1,520	6,000	2,740
Electricity tariff	INR 4.87/kWh			
Interest rate	12%			
Taxes and insurance	1% of principal			
Repair and maintenance	10% of principal			
Miscellaneous cost (INR/hr)	2% of principal			
Salvage value	10% of principal			

Source: Authors' analysis;(Purohit and Chaturvedi 2018; Punjab State Electricity Regulatory Commission 2019)

Table A6: Formulae used in the study

Farm to straw bank	
Parameter	**Formulae**
Chopper/ cutter-cum-spreader cost	$\frac{\text{(Rental + Labour + Fuel) X Field area (acre)}}{\text{Straw Produced in the field}} = \frac{293 \text{ X } 2.79}{7}$
Baler + raker cost	$\frac{\text{(Rental + Labour + Fuel) X Field area (acre)}}{\text{Straw Produced in the field}} = \frac{1367 \text{ X } 2.79}{7}$
Loading and unloading cost	2 X [(Man-hours required/tonne) X (Skilled labour charges/hour)] = 2×[3.48×52.29]
	$2 \text{ X } \left\{ \left[\frac{\text{(Distance from straw bank) X No. of trips required}}{\text{(average speed (km/hr))}} \text{ X (Tractor + Trailer rent/hr)} \right] + \right.$ [(Distance from straw bank) X (No. of trips required) X ((Diesel rate (per litre))/(Fuel mileage (km/l)] $+ \left. \left[\text{(Skilled labour charges/hour) X } \frac{\text{(Distance from straw bank) X No. of trips required}}{\text{(average speed (km/hr))}} \right] \right\}$ $= 2 \text{ X } \left\{ \left[\frac{10 \text{ X } 1}{15} \text{ X } 245 \right] + [10 \text{ X } 1 \text{ X } (72.21/0.71)] + \left[52.29 \frac{10 \text{ X } 1}{15} \right] \right\}$

Source: Authors' compilation

Farm to bank to end-user	
Parameter	**Formulae**
Residue cost	(Average farm to straw bank cost over 5 - 15kms) X 1.1(Moisture wight loss) X (Residue to Briquette or Pellet ratio) = 1228.63 X 1.1 X 1.3(Optional)
Storage cost	(Storage charges/tonne of residue) X (10% weight loss due to moisture) X (Residue of Briquette or Pellet ratio) = 800 X 1.1 X 1.3(Optional)
Processing cost (optional)	(Briquette processing cost/tonne) = 1003.83 (Pellet processing cost/tonne) = 1433.69
Transport charges	$\frac{\text{(Truck rental + Fuel + Driver(INR/km) X (Distance from straw bank to end-user)}}{\text{(Payload capacity for bales) or (Payload capacity for pellets and briquette)}} = \frac{40 \text{ X } 50}{5.77 \text{ or } 7}$
Loading and unloading cost	2 X [(Man-hours required/tonne) X (Skilled labour charges/hour)] = 2 X [3.48 X 52.29] (Bales) 2 X [(Labour charges/tonne)] = 2 X [67.32] (Briquette/Pellet)

Source: Authors' compilation



Image: iStock

Briquettes are preferred in large and medium-scale boilers, whereas pellets are widely burnt in small-sized devices, such as a pellet stove, furnace, and cooking range.



COUNCIL ON ENERGY, ENVIRONMENT AND WATER (CEEW)

Sanskrit Bhawan, A-10, Aruna Asaf Ali Marg
Qutab Institutional Area
New Delhi - 110 067, India
T: +91 11 4073 3300

info@ceew.in | ceew.in | @CEEWIndia

Multimedia Systems
https://doi.org/10.1007/s00530-020-00741-x

SPECIAL ISSUE PAPER

Leveraging crowd knowledge to curate documentation for agile software industry using deep learning and expert ranking

Akshi Kumar[1]

Received: 13 July 2020 / Accepted: 19 December 2020
© The Author(s), under exclusive licence to Springer-Verlag GmbH, DE part of Springer Nature 2021

Abstract
Agility gives the ability to be highly responsive to changes and make improvements in the way the agile project is documented. The just-in-time and just barely good enough documentation may miss out on important executable specifications. Interestingly, the frequently asked 'how-to-do' questions on the popular question answering (Q&A) websites like stack overflow are strong indicators of gaps in documentation and respondent answers can complement conventional software documentation practices. Social interaction within these QA websites generates partially structured content commonly referred to as crowd knowledge that can offer a peer-reviewed re-documentation by integrating the answers of these 'how-to-do' concerns. But finding the best, value-added answer to the question which can contribute toward an enriched and curated documentation is computationally difficult. Moreover, query duplicates can cause seekers to spend more time finding these best answers. As a solution, the research proffers a novel question-answering crowd documentation model (QACDoc) which is based on a socially mediated documentation mechanics involving the dynamics of community-based web. Firstly, duplicate questions are detected using Siamese neural architecture where two identical hierarchical attention networks are used to generate vectors for similarity matching. Semantic matching is done using Manhattan distance function, and a multi-layer perceptron is trained to output the predictions. Next, all respondents of semantically matched questions are grouped to form an intent-based crowd and lastly, top *k*-experts are identified using representative social presence features for expert ranking. The crowd documentation is then filtered to only include answers of these identified experts.

Keywords Crowd documentation · Question-answering websites · Agile software industry · Deep learning

1 Introduction

The software industry has significantly progressed in terms of open-source and digital collaboration technology use among businesses and stakeholders. Its impact on the global economy and across various industries can be gauged by the increase in innovations, technical progress, enhanced productivity, and the workforce. Undoubtedly, the software development practices have evolved to form a powerful arsenal of agile-oriented methodologies [1] which include 'application program interface,' 'user experience,' 'blockchain,' 'edge computing,' 'shorter feedback loops,' 'continuous integration, delivery and deployment,' 'cybersecurity,' among others [2–6]. These characterize some of the best practices adopted by software developers to successfully deliver bespoke projects that satisfy both the client and the end user.

Data curation is a long-established idea where subject matter experts (SMEs) make a collection of diverse data sources useful and integrated so that other users have rationalized data sets readily available to drive understanding and decision-making. At the same time, the current generation of ubiquitous and socially engaging web [7–9] has fostered open participation of volunteers to leverage the so-called 'wisdom of the crowd.' Organizations are contemplating the use of flexible on-demand workforce which supports an open innovation and collective intelligence ecosystem where customers are valuable co-creators of products and services. In particular, the term crowdsourcing has attracted significant attention and using the 'crowd' as an innovation partner is now an imperative component of the corporate innovation tool kit. The idea of crowdsourcing in various phases of software engineering process leads to seamless integration

✉ Akshi Kumar
akshikumar@dce.ac.in

[1] Department of Computer Science and Engineering, Delhi Technological University, New Delhi, India

of suitable experts into the software development process. It generally takes one of four distinct forms—contest, collaborative community, complementor or labor market—each best suited to a specific kind of challenge. Various crowdsourcing models have been applied to software development. Benefits of using crowdsourcing in software engineering include expansion of company's talent pool when there are insufficient internal resources or expertise to get a certain job done; reduce the time to market by modular tasking and generating a variety of solutions—effectively drawing on ideas from a range of people performing in parallel. Undeniably, crowdsourcing and curation are often complementary where wisdom of the crowd about data trust can be a great help to a data curator within a collaborative environment.

As a developer, it is always important to have reliable documentation to guide your work. Technical documentation in software engineering is the umbrella term that encompasses all written documents and materials dealing with software product development. Documentation exists to explain product functionality, unify project-related information and allow for discussing all significant questions arising between stakeholders and developers [10]. According to PMI's 9th Global Project Management Survey,[1] the agile approach is used by 71% of organizations for their projects in comparison to the conventional linear waterfall approach. The agile approach is based on teamwork, close collaboration with customers and stakeholders, flexibility and ability to quickly respond to changes. The basic building blocks of agile development are iterations; each one of them includes planning, analysis, design, development and testing. The agile method does not require comprehensive documentation at the beginning. Managers do not need to plan much in advance because things can change as the project evolves. This allows for just-in-time planning. The manifesto for agile software development [11] values 'a working software over comprehensive documentation' that is to produce documentation with information that is essential to move forward. A common agile practice is to defer the creation of all deliverable documentation as late as possible and create a just barely good enough (JBGE) document. But poor documentation is a key source of ambiguity and uncertainties. It is imperative for the documentation to be functional, updatable, flexible, transparent and easily accessible. Documentation errors can set gaps between the visions of stakeholders and engineers and as a result, a proposed solution would not meet stakeholder's expectations. Managerial issues in software documentation primarily include non-traceability, navigability, outdated and incomplete documentation with lack of good examples.

The development and popularity of question and answering (Q&A) websites as the gratis platform to search, provide and share knowledge are a practical use case which describes the opulence of crowdsourcing [12, 13]. As discussed, documentation is a key source of ambiguity and uncertainties and with the increased accessibility and availability of web, people seek advice and support from Q&A websites to report software application issues. The openness in sharing and learning, access to alternate perspectives, collaborations and interactions characterize the benefits of using a typical Q&A website. For example, stack overflow which is a popular Q&A website focusing on the topics specific to the domain of computer programming. As the largest, trusted, non-trivial toolset of programmers, its intent of use ranges from solving coding problems, developing new skills to finding job opportunities. Q&A websites can provide the apposite medium to identify gaps in application documentation as the frequently asked questions on these websites are strong indicators of predicament, challenges or unresolved issues faced by client and/or end users. Respondent answers can complement conventional software documentation practices as crowd knowledge contributes to a value-added answer to the question, thus facilitating an enriched and curated documentation. This partial documentation does not intend to substitute the formal documentation but intends peer-reviewed re-documentation by integrating the answers of 'how-to-do' concerns. Undoubtedly, getting user community involved in the documentation process can offer many benefits. Firstly, a peer-reviewed documentation would be created and secondly, the openness toward documentation improvement might help to reduce time to market and improve software quality in the current 'agile' software development generation. The social user community demands intelligent filters so that the external knowledge flow is harnessed in the best possible way.

But on the flip side, it is arduous and time-consuming to sift the pertinent information which includes finding preeminent answers and experts and also detecting duplicate questions for better user experience [14]. With the influx of questions and answers, high response time is expected, and the quality of answers is often compromised. Thus, finding the best, value-added answer to the question is computationally difficult. Duplicate content further corrupts the filtering mechanism. Query duplicates can cause seekers to spend more time finding these best answers making the task of selecting the documentation supporting answer challenging. In practice, stack overflow uses incentive-based and automated mechanisms to prevent duplicates but still excessive duplicates clutter up over time. Hence, mitigating information overload and filtering quality contributions from this shared knowledge base are a non-trivial task. The content and social interactions within the shared knowledge base need to be

[1] https://www.pmi.org/-/media/pmi/documents/public/pdf/learning/thought-leadership/pulse/pulse-of-the-profession-2017.pdf.

extracted for filtering quality contributions of the crowd. The aim should be that managers do not spend time finding the best answers but simply use the model to get the crowd expert's answer to complement documentation with concepts, purpose, usage scenarios and code examples. To clearly define the problem statement four research questions (RQs) were identified as follows:

- RQ1: How is semantic question matching done on stack overflow?
- RQ2: How is a virtual group of high-quality respondents on such Q&A websites modeled and how are the experts mined?
- RQ3: How can crowd experts be ranked to apprehend content which complements documentation?
- RQ4: Whether software developers find the crowd expert answer valuable?

As a solution, this research proffers a novel question-answering crowd documentation model (QACDoc) which is based on a socially mediated documentation mechanics involving the dynamics of community-based web. It aims to leverage crowd knowledge such that the documentation can be curated within the agile methodology of software development. The QACDoc model consists of three key modules, namely the semantic question pair matching module, intent-based group module and the expert mining module. Firstly, the duplicate questions are detected using deep Siamese network where two identical hierarchical attention networks (HAN) [15] are used to generate vectors for similarity matching. Semantic matching is done using Manhattan distance function and a multi-layer perceptron (MLP) is trained to output the predictions. This hybrid model is referred to as the HAN-Man-MLP model for sematic question pair matching. Subsequently, all respondents of semantically matched questions are grouped to form an intent-based crowd. Lastly, from this crowd, the respondents are ranked using representative social presence features (SPF) to finally identify top *k*-experts. The crowd documentation is then filtered to only include answers (opinion/suggestion/recommendation) of these identified experts. Thus, the primary contributions of this work are:

- Duplicate detection in stack overflow done using deep hybrid Siamese neural network built with two HANs and a MLP. Manhattan distance function is used for semantic similarity calculation.
- Creating an intent-based respondent crowd by grouping all respondents of semantically similar questions.
- Social expert ranking in crowd using profile features for including expert's answer to complement documentation.
- Investigate the effectiveness of value-added answers using crowd-sourced evaluation mechanism.

The rest of the paper is organized as follows: The next section, Sect. 2 discusses the background work within the domain of study. Section 3 illustrates the proposed QAC-Doc model with various sub-sections expounding the modules and working details of the model. Section 4 presents the results and confers a detailed discussion on the same followed by the last section, Sect. 5, which concludes the research. The potential future directions are also examined in the last section.

2 Background work

The agile development processes are adaptive rather than predictive and emphasize on just enough documentation, just-in-time documentation. Various studies have been reported on commonalities and differences in agile and document-driven approaches in requirement engineering and change management [16]. Ruping [17] reported that need to follow agile principles and still produce a light-weight documentation called the agile documentation. Selic [18] highlights the need of documentation owing to the complexity of modern software systems and the cryptic nature of current programming languages. Poniszewska-Marańda et al. [19] discuss the need of 'some documentation' in agile software development. Pertinent studies discuss about the right balance of documentation being complete and detailed and adaptive to the exceptionally dynamic development process.

Despite its success as a preferred or de-facto source of information, the web implicates two key challenges: firstly, to provide improved systems which retrieve the most relevant information available, and secondly, how to target search on information that satisfies user's need with accurate balance of novelty and relevance [20]. Information overload on the web has been a well-identified problem which has amplified with the advent of social web [21]. Detecting duplicate sentences is a well-known task in natural language processing (NLP). Deep learning models have achieved state-of-the-art results when applied to various NLP tasks with their power to adapt to new problems and perform automatic feature engineering [22]. Convolutional neural networks (CNNs), conventionally modeled for a variety of computer vision tasks such as image recognition and object classification, have been successfully adapted for NLP tasks [23]. Recurrent neural networks (RNNs) have also proven useful in NLP tasks to process sequential data (time-stamped) where a word's meaning in the sentence significantly relies on the context [24]. Further, both long short-term memory (LSTM) and gate recurrent unit (GRU) have been introduced with the goal of tracking long-term dependencies effectively while handling the vanishing/exploding gradient problems of Vanilla RNN. A typical extension of LSTMs, the bidirectional LSTMs (Bi-LSTMs)

have also been applied to sequence classification problems to improve the accuracy of models. Attention-based deep neural architectures facilitate sequential reasoning by processing subsets of the input in a sequential manner. Attention mechanisms can considerably improve the performance of RNNs. Yang et al. [15] developed a hierarchical attention network (HAN) for document classification that outperformed both RNNs and CNNs. Various deep learning architectures have been popularly used in semantic matching tasks. These use Siamese neural networks to detect duplicate sentences. Siamese neural networks take the two candidate sentences and encode them discretely using two similar architecture neural networks [25–27]. Primary studies on detecting query duplicates on quora [28–31] and on stack overflow [32–34] are available.

In recent years, the studies on expert finding in Q&A websites or community question answering (CQA) have increased. Zhou et al. [35] proposed a support vector machine model for connecting the gap between posted questions and potential user that could answer. Pal et al. [36] came up with new model called probabilistic models for finding the expert in CQA. Kumar et al. [37] had proposed a ComEx Miner System had used for ranking the blogs and then find the experts based on ranks of the blogs. In addition, work presented by Zhou et al. [38] proposed a topic-sensitive probabilistic model. In this model, they only considered the link structure technique rather than link analysis techniques. According to Bozzon et al. [39] vector space model gave the better result for finding the expert in the CQA. Chang et al. [40] also proposed an expert recommendation model. However, in this model, they had focused on routing a question to a group of experts rather than routing it to a particular expert. Yang et al. [41] proposed tag-based expert recommendation model. Rafiei et al. [42] developed a hybrid method for social network. Zhao et al. [43] developed a graph regularized matrix model for finding experts in CQA. Mandal et al. [44] proposed a query likelihood language (QLL) model for solving the problem by routing a novel question to the expert users that have the potential to answer the questions. Geerthik et al. [45] proposed a model called domain expert ranking model that identified the expert in the CQA. Yang et al. [46] proposed a NEWHITS model and compared it with traditional link analysis model. Alharthi et al. [47] proposed a method called linear prediction model. However, their model predicting the score of the questions. In 2017, Liu et al. [48] proposed a coupled semi-supervised mutual reinforcement-based label propagation (CSMRLP) model. Few studies on sentence similarity have been studied as a part of text summarization. The prominent ones include state-of-the-art LexRank [49] and maximal marginal relevance (MMR) [50].

Very less work has been done to comprehend the use of Q&A websites in improving documentation. In 2016, Treude et al. [51] presented supervised insight sentence extractor (SISE) approach to identify insightful sentences from stack overflow about missing components which can be augmented to API documentation. Delfim et al. [52] also worked on the re-documentation of APIs with knowledge generated by crowd in the form of stack overflow questions and answers.

3 The proposed QACDoc model

The proposed QACDoc model leverages the wisdom of the crowd to complement conventional software documentation practices to finally produce a curated, value-added documentation. Figure 1 depicts the working architecture of the proposed model.

As agile and DevOps have become an integral part of the digital transformation journey, value stream mapping which combines people, process and technology has emerged as a key practice to significantly enhance software delivery and customer experience and the proposed model endorses this mapping. The model works as follows: Firstly, intent matching of questions is done to identify similar intent respondents who form a virtual crowd followed by expert mining using social presence features of respondents of each group to rank the top experts. Lastly, the answers of these top experts are filtered to aid and add value to the documentation. The model is divided into three main modules, namely the semantic question pair matching module, intent-based group module and the expert mining module. The details of each module are giving the following sub-sections.

3.1 Dataset

The dataset used for this study is the stack overflow data dump given in the MSR 2015 challenge [53]. These data include the publicly available history of question and answer posts, tags, votes on the posts and the reputation of the users. The data set is divided into three files, namely posts, badges and users file. The 'how-to-do' questions are extracted from the posts using qualitative coding to a sample of questions as given by Treude et al. [51]

Duplicate questions in stack overflow are marked with a special marker. To collect these questions, the posts database file is parsed, the questions that are closed are identified as duplicates (whose *closedDate* value is not null) and considered as 'non-master' question. Next, the questions to which these non-master questions were linked are considered and are referred as 'master' question. The post links and the posts database files are parsed in order to collect the master questions. The non-master to master mapping is n:1 in the original dataset. However, in this work, hybrid Siamese neural network with MLP has been used for semantic question

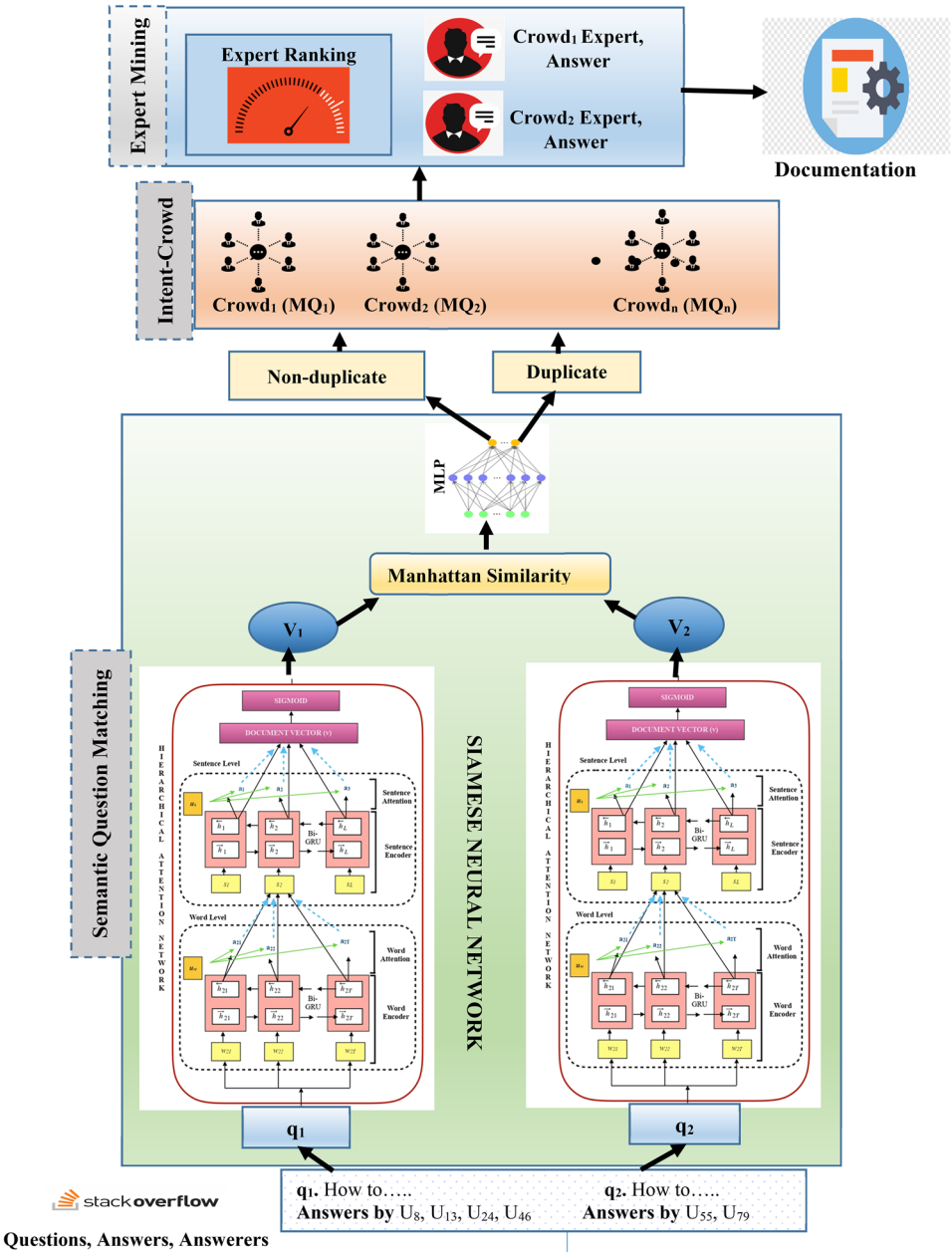


Fig. 1 Proposed QACDoc model for curated documentation

matching. It takes question pairs as input so the dataset has been modified to create as 1:1 mapping between the master and the non-master question or the master and master question. The further labeling of the dataset has been done as shown in Table 1.

The transaction files are built using the three XML files with the stack overflow dataset with relevant features required to build the QACDoc model.

3.2 Semantic question pair matching

The QACDoc model at its core targets to form an intent-based crowd which is a virtual group arranged by congregating respondents who answer similar questions. Thus, as a preliminary, it is imperative to collect question pairs that match semantically. Once the similar questions are grouped, all the respondents of these questions are put together. This way, the intent matching of questions is done to match similar intent respondents, thus forming a

Table 1 Dataset labels

$Question_1$	$Question_2$	Label
Master question (MQ)	Non-master (NMQ)	Duplicate
Master question (MQ)	Non-master (NMQ)	Non-duplicate

virtual crowd. In this research, a hybrid of Siamese neural network and multi-layer perceptron network is used to detect duplicate question pairs. The Siamese neural network contains two identical HANs. Manhattan distance metric is used for measuring similarity. This hybrid deep architecture is referred to the HAN-Man-MLP model for sematic question pair matching.

3.2.1 Hierarchical attention network (HAN)

The HAN model was put forward by Yang et al. [15] in 2017 for the purpose of classification of documents. HAN attempts to classify a document based on the knowledge it can infer about the document from its constituent parts, that is, the sentences and words that make up the document. It is based on the notion that a document has a hierarchical structure, i.e., words from various sentences, and then, those different sentences form the complete document. It takes this hierarchy into consideration while formulating the vector representation of the document. In order to accurately encapsulate the unerring thought of the document, the HAN model employs two distinct levels of attention mechanisms, one at the word level and the other at the sentence level. Different parts of a document contribute distinctively toward its true meaning. These unique contributions of the several disparate parts of the document along with the context-dependent significance of these parts are taken into account while formulating the vector representation of the document. The HAN model constitutes of an embedding layer, attention layers and encoders which together aids the model in comprehending the textual features. The encoders are responsible for the extraction of relevant context. The attention layers evaluate how relevant a sequence of tokens is with reference to the document. The framework of HAN is made up of five distinct layers, namely the embedding layer, word sequence encoder, word-level attention layer, sentence encoder and sentence-level attention layer. The next generation embeddings from language models (ELMo) word embedding [54] are used as the word vector learning technique to provide the classifier with the learned representations of the text. The word vectors generated by the embedding layer are fed to the word encoder which generates the annotations of those words. These annotations are then filtered by the attention mechanism based on their contribution to the essential meaning of the sentence. This whole process is repeated once again at the sentence level where the inputs are sentence vectors. Thus, the HAN classifier tackles each document at the word level and then again at the sentence level, finally forming a document vector. This document vector is then passed through a sigmoid activation function to generate the output. Figure 2 depicts the HAN architecture used in the Siamese network.

3.2.1.1 Embedding layer Word embeddings extract meaning from the text while preserving the contextual similarity of the words meaning that words with similar context end up having similar vector representations. It essentially provides us with a mapping of the learned representations of all the words in a document. In the text that is input, each word is encoded into an integer. Therefore, a unique integer represents each word of the input text. The embedding layer is initialized with random weights, and it learns the embedding for all of the words present in the training set. In our work, the ELMo 5.5B word embeddings model [54] is used to produce the vector representation for the words. ELMo was preferred over the conventional models such as Word2Vec, fastText or GloVe, because ELMo provided with contextualized word representations, which essentially means that the vector representation generated by ELMo for each word, entirely depends on the context in which the word has been used. It is possible for the exact same word to have two distinct vector representations depending on the context in which it is used. ELMo produces vectors on the spot by feeding words into the deep learning model instead of having a dictionary of words and their matching vectors, as is the case with conventional word embedding models. Additionally, ELMo representations are purely character-based, which permits our model to generate representations for words that are absent from the training set.

3.2.1.2 Encoder A GRU based sequence encoder is used in HAN. The GRU uses a gating mechanism without employing any separate cells for memory to track the sequences of the state. It comprises of the following two gates: the update gate z_t and the reset gate r_t. The reset gate regulates the amount of contribution provided by the previous state to the current state. The update state z_t defines the extent of the past information to be added and also of the new information to be added to the current state. Both, the reset gate r_t and the update gate z_t jointly control the updating of information in the current state.

At any time, t, the new state h_t is calculated as given in Eq. (1):

$$h_t = \left(1 - z_t\right) \odot h_{t-1} + z_t \odot \tilde{h}_t, \tag{1}$$

where h_{t-1} stands for the previous state, and $\tilde{h}$ stands for the current state.

The update state z_t is computed as given in Eq. (2):

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z), \tag{2}$$

where x_t stands for the sequence vector at any time t.

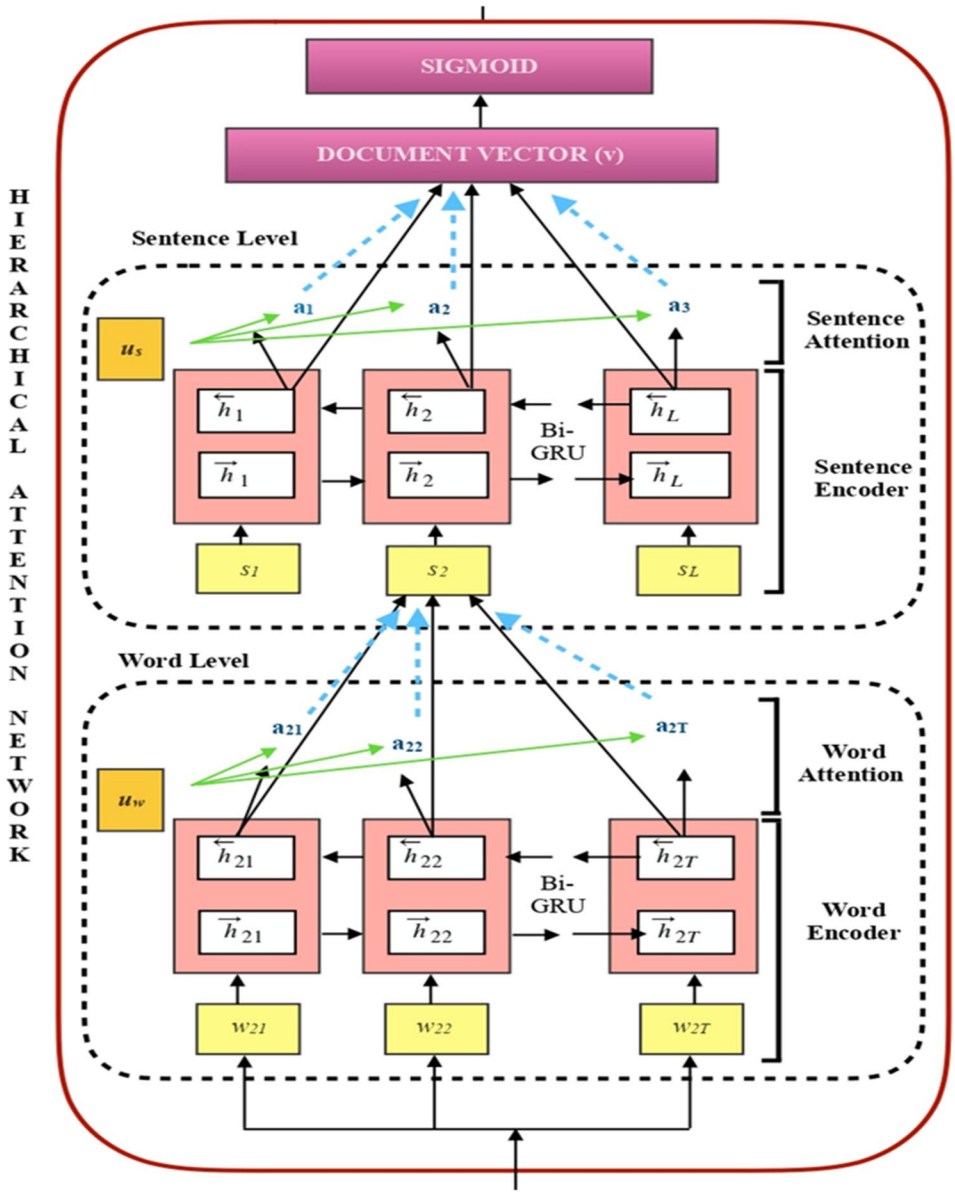
The candidate state $\tilde{h}_t$ is computed as given in Eq. (3):

$$\tilde{h}_t = \tanh(W_h x_t + r_t \odot (U_h h_{t-1}) + b_h), \tag{3}$$

where r_t stands for the reset gate.

The reset gate is computed as using Eq. (4):

Fig. 2 HAN architecture



$$r_t = \sigma\left(W_r x_t + U_r h_{t-1} + b_r\right). \tag{4}$$

- Word encoder

  A mapping of discrete variables to a vector of continuous numbers is termed as an embedding matrix. Neural network embeddings prove to be utilitarian as they are able to lower the dimensionality of categorical variables and pertinently represent categories in the transformed space. Provided a sentence with words w_{it}, $t \in [0, T]$, it starts by embedding the words to vectors with the help of an embedding matrix W_e, $x_{it} = W_e w_{it}$.

  The information is summarized from both the directions, namely forward and backward, using a bidirectional GRU in order to obtain annotations of the words. The forward GRU $\vec{f}$ reads the sentence s_i from w_{i1} to w_{iT} as given in Eq. (5):

$$\vec{h}_{it} = \overrightarrow{\mathrm{GRU}}\left(x_{it}\right), t \in [1, T]. \tag{5}$$

  The backward GRU $\overleftarrow{f}$ reads the sentence s_i from w_{iT} to W_{i1}. The concatenation of the backward hidden state and the forward hidden state helps us in acquiring the annotation of the word w_{it}, i.e., $h_{it} = [\vec{h}_{it}, \overleftarrow{h}_{it}]$ as given in Eq. (6):

$$\overleftarrow{h}_{it} = \overleftarrow{\mathrm{GRU}}\left(x_{it}\right), t \in [T, 1]. \tag{6}$$

- Sentence encoder

  Analogous to a word encoder, the sentence encoder is utilized to obtain the document vector from the given sentence vectors s_i. A bidirectional GRU is employed for encoding the sentence as given in Eqs. (7) and (8):

$$\vec{h}_i = \overrightarrow{\mathrm{GRU}}(s_i), i \in [1, L], \tag{7}$$

$$\overleftarrow{h}_i = \overleftarrow{\mathrm{GRU}}(s_i), t \in [L, 1]. \tag{8}$$

  To get the annotation of a sentence i, both the backward hidden state h_j and the forward hidden state h_i, are concatenated, i.e., $h_i = [\vec{h}_i, \overleftarrow{h}_i]$.

3.2.1.3 Attention layer The hierarchical attention layers are used for the document level classification. Suppose the document has L sentences s_i, and each sentence s_i contains T_i words. The notion of 'attention' is based on the fact the words in a sentence do not contribute equally to its meaning. Similarly, not all sentences contribute equally toward the overall meaning of the document. The word-level and sentence-level attention mechanism is described next.

- Word attention layer

  In trying to understand the meaning of the sentences, it becomes clear that all the words present in a sentence do not contribute equally to its meaning. Therefore, to find out those words which are key to the meaning of the sentences, the word attention layer is used. The extracted words, then, form the sentence vector. To get the hidden representation of the h_{it}, u_{it} is calculated using a one-layer MLP using Eq. (9):

$$u_{it} = \tanh\left(W_w h_{it} + b_w\right). \tag{9}$$

  The normalized importance a_{it} is then generated through a sigmoid function to calculate the significance of the word as a similitude of u_{it} with a word-level context vector u_w using Eq. (10):

$$a_{it} = \frac{\exp\left(u_{it}^T u_w\right)}{\sum_t \exp\left(u_{it}^T u_w\right)}. \tag{10}$$

  After that, the weighted sum of the word annotations based on their weights is computed which corresponds to the sentence vector s_i as given in Eq. (11):

$$s_i = \sum_t a_{it} h_{it}. \tag{11}$$

- Sentence attention layer

  Since all the sentences in the document are not important to understand the meaning of the document, it is necessary to extract such sentences that are of more importance compared to the others. To get the hidden representation of h_i, u_i is calculated using a one-layer MLP as given in Eq. (12):

$$u_i = \tanh\left(W_s h_i + b_s\right). \tag{12}$$

  The normalized importance a_i is then generated through a sigmoid function to calculate the significance of the sentence as a similitude of u_i with a sentence-level context vector u_s using Eq. (13):

$$a_i = \frac{\exp\left(u_i^T u_s\right)}{\Sigma_i \exp\left(u_i^T u_s\right)}. \tag{13}$$

  After that, the document vector v_i is computed, which is the weighted sum of the sentence annotations depending on their weights. The summarization of all the information gathered from all the sentences of a document is present in the document vector v_i. as given in Eq. (14):

$$v_i = \sum_i a_i h_i. \tag{14}$$

3.2.1.4 Document classification Finally, the classification of the document into one of the two given classes is done using the document vector v. The document vector v passes through the last layer, i.e., the output layer, which employs a sigmoid function as the semantic matching problem is a typical binary classification problem. This leads to the generation of the final output of the HAN classifier, given by Eq. (15):

$$p = \text{ sigmoid } \left(W_c v + b_c\right). \tag{15}$$

The negative of the log-likelihood of the correct labels is taken as the training loss using Eq. (16):

$$L = -\sum_d \log p d_j, \tag{16}$$

where j stands for the label of document d.

3.2.2 Similarity distance function

Varied similarity metrics (cosine, Euclidean and Manhattan distances) were assessed and the best results were achieved using the Manhattan distance as given in Eq. (17):

$$\exp\left(-||v_1 - v_2||\right) \in [0, 1], \tag{17}$$

where v_1 and v_2 represent the outputs of each sub-network. The exponential function on the negative value of distance

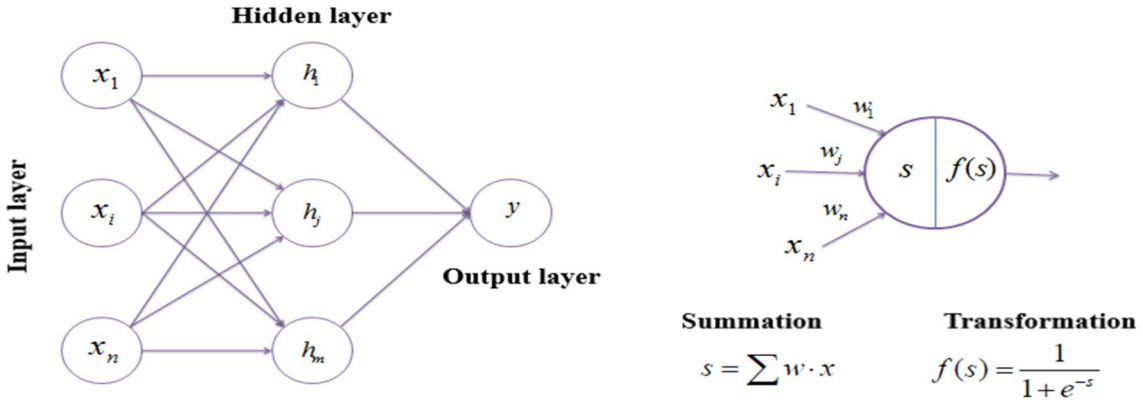


Fig. 3 MLP architecture [55]

measured by Manhattan distance function is thus used to find the similarity distance between the vector representations.

3.2.3 Output layer

The MLP model with a single hidden layer of nodes is used in the output layer to make the final prediction. A multi-layer perceptron is a deep, artificial neural network. It consists of multiple perceptrons. A perceptron on its own is merely a building block that turns out to be useful only in larger functions such as the multi-layer perceptron. An MLP is essentially a linear classifier but it is also adept in approximating various non-linear functions as well. It is deemed as a linear classifier because it can accomplish the task of binary classification by separating two distinct classes or entities from each other with the aid of a straight line. A perceptron takes a feature vector x as its input, which is multiplied to a weight w before adding a bias b to it as given in Eq. (18):

$$y = w * x + b. \tag{18}$$

A perceptron is a shallow neural network and has limited capabilities when it comes to classification tasks. The perceptron was designed to perform binary classification tasks only. Classification problems containing more than two classes are above the level of comprehension of a perceptron. Various inputs can be provided to a perceptron and it produces an output by forming a linear combination by utilizing its input weights. Sometimes, it passes the output through a non-linear activation function. This can be expressed as given in Eq. (19):

$$y = \varphi\left(\sum_{i=1}^{n} w_i x_i + b\right), \tag{19}$$

where w denotes the weight vector, x denotes the input vector, b denotes the bias and the symbol phi stands for the non-linear activation function.

A MLP typically consists of an input layer, an output layer and a fixed number of hidden layers depending on the task. Figure 3 depicts the typical layers of MLP [53].

Therefore to output the probability of a match between the two questions as duplicate or non-duplicate, the MLP takes the concatenation of the high-level vector representations and their distance. That is, the final output layer consists of a single (1×2) vector, with the two elements corresponding to each of the two classes: non-duplicate and duplicate and *argmax* is taken over this vector to produce the final prediction.

3.3 Intent-based crowd

Formally, a crowd is a set of all respondents who answer questions with similar intent, where intent is matched using the semantic matching module. That is, if q_1 is the master question MQ_1 and q_2, q_3, q_4 are its duplicates then all the respondents of q_1, q_2, q_3 and q_4 are combined to form this similar intent virtual crowd as shown in Fig. 4. The crowd is labeled with the < tag > of the master question. The social presence features of these grouped respondents are then extracted to identify the top k-experts in each crowd. Ranking criteria have been used to find the top experts.

Table 2 depicts the sample crowd formation, where each user who responds to semantically similar question of the master question.

3.4 Expert mining (expertRank score)

Each crowd is built with users who answer semantically matched question and now for each user in a crowd his/her social presence features (SPF) can be extracted to mine crowd experts and rank them. SPF defines the profile features of user which index his social existence and activity

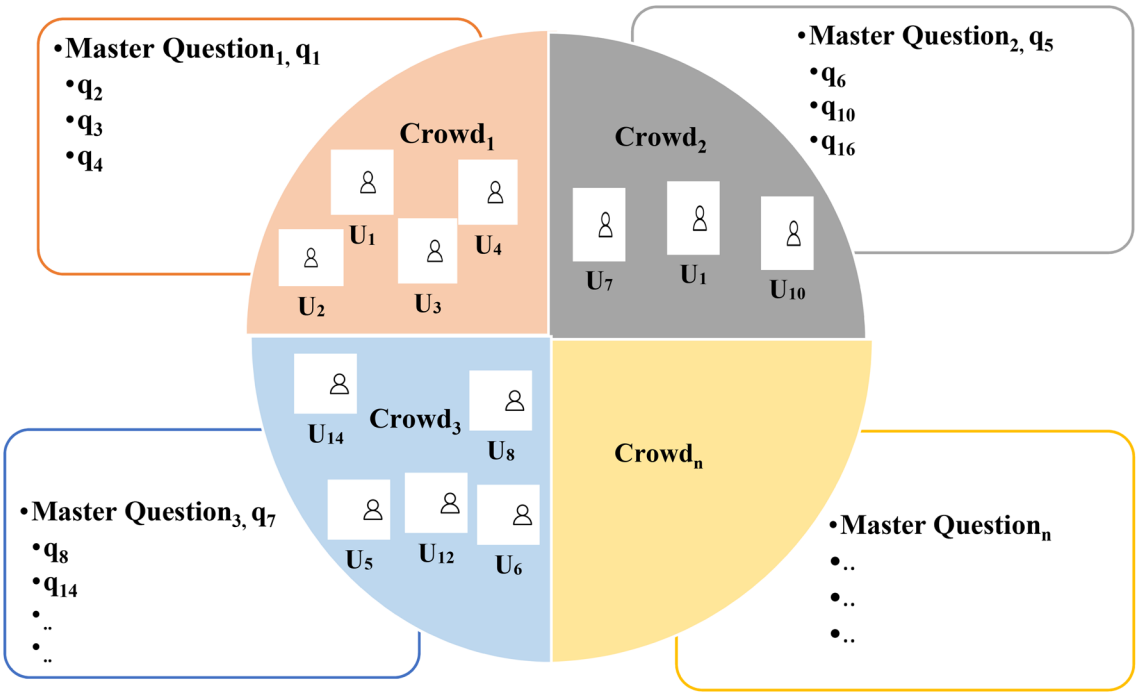


Fig. 4 Intent-based crowd

Table 2 Sample crowd formation

Master question$_1$	
Crowd$_1$	User$_1$
	User$_2$
	User$_3$
	User$_4$
	User$_5$
Master question$_2$	
Crowd$_2$	User$_1$
	User$_7$
	User$_{10}$
Master question$_n$	
Crowd$_n$	User$_x$

in the Q&A website. A transaction file with the following features relevant for identification of expert is made. For each used Id, unique to a user profile we define:

- *Reputation:* Reputation is the extent to which the user is trusted by the community of stack overflow.
- *Up_vote:* Number of votes given by users if the answer is relevant and accurate to the asked query.
- *Down_vote:* The number of votes given by users for irrelevant or inaccurate answer.
- *Tenure:* The time duration since the user is member of stack overflow.
- *No. of answers:* The total number of answers given by the user on stack overflow. The activity of user and that too a consistent one is an important indicator.
- *Answers badges:* Badges earned by a user for being helpful by answering the queries of other users. That is, besides gaining reputation with questions and answers, users receive badges for being especially helpful. Badges appear on the profile page, flair and your posts. There are three types of badges, i.e., bronze, silver and gold badges. Users can receive these badges within various categories defined: question badges, answer badges, participation badges, moderation badges, documentation badges and other badges. In this research, we consider on answer badges and our primary intent is to collate respondents and find experts. Also, higher weightage has been given to gold badge holder as it recognizes important contributions from members of the community and is rarely awarded. Likewise, silver badges are given lesser weightage than gold and higher than bronze badges.
- *People reached:* Number of views on all the posts of the user. This is a direct marker of user's social presence on the stack overflow.

Using these features, the *expertRank* score of each respondent user in the crowd is calculated using Eq. (20) as given:

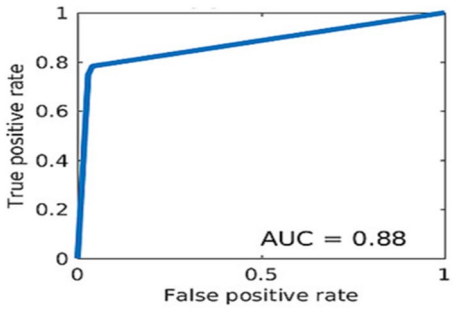


Fig. 5 AUC-ROC curve of HAN-Man-MLP model

$$expertRank\text{score} = \sum_{i=1}^{n} \frac{(up_i - down_i) * \text{PT}}{n^2 * \delta} + \frac{R * (100G + 10S + B)}{T^2}, \quad (20)$$

where n: no. of answers, up_i: number of up_votes given by users to answer i, $down_i$: number of down_votes given by users to answer i, PT: total people reached as per user profile, δ is the damping factor set to 10,000 (as the people's reach is in millions), R: reputation of user, G: gold answer badges, S: silver answer badges, B: bronze answer badges and T: tenure of user.

Once the rank list using this *expertRank* score is generated, the top $k=3$ experts are extracted. Answers of these identified experts are then included as crowd documentation. Their answers to the 'how-to-do' concepts, purpose, usage scenarios and code examples can be a value addition to the documentation. The next section elaborates the results of the QACDoc model.

4 Results and discussion

In this work, we propose a model to detect duplicate question in stack overflow to further form an intent-based crowd of user profiles. User profile modeling is done to extract the features of each respondent in the crowd and further ranking the user profiles based on *expertRank* score is done to find the top k-experts for each crowd. The results have been shown in terms of: (i) performance of HAN-Mann-MLP model for sematic question pair matching (ii) *expertRank* score evaluation (iii) effectiveness of value-added answers that complement documentation. The following sub-sections present the results.

4.1 Performance of semantic question pair matching

The performance of the semantic matching module has been evaluated in terms of AUC-ROC curve. The AUC-ROC curve is an elementary metric for analyzing binary classification problems. It helps to understand the degree to which a learning model is able to distinguish between the given classes. The value of AUC ranges from 0.5 to 1. If the AUC is closer to 1, the model has a high potential for separating classes from one another. On the other hands, if the AUC is 0.5, the model has zero potential for segregating the given classes. The AUC score attained by the model is 0.88. The ROC curve of the deep learning-based semantic matching module is shown in Fig. 5.

The semantic matching module is also evaluated using Euclidean distance and cosine distance instead of Manhattan distance function. The AUC score attained using Cosine is 0.79 and that by Euclidean is 0.82. The respective ROC curves are given in Fig. 6.

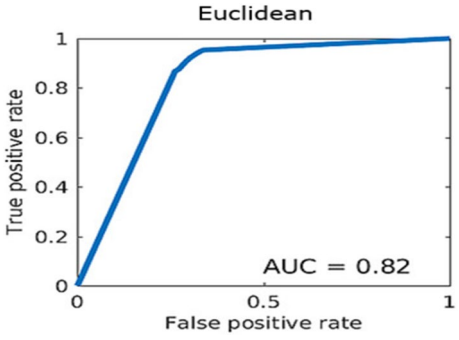


Fig. 7 Performance accuracy of ranking mechanisms

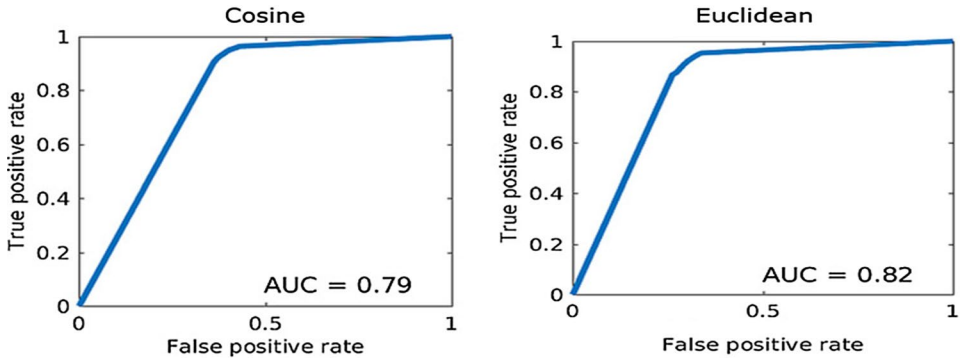


Fig. 6 AUC-ROC curve of HAN-*-MLP model

Table 3 Crowd-sourced evaluation

	Java API (%)	Rest API (%)	Twitter API (%)	Android API (%)
Relevant, accurate, adds information (RAV)	45	42	54	58
Relevant, accurate, no added information (RANV)	36	29	24	31
More context required (CON)	12	18	16	8
No sense (NS)	7	11	6	3

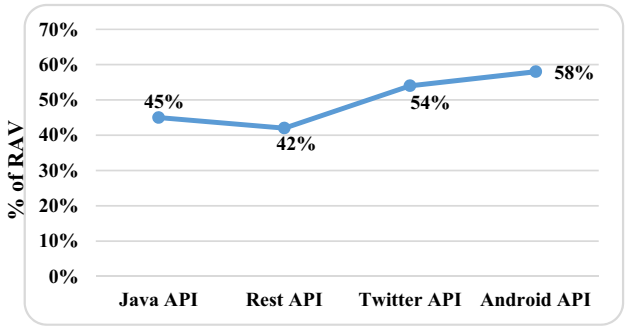


Fig. 8 Relevant, accurate and value-added answers (RAV) as per crowd-sourced evaluation

4.2 expertRank score evaluation

The *expertRank* score has been characterized using the social presence features of crowd members. We assess the performance accuracy of ranking mechanism by comparing the results with ML-kNN [56] and rankSVM [57] models. In stack overflow, expertise of a user is recognized through a reputation system. This reputation scheme provides easy identification of experts as users with high reputation [58]. The proposed *expertRank* score performs reasonably well in comparison to rankSVM and has minor improvement in contrast to ranking based on ML-*k*NN. The accuracy results are shown in Fig. 7.

4.3 Effectiveness of value-added answers to developers for documentation

The final research question was to find whether software developers find the crowd expert answer valuable or not. Thus, to determine the effectiveness of the proposed QACDoc model, crowd sourcing-based evaluation was done. Four crowd tags are considered to carry the study, Java API, Rest API, Twitter API and Android API. As participants, 20 Delhi Technological University alumni who are currently working as software developers in companies like Amazon, Oracle, Flipkart and Samsung were selected. Each participant was shown the answers of the top *k*-experts belonging to the respective crowd tags. For the respective < tag > type, the participants were asked to choose one of the following options:

- The answer is relevant, accurate and adds useful information not found in the documentation.
- The answer is relevant and accurate but does not add information to the documentation (already exists in the documentation).
- The answer requires more context to understand.
- The answer does not make sense to me.

Table 3 shows the results of the crowd-sourced evaluation. All the values are expressed in percentage.

Figure 8 represents the percentage for answers that are found relevant, accurate and add value (RAV) to the documentation for each API. An average of 49.75% answers provided by the crowd experts is relevant and accurate. This clearly demonstrates the how crowd knowledge and expert mining within a real-time social setting can complement conventional software documentation practices.

The results of value-added answers generated by the proposed QACDoc model were also compared with the existing techniques used to detect useful insightful sentences. As shown in Table 4, the QACDoc model outperforms the existing techniques by identifying the most relevant and accurate information for value addition in the API documentation. Also, 30% of the information was relevant and accurate but did not have any additional value to the documentation, which implies it was

Table 4 Comparison with existing techniques

	LexRank [47] (%)	MMR [48] (%)	SISE [49] (%)	QACDoc (%)
Relevant, accurate, adds information	22.5	32.5	47.5	49.75
Relevant, accurate, no added information	12.5	10.0	22.5	30
More context required	37.5	42.5	15.0	13.5
No sense	27.5	15.0	15.0	6.75

already existent in the documentation. It is therefore interesting to note that nearly 80% of the answers extracted were relevant, accurate and useful to API documentation, some of which were already existent and some provided value addition.

5 Conclusion

The agile manifesto values documentation, but it values working software more, where agile documentation is simply an approach to create concise documents that serve the situation at hand. But missed out feature or a quirky API call that is not thoroughly documented can cause all sorts of support problems. The frequently asked 'How-to-do' questions strongly indicate missing links in documentation. But the overload of questions, answers and users on Q&A websites like stack overflow make it challenging and expensive to filter knowledge which can complement documentation. This research builds a virtual crowd by mining a subset of respondents who answer similar questions and finds experts in the crowd to extract their answer adding value to documentation. The work demonstrates the ability of deep learning to match semantic similarity in question of stack overflow. This semantic matching is a preliminary step to identify and extract all respondents with similar intent typified by the fact that they answer similar questions. The implementation of hierarchical attention Siamese network with Manhattan similarity is used to train the MLP for similarity matching. Virtual crowd with respondents is then built, and a ranking mechanism is proposed by modeling the user profile. This way a two-level intelligent filter is implemented, where firstly, duplicate questions are combined to reduce question overload and secondly, experts are mined and ranked to filter top experts and their answers to improve documentation. The results of the proposed QACDoc model are motivating and validated with crowd-sourced evaluation method. It is important to note that the value-added answers partially offer a socially mediated documentation and not a replacement to the technical documentation. Pre-trained networks and transfer learning of expertise between different knowledge domains can be examined as a possible future direction. Also, integrating feedbacks, especially implicit ones need to be investigated.

References

1. Huber, T.L., Winkler, M.A., Dibbern, J., Brown, C.V.: The use of prototypes to bridge knowledge boundaries in agile software development. Inf. Syst. J. **30**(2), 270–294 (2020)
2. Abou-Nassar, E.M., Iliyasu, A.M., El-Kafrawy, P.M., Song, O.Y., Bashir, A.K., Abd El-Latif, A.A.: DITrust chain: towards blockchain-based trust models for sustainable healthcare IoT systems. IEEE Access **8**, 111223–111238 (2020)
3. Lou, K., Yang, Y., Wang, E., Liu, Z., Baker, T., Bashir, A.K.: Reinforcement Learning Based Advertising Strategy Using Crowdsensing Vehicular Data. IEEE Trans. Intell. Transp. Syst. (2020). https://doi.org/10.1109/TITS.2020.2991029
4. Rafique, H., Almagrabi, A.O., Shamim, A., Anwar, F., Bashir, A.K.: Investigating the acceptance of mobile library applications with an extended technology acceptance model (TAM). Comput. Educ. **145**, 103732 (2020)
5. Abbasi, M., Shokrollahi, A., Khosravi, M.R., Menon, V.G.: High-performance flow classification using hybrid clusters in software defined mobile edge computing. Comput. Commun. **160**, 643–660 (2020)
6. Menon, V.G., Jacob, S., Joseph, S., Almagrabi, A.O.: SDN powered humanoid with edge computing for assisting paralyzed patients. IEEE Internet Things J. **7**(7), 5874–5881 (2019). https://doi.org/10.1109/JIOT.2019.2963288
7. Kumar, A., Jaiswal, A.: A deep swarm-optimized model for leveraging industrial data analytics in cognitive manufacturing. IEEE Trans. Ind. Inf. (2020). https://doi.org/10.1109/TII.2020.3005532
8. Tripathi, A.K., Sharma, K., Bala, M., Kumar, A., Menon, V.G., Bashir, A.K.: A parallel military dog based algorithm for clustering big data in cognitive industrial internet of things. IEEE Trans. Ind. Inf. (2020). https://doi.org/10.1109/TII.2020.2995680
9. Jain, D.K., Kumar, A., Sharma, V.: Tweet recommender model using adaptive neuro-fuzzy inference system. Future Gener. Comput. Syst. **112**, 996–1009 (2020)
10. Parnas, D.L.: Precise documentation: the key to better software. In: The future of software engineering, pp. 125–148. Springer, Berlin (2011)
11. Poth, A., Sasabe, S., Mas, A., Mesquida, A.L.: Lean and agile software process improvement in traditional and agile environments. J. Softw. Evol. Process **31**(1), e1986 (2019)
12. Kumar, A., Sangwan, S.R.: Expert finding in community question-answering for post recommendation. Int. J. Eng. Technol. **7**(34), 151–159 (2018)
13. Anderson, A., Huttenlocher, D., Kleinberg, J., Leskovec, J.: Discovering value from community activity on focused question answering sites: a case study of stack overflow. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining, pp. 850–858 (2012)
14. Hassenzahl, M.: User experience (UX) towards an experiential perspective on product quality. In: Proceedings of the 20th conference on l'Interaction Homme-machine, pp. 11–15 (2008)
15. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, pp. 1480–1489 (2016)
16. Sillitti, A., Ceschi, M., Russo, B., Succi, G.: Managing uncertainty in requirements: a survey in documentation-driven and agile companies. In: 11th IEEE international software metrics symposium (METRICS'05), p. 10 (2005)
17. Rüping, A.: Agile documentation: a pattern guide to producing lightweight documents for software projects. John Wiley & Sons, Hoboken (2005)
18. Selic, B.: Agile documentation, anyone? IEEE Softw. **26**(6), 11–12 (2009)
19. Poniszewska-Marańda, A., Zieliski, A., Marańda, W.: Towards project documentation in agile software development methods. In: Data-centric business and applications, pp. 1–18. Springer, Cham (2020)

20. Bhatia, M.P.S., Khalid, A.K.: Contextual proximity based term-weighting for improved web information retrieval. In: International conference on knowledge science, engineering and management, pp. 267–278. Springer, Berlin, Heidelberg (2007)
21. Kumar, A., Sangwan, S.R.: Rumor detection using machine learning techniques on social media. In: International conference on innovative computing and communications, pp. 213–221. Springer, Singapore (2019)
22. Young, T., Hazarika, D., Poria, S., Cambria, E.: Recent trends in deep learning based natural language processing. IEEE Comput. Intell. Mag. **13**(3), 55–75 (2018)
23. Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., et al.: Recent advances in convolutional neural networks. Pattern Recognit. **77**, 354–377 (2018)
24. Otter, D.W., Medina, J.R., Kalita, J.K.: A survey of the usages of deep learning for natural language processing. IEEE Trans. Neural Netw. Learn. Syst. (2020). https://doi.org/10.1109/TNNLS.2020.2979670
25. Kumar, A.: Using cognition to resolve duplicacy issues in socially connected healthcare for smart cities. Comput. Commun. **152**, 272–281 (2020)
26. Bogdanova, D., dos Santos, C., Barbosa, L., Zadrozny, B.: Detecting semantically equivalent questions in online user forums. In: Proceedings of the nineteenth conference on computational natural language learning, pp. 123–131 (2015)
27. Bromley, J., Bentz, J.W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., et al.: Signature verification using a "siamese" time delay neural network. Int. J. Pattern Recognit. Artif. Intell. **7**(04), 669–688 (1993)
28. Wang, G., Gill, K., Mohanlal, M., Zheng, H., Zhao, B.Y.: Wisdom in the social crowd: an analysis of quora. In: Proceedings of the 22nd international conference on world wide web, pp. 1341–1352 (2013)
29. Saedi, C., Rodrigues, J., Silva, J., Branco, A., Maraev, V.: Learning profiles in duplicate question detection. In: 2017 IEEE international conference on information reuse and integration (IRI), pp. 544–550 (2017)
30. Rodrigues, J., Saedi, C., Maraev, V., Silva, J., Branco, A.: Ways of asking and replying in duplicate question detection. In: Proceedings of the 6th joint conference on lexical and computational semantics (* SEM 2017), pp. 262–270 (2017)
31. Chen, Z., Zhang, H., Zhang, X., Zhao, L.: Quora question pairs. (2018). http://static.hongbozhang.me/doc/STAT_441_Report.pdf
32. Zhang, Y., Lo, D., Xia, X., Sun, J.L.: Multi-factor duplicate question detection in stack overflow. J. Comput. Sci. Technol. **30**(5), 981–997 (2015)
33. Ahasanuzzaman, M., Asaduzzaman, M., Roy, C.K., Schneider, K.A.: Mining duplicate questions of stack overflow. In: 2016 IEEE/ACM 13th working conference on mining software repositories (MSR), pp. 402–412 (2016)
34. Silva, R.F., Paixão, K., de Almeida Maia, M.: Duplicate question detection in stack overflow: a reproducibility study. In: 2018 IEEE 25th international conference on software analysis, evolution and reengineering (SANER), pp. 572–581 (2018)
35. Zhou, T.C., Lyu, M.R., King, I.: A classification-based approach to question routing in community question answering. In: Proceedings of the 21st international conference on world wide web, pp. 783–790 (2012)
36. Pal, A., Harper, F.M., Konstan, J.A.: Exploring question selection bias to identify experts and potential experts in community question answering. ACM Trans. Inf. Syst. **30**(2), 1–28 (2012)
37. Kumar, A., Ahmad, N.: ComEx miner: Expert mining in virtual communities. Int. J. Adv. Comput. Sci. Appl. **3**(6) (2012). https://doi.org/10.14569/IJACSA.2012.030610
38. Zhou, G., Lai, S., Liu, K., Zhao, J.: Topic-sensitive probabilistic model for expert finding in question answer communities. In: Proceedings of the 21st ACM international conference on information and knowledge management, pp. 1662–1666 (2012)
39. Bozzon, A., Brambilla, M., Ceri, S., Silvestri, M., Vesci, G.: Choosing the right crowd: expert finding in social networks. In: Proceedings of the 16th international conference on extending database technology, pp. 637–648 (2013)
40. Chang, S., Pal, A.: Routing questions for collaborative answering in community question answering. In: 2013 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM 2013), pp. 494–501 (2013)
41. Yang, B., Manandhar, S.: Tag-based expert recommendation in community question answering. In: 2014 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM 2014), pp. 960–963 (2014)
42. Rafiei, M., Kardan, A.A.: A novel method for expert finding in online communities based on concept map and PageRank. Hum. Centric Comput. Inf. Sci. **5**(1), 10 (2015)
43. Zhao, Z., Zhang, L., He, X., Ng, W.: Expert finding for question answering via graph regularized matrix completion. IEEE Trans. Knowl. Data Eng. **27**(4), 993–1004 (2014)
44. Mandal, D.P., Kundu, D., Maiti, S.: Finding experts in community question answering services: a theme based query likelihood language approach. In: 2015 International conference on advances in computer engineering and applications, pp. 423–427 (2015)
45. Geerthik, S., Gandhi, K.R., Venkatraman, S.: Domain expert ranking for finding domain authoritative users on community question answering sites. In: 2016 IEEE international conference on computational intelligence and computing research (ICCIC), pp. 1–5 (2016)
46. Yang, J., Peng, S., Wang, L., Wu, B.: Finding experts in community question answering based on topic-sensitive link analysis. In: 2016 IEEE first international conference on data science in cyberspace (DSC), pp. 54–60 (2016)
47. Alharthi, H., Outioua, D., Baysal, O.: Predicting questions' scores on stack overflow. In: 2016 IEEE/ACM 3rd international workshop on crowd sourcing in software engineering (CSI-SE), pp. 1–7 (2016)
48. Liu, J., Shen, H., Yu, L.: Question quality analysis and prediction in community question answering services with coupled mutual reinforcement. IEEE Trans. Serv. Comput. **10**(2), 286–301 (2015)
49. Erkan, G., Radev, D.R.: Lexrank: graph-based lexical centrality as salience in text summarization. J. Artif. Intel. Res. **22**, 457–479 (2004)
50. Boudin, F., El-Bèze, M., Torres-Moreno, J.M.: A scalable MMR approach to sentence scoring for multi-document update summarization. In: Coling 2008: companion volume: posters, pp. 23–26 (2008)
51. Treude, C., Robillard, M.P.: Augmenting api documentation with insights from stack overflow. In: 2016 IEEE/ACM 38th international conference on software engineering (ICSE), pp. 392–403 (2016)
52. Delfim, F.M., Paixão, K.V., Cassou, D., de Almeida Maia, M.: Redocumenting APIs with crowd knowledge: a coverage analysis based on question types. J. Braz. Comput. Soc. **22**(1), 9 (2016)
53. Ying, A.T.: Mining challenge 2015: comparing and combining different information sources on the stack overflow data set. In: The 12th working conference on mining software repositories (2015)
54. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. arXiv preprint arXiv:1802.05365 (2018)
55. Kumar, A., Sachdeva, N.: Multi-input integrative learning using deep neural networks and transfer learning for cyberbullying detection in real-time code-mix data. In: Multimedia systems. Springer, Berlin (2020). https://doi.org/10.1007/s00530-020-00672-7
56. Zhang, M.L., Zhou, Z.H.: ML-KNN: a lazy learning approach to multi-label learning. Pattern Recognit. **40**(7), 2038–2048 (2007)

57. Cao, Y., Xu, J., Liu, T.Y., Li, H., Huang, Y., Hon, H.W.: Adapting ranking SVM to document retrieval. In: Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, pp. 186–193 (2006)
58. Singh, A.K., Nagwani, N.K., Pandey, S.: RANKuser: a folksonomy and user profile based algorithm to identify experts in community question answering sites. Data Technol. Appl. **52**(3), 329–350 (2018)

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

Low Fidelity VTOL UAV Design Optimization Using an Open Source Framework

Nikhil Sethi
Mechanical Engineering Undergraduate
(Delhi Technological University)

Saurav Ahlawat
Applied Physics Undergraduate
(Delhi Technological University)

Prof. N. S. Raghava
HoD ECE Department
(Delhi Technological University)

NOMENCLATURE

W	Weight	P	Power
T	Thrust	ROC	Rate of Climb
$GTOW$	Gross Take-off Weight	FOM	Figure of Merit
C_D	Coefficient of drag	S	Wing reference area
C_L	Coefficient of lift	η	Efficiency
C_M	Pitching moment	μ	-
ρ	Density of air	θ	Motor tilt angle
σ	Areal density	K	-
W/S	Wing loading	$_{vt}$	Vertical tail components
W/P	Power loading	$_{ht}$	Horizontal tail components
L	Lift (N)	$_{vtol}$	VTOL assisting components
D	Drag (N)	C	Rated charge capacity

***Abstract*—An all-electric unmanned aerial system with both VTOL and Fixed wing capabilities is designed and optimized for long range surveillance and relief operations. The UAV is equipped with onboard computers and sensors and is capable of carrying 1kg of relief payload upto 100 Km. The entire low fidelity design process- from concept to render is carried out using completely open source tools, libraries and in-house code. The challenges faced and primary differences are discussed parallelly. A comparison with commercial codes and programs is also done in some areas to give an overview of key capabilities and caveats.**

***Keywords*—FreeCAD; OpenVSP; OpenFOAM, VTOL; Hybrid; Tricopter; UAS-DTU**

I. INTRODUCTION

The development of Unmanned Aerial Vehicles is gaining popularity with the advent of more research and robust tools for the same. An even more recent development is the VTOL hybrid concept which combines the advantages of both fixed wing and vertical thrust configurations. They can adapt to the mission scenario while keeping good performance characteristics throughout. One of the primary challenges in developing such vehicles is the added weight and parasite drag of the vertical thrust components as compared to conventional designs. The low operating Reynolds number of MAVs makes this an even more difficult problem to tackle.

A key challenge that undergraduate students and inexperienced designers face is the unavailability of consolidated resources to learn and develop good designs. This is even more pronounced in the case of aircraft design where the intertwined relationships of multiple disciplines are profound. The use of commercial tools act as a proponent in such a case as they do not provide the user with the freedom to learn and customize accordingly. Opensource tools and the development of custom libraries/components helps in learning from scratch and the leverage to create and use such tools has never been greater than before.

The authors recognize the above two challenges and aim to alleviate them by developing a sound methodology spread across a wide variety of completely free tools and in-house code. Our research is aimed primarily to help students and upcoming researchers developing feasible designs for student competitions or learning to design in general. Formally, the rationale of our research is centered around the following outcomes:

Tangible: The development of a robust and stable design for a VTOL hybrid UAV within the defined design constraints.

Intangible: Inculcation of completely open source tools and libraries for the entire design and testing routine so as to move towards “free design” methodologies.

The paper is organized as follows- beginning with a background and literature review of the current state of research carried out by others. This is followed by our planning and execution explained in the Methods section and the entire design process. A brief section on the fabrication and testing is also given. To conclude, some final comments and discussion on the design and methods used is carried out.

Note: *This research extensively uses the term “hybrid” which is not to be confused with hybrid electric propulsion which relies on the use of gasoline powered generators for the production of onboard power. However, “hybrid” here would only refer to the operational capabilities of the aircraft*

© Nikhil Sethi, Saurav Ahlawat, N. S. Raghava 2020

as both a multirotor and a fixed wing craft unless stated otherwise.

II. Background

A VTOL fixed wing UAV has the ability to take off and land vertically as well as hover in place along with achieving a sustained cruised flight. These kind of hybrid UAVs combine VTOL capability with the standard forward propulsion of a fixed wing UAV. In many hybrid VTOL UAVs, rotary lift propellers are typically incorporated into the aircraft's wings, which then transition for forward flight. VTOL fixed wings have several advantages over typical fixed-wing unmanned aircraft. They require much less space to launch and recover, as they do not need to use a specified area for Take-Off. They are suited to applications where aerial inspection and monitoring is needed, making the aircraft maintain a fixed position for a period of time

A study of [1], covers the use of SUAVE, a software used to analyze, optimize, and design a range of small UAVs which has been shown to be a flexible aircraft design tool with the unique ability to handle alternative energy systems and unconventional designs. In [2], a similar configuration to this paper is explored using standard low fidelity analyses. The nature of the software and tools used is not stated however. In [3], potential design strategies are presented and compared, identifying various design variables and constraints which take precedence during a particular optimization while taking a novel weight build-up approach to estimating operating empty weight. This work simplifies the larger study of optimizing sensitivity of constraints while removing implicitly resolved parameters alongside reducing computational cost.

In this work, a wide variety of tools and resources were used in the design and development and a brief description of each and their capabilities is highlighted below:

1) FreeCAD: It is a powerful open source solid modelling software written in C++ with a highly interoperable Python API. It has multiple workbenches to deal with different kinds of geometry. Being a relatively old software, the documentation is sufficient and has a well-maintained forum.[1]

2) OpenFoam+ParaView: Another toolbox written C++ for the development of sutomized numerical solvers containing all pre as well as post-processing utilities for the solution of continuum mechanics problems in computational fluid dynamics.[1] The library is also distributed with Paraview which is a verstatile post processing software to visualise the computed fields.

3) LuxCoreRender: This software provide artists with a powerful tool to create extremely realistic and accurate images. It is very flexible to the user and does not compromise on quality and physical correctness where the calculations are done according to mathematical models based on physical phenomena aiming to make 'unbiased' design choices.

4) OpenVSP: It is a parametric aircraft geometry tool where the user is allowed to create a 3D model of an aircraft defined by common engineering parameters. This model can be processed into formats suitable for engineering analysis. This software was released as an open source project under the NASA's Open Source Agreement (NOSA) version 1.3. giving the advanced tools of aircraft design for use to the general public.[5]

5) XFLR5: It is an analysis tool for airfoils, wings and planes operating at low Reynolds Numbers which includes XFoil's Direct and Inverse analysis capabilities with wing design and analysis capabilities based on the Lifiting Line Theory, on the Vortex Lattice Method, and on a 3D Panel Method.

6) Ardupilot SITL/Mission Planner: The SITL (software in the loop) simulator allows us to run Plane and Copter simulations without any actual flights. It is a build of the autopilot code using the original C++ compiler, giving us a native executable that allows one to test the behaviour of the code without implementation on the hardware itself. The SITL uses the sensor data coming from various flight dynamics models in a flight simulator. This allows ArduPilot SITL to be tested on a very wide variety of vehicle types, essential for testing a hybrid system such as ours.

III. Design Methodology

Incorporating completely open source tools for the conceptual preliminary and detailed design is a challenge because of three main factors:

1) *The accuracy of results*

2) *Lack of documentation*

3) *Lack of experienced users/populated forums*

Considering the above challenges and the fidelity required for the design, the tools along with the design process are shown in fig 1.

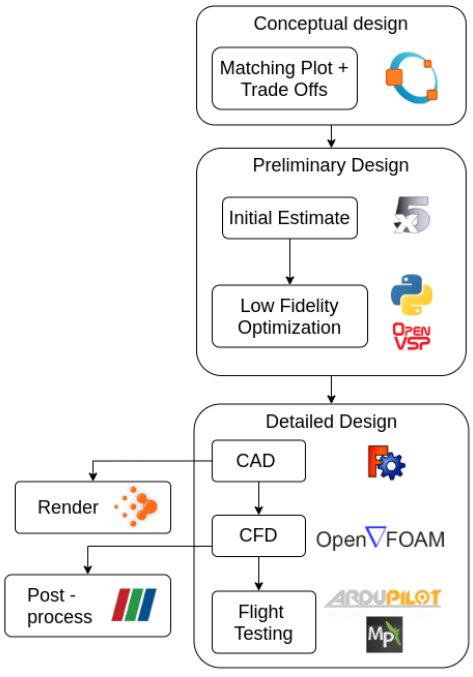


Fig 1. High Level Design Process.

The high degree of inter-related disciplines in aircraft design make it difficult to dodge locally optimal designs developed by intuition. Multidisciplinary design optimization has therefore become a necessary part of the toolbox of aerospace designers. The approach used in the following sections is not a full scale MDO solution by any means but a novel low fidelity attempt at design exploration. A human is kept in the loop at all stages to guide the design and perform validation checks. Further methods pertaining to the same are explained in the following sections

IV. Conceptual Design

For the purpose of this research, we model our aircraft for maximizing its range of flight to a 100kms, cruising at 25m/s while maintaining its VTOL functionality. The aircraft will keep a standard mission profile with vertical take-off, cruised flight and vertical landing.

A. Configuration selection:

A hybrid of a blended flying wing configuration supplemented with a tricopter configuration with two forward tilting motors is arrived at. The forward motors tilt forward for transition to cruise flight after a vertical take-off and tilt back to multirotor while landing. The rear motor is operational for the multirotor segment of flight.

$$(C_{Do})_{Total}$$

$$= (C_{Do})_W + (C_{Do})_{ht}\frac{S_{ht}}{S} + (C_{Do})_{vt}\frac{S_{vt}}{S} + (C_{Do})_{vtol}\frac{S_{vtol}}{S}$$

A blended flying wing will prove to be a better candidate for long endurance missions with significantly reduced $(C_{Do})_{ht}$ component and better performance of winglets to reduce vortex induced drag and an overall lower $(C_{Do})_{vt}$ component.

Most commercially available products use a quad-plane configuration, which is basically a quadcopter attached to the frame of an independent fixed wing aircraft. This solution, though the least complicated, added redundant structures and additional motors which contribute more to $(C_{Do})_{vtol}$ component for additional drag and are basically dead weight in cruised flight. This would not be an optimal choice for a high endurance mission.

A tri plane with two forward tilt motors however is a highly favorable choice in terms of weight and lower $(C_{Do})_{vtol}$. Even though planes with a greater number of rotors provide better stability and airworthiness, their cost and weight diminish the advantages. Owing to a tri copter's superior stability and airworthiness along with its cost effectiveness, it is chosen.

B. Propulsion:

The propulsion selection is the most essential tradeoff to obtain the most out of the aircraft's endurance and reach the desired range.

During cruise flight, all the load will be on the forward motors whereas during hover, all the three motors would share the load. This introduces the issue of transition where the forward motors have to be chosen in such a way so as to provide sufficient thrust during hover as well as have optimal RPM and pitch speed for a sustained cruised flight.

The thrust margin of the all the motors was arrived at an *x + x + 2x* ratio where the max thrust of the forward motors is capped at *x*. This distribution was a necessary step to ensure that the forward motors operate at their peak optimum during cruise and the rear motor could handle the remaining thrust required for hover. The rear motor, only operational during hover segments, is optimized for hover and the ratio of thrust distributed on it is calculated based on the available thrust margin from the forward motors after they have been finalized during selection for cruised flight.

C. Weight Estimation:

The initial estimation is done by analyzing historical data, collecting the necessary features and arriving at a hypothesis akin to the methods in [7]. Since the payload is defined within the constraints, the payload fraction was used as the most important feature at this stage and a linear regression curve is fitted to the data. This trendline along with the equations is solved analytically to give a prediction.

$$W_0 = a\left(\frac{W_p}{W_0}\right) + b$$

$$W_0 = W_\mathrm{e} + W_\mathrm{p}$$

where a = - 0.00468; b = 0.878

The regression analysis yields the following total mass of the aircraft required for 1 kg of payload. A 200g safety margin brings this up to an estimated 7kg aircraft.

Payload = 1 kg

Total = 6.8 + 0.2 = 7kgs

V. Preliminary Design

The preliminary design stage makes use of extensive mathematical optimization. The objective of the optimization process is to design a geometry and propulsion system to maximize the range which implies minimizing power and maximizing cruise efficiency. In our case, we have divided the optimization and design process into three stages.

A. Initial Sizing:

The first stage gives a good estimate of the two driving factors for any design- wing loading and power loading. Appropriate relations for the transition, hover and climb phases for the tricopter sizing are also included to account for disc loading using an approach very similar to the one in [8]. Since a rough estimate is desired at this stage, a design space is created using Linear programming based on the following formulae:

$$\left(\frac{W}{S}\right)_{stall} = 0.5\rho(V_{stall})^2 C_{Lmax}$$

$$\left(\frac{W}{S}\right)_{\max range} = q\sqrt{\pi A e C_{D_o}}$$

$$\left(\frac{W}{S}\right)_{\max loiter} = q\sqrt{3\pi A e C_{D_o}}$$

$$\left(\frac{W}{P}\right)_{FW\ ROC} = \frac{1}{\frac{ROC}{\eta_p} + \sqrt{\frac{2}{\rho\sqrt{\frac{3C_{D_o}}{K}}}\left(\frac{W}{S}\right)}\left(\frac{1.155}{\left(\frac{L}{D}\right)_{max}\eta_p}\right)}$$

$$\left(\frac{W}{P_{SL}}\right)_{V_{max}} = \frac{\eta_p}{\frac{1}{2}\rho_o V_{max}^3 C_{D_o}\frac{1}{\left(\frac{W}{S}\right)} + \frac{2K}{\rho\sigma V_{max}}\left(\frac{W}{S}\right)}$$

$$\left(\frac{W}{P}\right)_{hover} = FOM \times \left(\frac{\sqrt{2\times\rho}}{DL}\right)$$

$$\left(\frac{W}{P}\right)_{climb} = \frac{1}{V_y - \frac{k_1 V_y}{2} + \frac{k_1}{2}\sqrt{V_y^2 + \frac{2(DL)}{\rho_o} + \frac{\rho_o V_{tip}^3}{(DL)}\left(\frac{\rho C_d}{8}\right)}}$$

Hence,

$$\left(\frac{W}{P}\right)_{trans} = \frac{1}{d_1 + d_2 + d_3}$$

$$d_1 = \frac{k_1}{\sin(\theta_{tilt})}\sqrt{\frac{-V_\infty^2}{2} + \sqrt{\left(\frac{-V_\infty^2}{2}\right)^2 + \left(\frac{DL}{2\rho\sin(\theta_{tilt})}\right)^2}}$$

$$d_2 = \frac{\rho V_{Tip}^3}{DL}\left(\frac{\sigma C_d}{8}(1 + 4.6\mu^2)\right)$$

$$d_3 = \left(\frac{1}{2}\rho_o V_\infty^3 C_{D_o}\frac{1}{\left(\frac{W}{S}\right)} + \frac{2K}{\rho_o V_\infty}\left(\frac{W}{S}\right)\right)$$

The LPP outputs a comfortable design space to choose points from depending on desired performance. To make the analysis more constrained, minimum and maximum values of disc loading are also used depending on the motors and their rated propeller sizes. As shown, the VTOL and fixed wing design points are far apart, both constrained by their climb limits. For each configuration, the maximum power points are chosen to act as a safety factor.

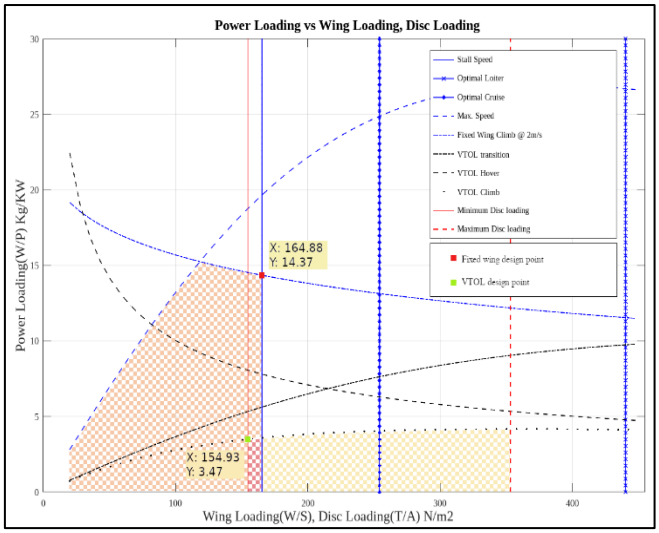


Fig 2. Preliminary Design Space

$$S_{wing} = \frac{W_o}{\left(W/S\right)_{opt}}$$

$$P_{max} = \frac{W_o}{max\left\{\left(W/P\right)_{FW}, \left(W/P\right)_{vtol}\right\}}$$

B. *Geometry optimization*

1) *Initial 'Human' estimate*

An initial estimation of aircraft performance is first analyzed on XFLR5 focusing solely to obtain initial but educated estimates of static stability parameters.

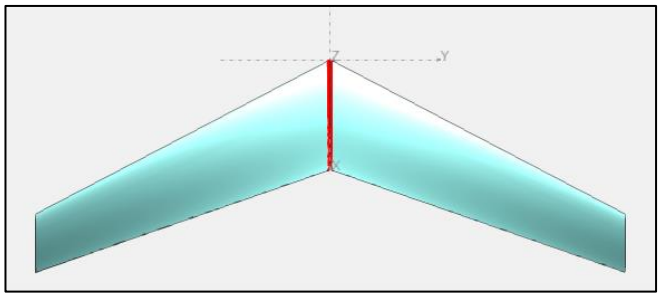


Fig 3. Initial Wing Planform

Wing planform area S is initially estimated by using crude C_l values obtained from observing trends in standard use reflexed airfoils on angle of attack at peak C_l/C_d values operating at cruising altitude and speed Reynold's numbers.[10]

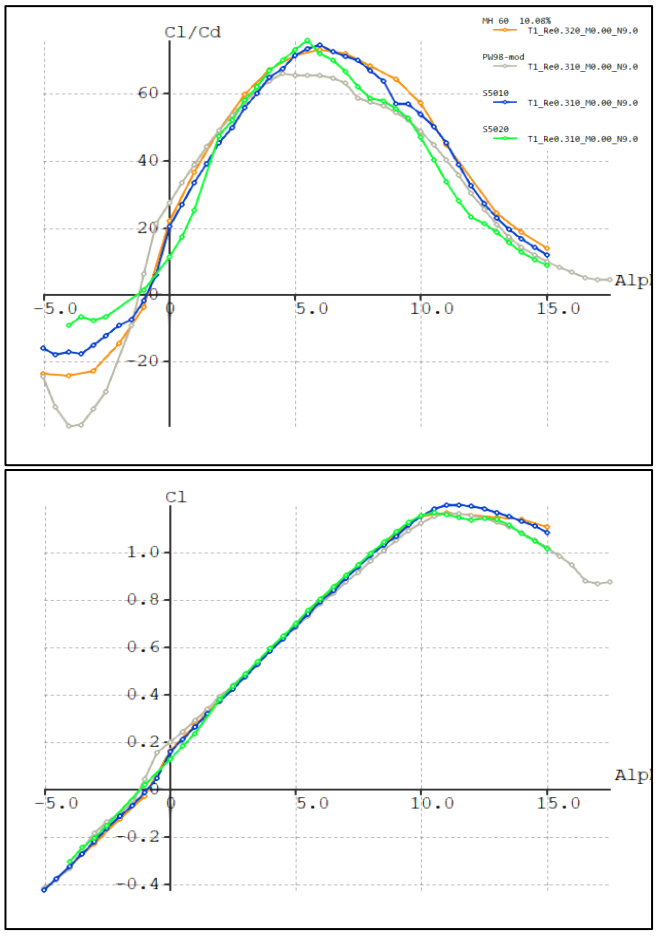


Fig 4. Initial Airfoil Study

Fixed lift analysis is then conducted on the operating GTOW of the flying wing now to introduce static stability into the craft by tweaking the tip twist in the wing with an optimal sweep backwards, aiming to nullify pitching moment C_m of the flying wing.

Iterations were performed on the design variables to optimize the cruise C_l/C_d of the whole wing while balancing out the placement of the CoG to ensure close to null pitching moment. The resultant twist in the wing subsequently reduced the C_L of the whole wing, making the wing area more than initial calculations.

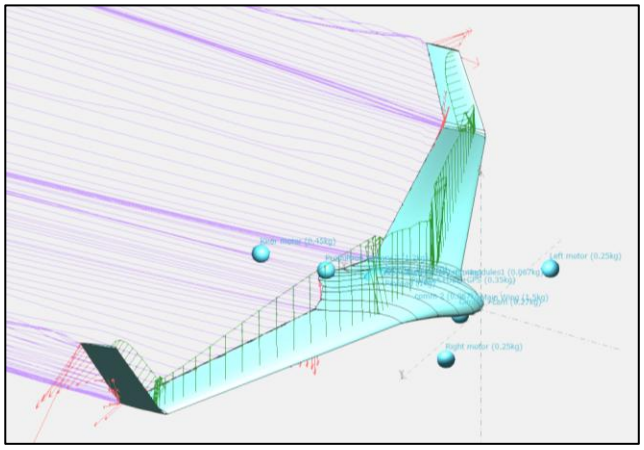


Fig 5. XFLR5 Preliminary Design

A rough and conceptual estimate of the intended design is also analysed but not optimized to ensure consistency in performance estimates.

2) Problem Setup

$$Min_x \quad w^T.J$$

$$w^T = [w_1, w_2 \ldots w_{\#J}]$$

$$J = \left[C_D,\ \frac{1}{n}\sum_{i=1}^{n}\left(\frac{C_l}{C_{ref}}(x)^i_{actual} - \frac{C_l}{C_{ref}}(x)^i_{target}\right)^2,\ \zeta_{sp}, \zeta_{dr}, \zeta_{ph}\right]$$

$$0 \le x \le b/2$$

$$x = [AR, \theta_{rel}, \lambda, \Lambda_{LE}]^T$$

Subject to

$$C_L = 0.9 \times C_l^c$$

$$S.M. = 4\%$$

$$C_M \approx 0$$

$$5 \le AR \le 10$$

$$-1 \le \theta_{rel} \le -5$$

$$0.4 \le \lambda \le 0.9$$

$$20 \le \Lambda_{LE} \le 30$$

The second stage has optimization geometry selection. While being decoupled from each propulsion selection is not optimal, it makes the development of the interface easier. Owing to the small number of variables and ease of writing code, a particle swarm optimizer is used.

3) Interface

Python was chosen as the scripting language to carry out the optimization. The libraries/software used for the design all have a well exposed API which enables the use of already developed macros to carry out efficient numerical computation.

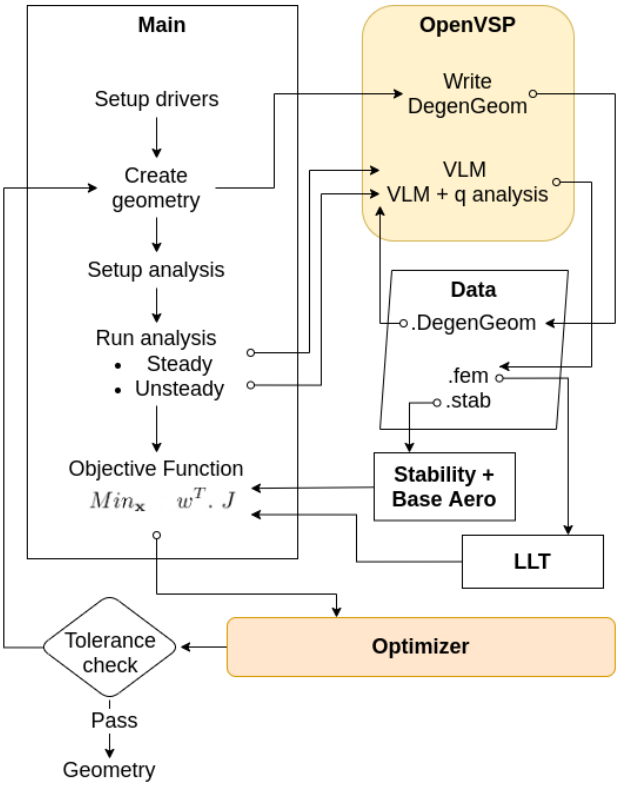


Fig 6. Preliminary Design Space

The geometry is arrived upon by interfacing a python script with OpenVSP and conducting a Vortex Lattice method analysis. The objective of the optimization accounts for three primary factors:

Drag: the analysis is done at constant lift coefficient which is equal to the lift at cruise condition.

Lift distribution; The lift distribution is a function of twist, taper ratio and sweep. A least squares cost function is used to match the target curve of the load distribution. A separate python script calculates this.

Stability: Since optimizing for stability often requires compromise on performance, the damping ratios of the short period, Dutch roll and phugoid mode are incorporated within the objective. For static stability however, a constraint for the design lift coefficient at zero moment is enough. The stability coefficients and BaseAero values are sought from the .stab file output by OpenVSP and converted to a data frame for further processing.

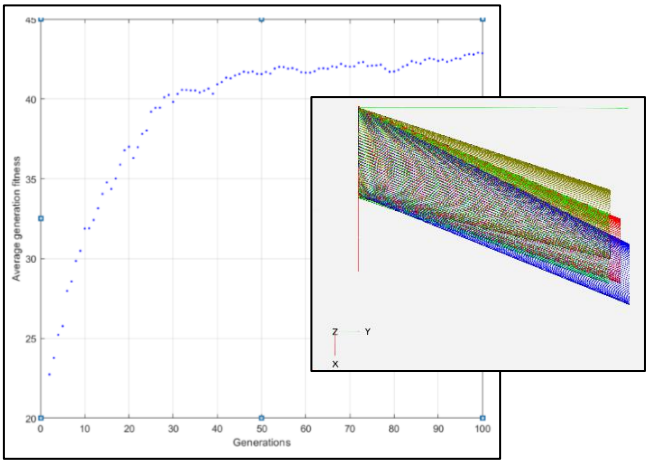


Fig 7. Optimization tradeoffs(right), Avg. Fitness vs Generations(left)

Some comments on each of design areas with respect to the tools are written to help the reader understand the level of sophistication available and possible caveats.

- *Vortex lattice method is not an accurate method for determining the overall drag and therefore the values are further verified by CFD analysis.*
- *The focus of the geometry optimization is primarily on the cruise phase.*
- *Since this is not a full-scale MDO solution, A human in the loop is necessary to guide and remove obviously suboptimal designs.*

Table 1. AIRCRAFT PARAMETERS

Parameter	**Values**
Wing loading	164.88 N/m²
Power loading	3.47 Kg/Kw
Wing Area	0.418 m²
Aspect Ratio	7.35
Taper Ratio	0.689
Wing incidence	5°
Relative Twist	-2
Leading Edge sweep	26.5°

C. Propulsion selection:

To drive the aircraft with its airframe designed at its peak performance with reduced drag to achieve the most of endurance, a propulsion architecture was conceived.

1) Forward Motor

These motors were the most essential part in optimizing cruise segment of the flight. The right combination of propeller size, pitch and motor combination was combed through to achieve the peak optimized performance of the motors while cruising at 25 m/s. This was matched with the pitch speed of the motor operating on the thrust needed at cruise and hence maximizing endurance.

The initial design of our fixed wing yielded a cruise drag $C_{d\ cruise}$ whose drag would have to be overcome by over forward cruise motors.

Get KV

$$v_{pitch} = [x_{pitch} \times \omega_{cruise}]$$

$$KV_{cruise} = \left[\left(\frac{v_{pitch}}{x_{pitch}}\right) \times \frac{1}{V_p} \times 60 \times \emptyset_c\right]$$

$$\emptyset_c = \sqrt{\left(\frac{T_{max}}{D_{cruise}}\right)}$$

Subject to

$$v_{pitch} = 25m/s$$

$$2.36 \leq \omega_{cruise} \leq 4.92$$

$$4" \leq x_{pitch} \leq 8"$$

$$0.5 \leq T_{max} \leq 0.8\ GTOW$$

$$D_{cruise} = \frac{1}{2}\rho v^2 S C_{d\ cruise}$$

$$V_p = 22.2V\ (6S\ Battery)$$

An RPM scaling factor $\emptyset_c$ is introduced keeping in mind the fact that the Required RPM increases quadratically with the required Thrust for a constant pitch propelled BLDC motor. This factor scales the motor RPM at cruise to what its RPM would likely be at Maximum of the basis of our requirements of D_{cruise} and T_{max}.

BLDC motors operating at 22.2V (6S) battery packs were searched upon which could give the required thrust of 17N at 500m and a pitch speed of 25m/s. EMAX MT3515 650KV motor coupled with a 10"x6" propeller was finalized upon after iterations of various KV rating of motors.

The max load the motor could handle was 26N of thrust, comfortable for flying at max T/W of 0.74 and at *x + x + 2x* configuration of motor thrust distribution can easily hover with a thrust margin of 1.48:1 during hover segments.

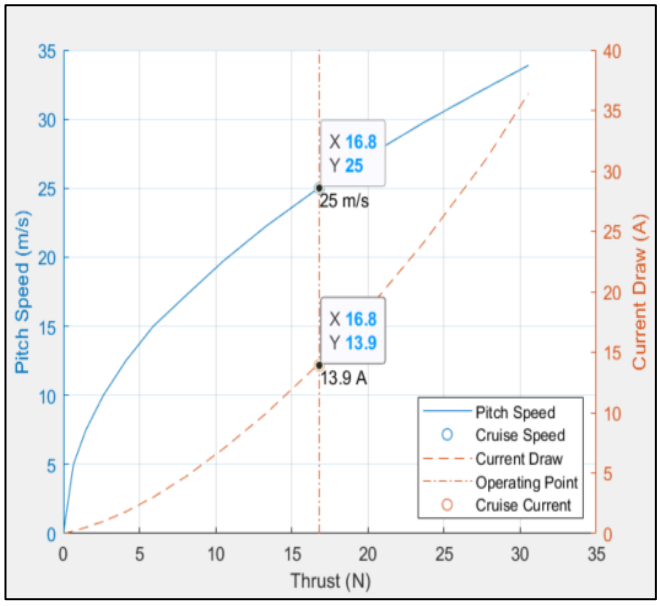


Fig 8Forward Motor Optimal Thrust

2) Battery Selection

Observing modern trends in electric propulsion in aviation, Lithium Ion battery packs prove to be the best in terms of energy density, holding more amperage compared to their Lithium Polymer counterparts. Lithium Ion battery packs do have their drawbacks of a limited current discharge. Current was evaluated for all segments of flight and was found sufficient to use.

Table 2. BATTERY DATA

Parameter	**Values**
Form factor	18650
Cell Voltage (nominal)	22.2V
Cell Capacity	3300mAh
Cell Discharge	3C (9.9A)
Cell Discharge (Peak)	4C (13.2A)
Weight (cell)	48g

$W_{battery}(\max) = 3Kg$

$6S\ 10P = 2880g$

Hence for operating at a 6S 10P Configuration:

Flight range of more than 100Kms was subsequently verified by calculating the battery usage in cruised flight.

$$Flight\ Time = \frac{C_{cell} \times n_{Parallel\ cells}}{I_{cruise} \times n_{motors}} \times \eta_{total}$$

$Flight\ Time = 71\ mins$

$Range = v_{cruise} \times t_{cruise}$

$Range = 107\ Kms\ (Theoretical)$

3) Rear Motor

With the forward motors handling half of the thrust required during hover segments the remaining is handled by the rear motor. Its sole optimization goal is to draw the least amount of current so as to be as efficient as possible in doing its job. Motors with larger propeller size and lower RPM (KV rating) draw lesser current as the induced pitch speed is lesser than those with shorter propellers and higher RPM.

The rear motor was hence finalized by limiting the propeller diameter to the length of available clean area under the rear propeller wash disc keeping in mind the placement of the rear motor according to the CoG.

The collective propulsion results are validated on and electric propulsion calculator which is not free. The results are in close agreement with the developed formulae above.

Table 3. FINAL MOTORS SELECTION

Forward Motor (×2)	**Value**
EMAX MT3515	
Type	BLDC 22.2V
KV rated	650 RPM/V
Maximum Power	630 W
Propeller	10” x 6”
Max Thrust	2660g
Weight	131g
Rear Motor (×1)	**Value**
T-MOTOR MN501-S	
Type	BLDC 22.2V
KV rated	360 RPM/V
Maximum Power	1000 W
Propeller	20” x 6”
Max Thrust	5372g
Weight	171g

VI. DETAILED DESIGN

The preliminary design results in a low fidelity flyable aircraft capable of satisfying the theoretical performance objectives. More sophisticated software for Computational fluid dynamics and aircraft control simulation engines are used at this point to validate the design.

A. CAD and preprocessing:

The solid model with all internal features is designed on FreeCAD. Extensive use of the part design and draft workbenches allows the creation of complex models and geometries. The commercial tool used for comparison is SolidWorks. As shown below, FreeCAD can capture most of the geometry but the tail is off design due to the lack of sophisticated lofting features with guide curves as compared to Solidworks.

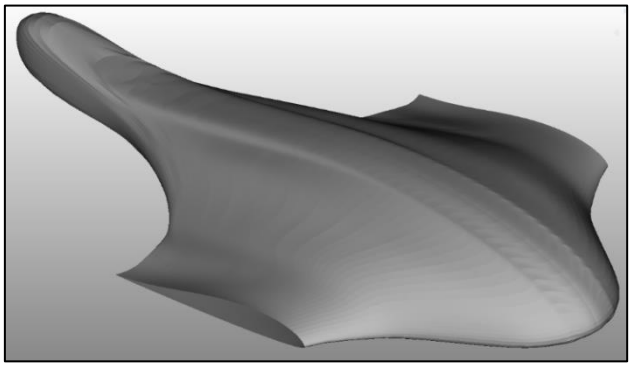


Fig 9. FreeCAD Body

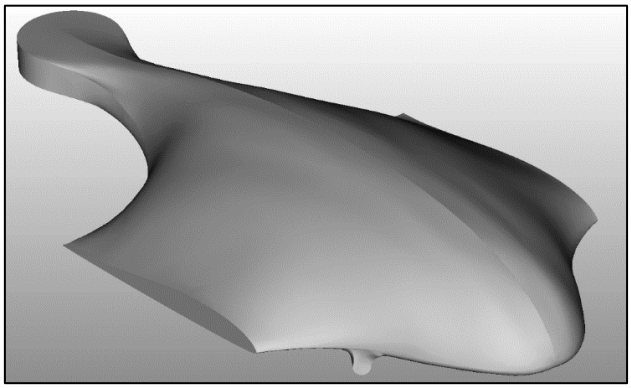


Fig 10. SolidWorks Body

Triangulation for the CFD model is done using the Draft, part and mesh design workbench. Each part and face

are sequentially upgraded using the draft upgrade, fused using the boolean fuse option and finally the entire fusion is converted to a mesh. Owing to the tedious processing time of a Netgen mesh, the standard mesh was used and further mesh optimization and quality improvement was conducted in the CFD tool.

B. CFD and postprocessing:

A computation fluid dynamics analysis was setup to calculate the lift and drag more accurately than the low fidelity analysis. OpenFOAM was used to do analysis.

For preprocessing, blockMesh and snappyHexMesh were used with the parameters given below. Some refinement regions were also added in particular areas like the leading edge and winglet-wing blend. The mesh quality dictionary was also edited to refine some areas. The following table shows only the edited values.

Table 4. MESH PARAMETERS

Parameter	Value
addLayers	false
Leading edge refinement levels	(1E15 8)
Winglet refinement levels	(1E15 7)
Front motor refinement levels	(1E15 7)
Local refinement box levels	(1E15 5)
implicitFeatureSnap	false
mergeTolerance	1e-6
maxNonOrtho	60
minTwist	-1e+30
minVol	-1e+30
minDeterminant	-1
minFaceWeight	-1
minTriangleTwist	-1

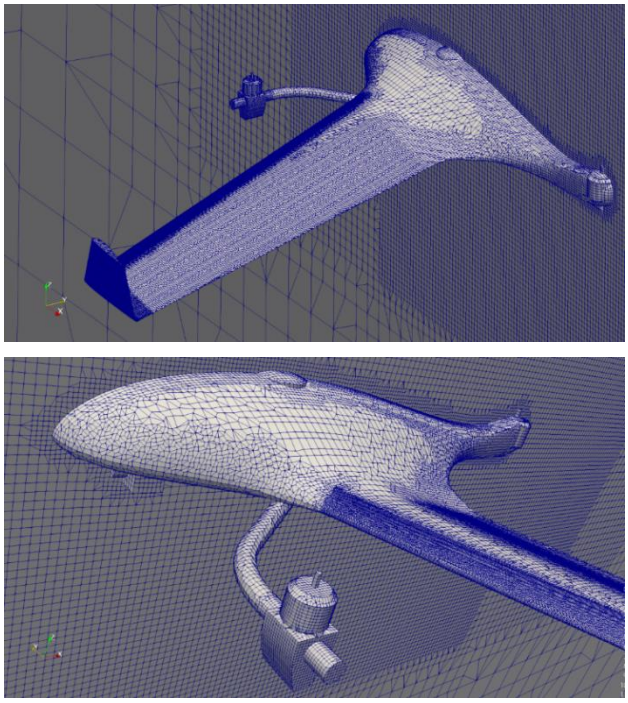


Fig 11. snappyHexMesh results

The basic steady state simpleFoam solver was used to obtain the coefficients at cruise state. The solver was run for 500 iterations but resulted in satisfactory residual convergence in 200 iterations (fig 13.)

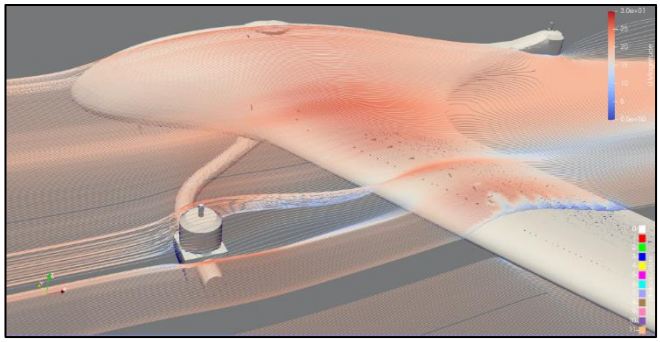


Fig 12. Streamlines coloured by velocity

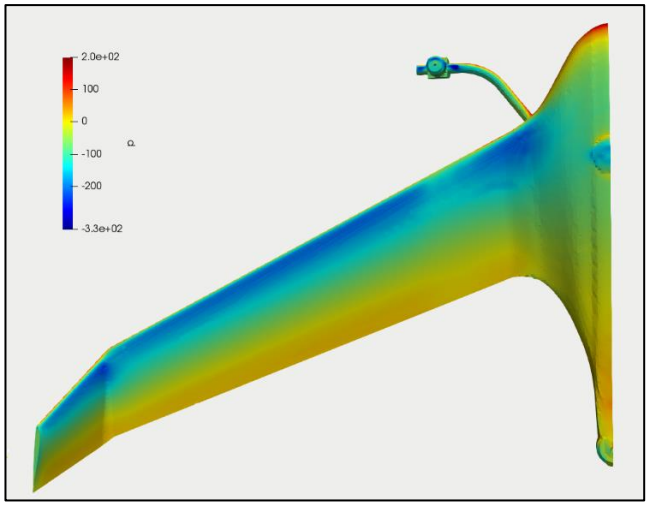


Fig 13. Pressure Contours

Table 5. OBTAINED COEFFICIENTS COMPARISON

Coefficient	XFLR5	OpenFOAM
C_D	0.018	0.038
C_L	0.477	0.403

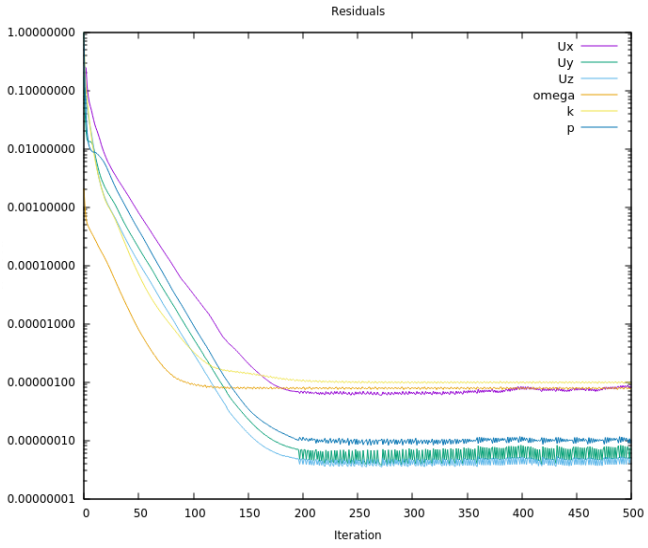


Fig 14. Residual convergence

C. Render

While rendering is obviously not essential to the design process, it is required in some commercial sectors. Conventionally it is done using expensive tools and software. Contrarily, we have used LuxCoreRenderer as the opensource tool to do the same and included it in the design loop.

The raytracing module of FreeCAD is used to output a .lxs configuration file which is modified accordingly. Note that the default .lxs file template has preset sizes and transformations of area light sources and one can either change the config file or the CAD itself. Our option was to choose the former which again highlights the freedom offered by opensource software. The changes to the default configuration are given. The entire render with a final noise pass error of about 0.001 took about 15 minutes to complete.

```
1. Camera
2. LookAt
        42.9652         27.7823         39.252
        -7.02438        -22.438         -0.599087
        -0.336994       -0.35615        0.871547
3. Flex image
4. "integer xresolution" [1080]
5. "integer yresolution" [1360]
6. Groundplane
7. Transform
        [4.000000000000000      0.000000000000000
        0.000000000000000       0.000000000000000
        0.000000000000000       4.000000000000000
        0.000000000000000       0.000000000000000
        0.000000000000000       0.000000000000000
        4.000000000000000       0.000000000000000
        0.000000000000000       0.000000000000000
        0.000000000000000       4.000000000000000]
8. Area Lightsource : sun
9. Transform
        [15.000000000000000     0.000000000000000
        0.000000000000000       0.000000000000000
        0.000000000000000       15.000000000000000
        0.000000000000000       0.000000000000000
        0.000000000000000       0.000000000000000
        15.000000000000000      0.000000000000000
        0.000000000000000       0.000000000000000
        40.000000000000000      1.000000000000000]
10."float power" [200.000000000000000]
11."color L" [0.50000000 0.95364237 0.70636380]
```

A comparison between Solidworks PhotoView and LuxCore is also done:

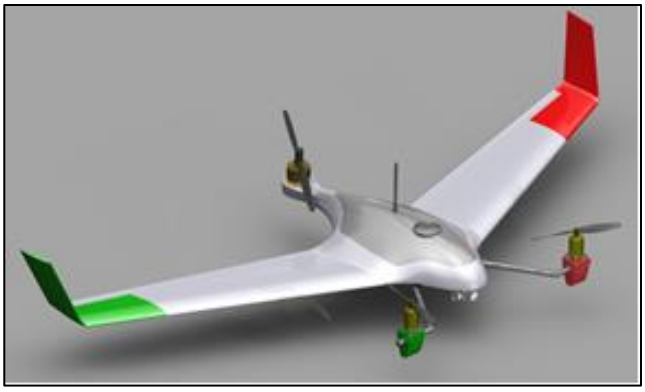


Fig 15. Solidworks+PhotoView Render (top)

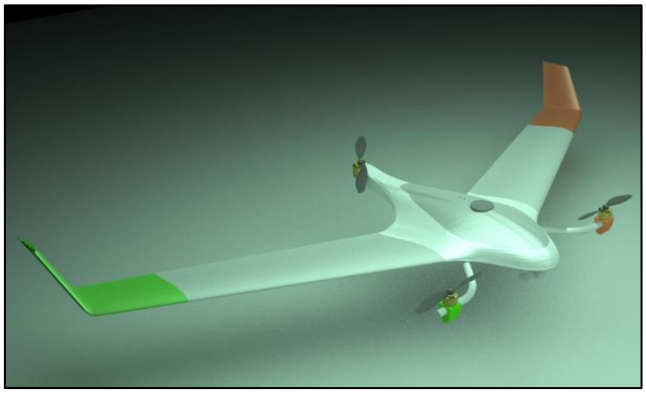


Fig 16. FreeCAD + LuxCore Render (bottom)

VII. Development and Integration

A. Fabrication:

A prototype body was fabricated to validate some performance and stability characteristics of the aircraft. The body is made of Extruded polystyrene foam (XPS) and glass fiber fabric through a vacuum assisted wet layup method. Some aspects of the configuration were altered depending on the resources available.

The fuselage was made by hotwiring along each orthographic projection plane. This method was highly efficient in terms of time and ensured minimum sanding of the foam to get the final shape. Aluminum pipes were used to define the structure of the tricopter assembly and the foam blended wing body was integrated along.



Fig 17. Fabrication and Vaccum bagging



Fig 18. Assembled body (1st iteration)

B. Integration

A well-defined architecture is shown in the fig below highlighting data flow between the subsystems.

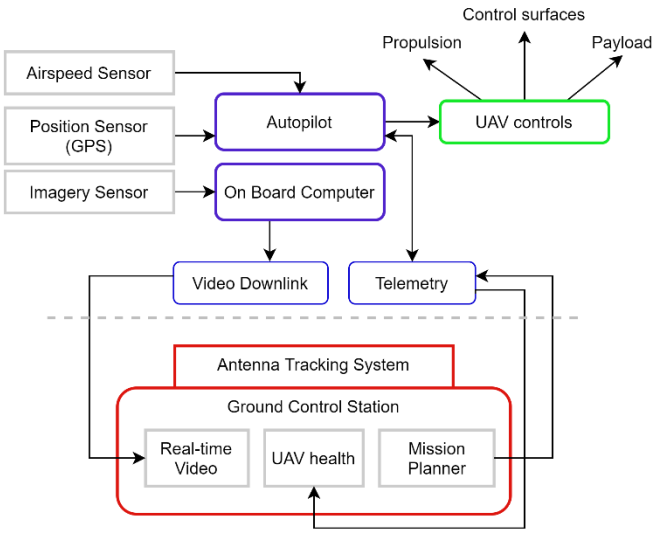


Fig 19. System Architecture

A brief description of some of the components is given:

1) Pixhawk 2 Cube: An autopilot is necessary to carry out custom missions and provide additional stability during flight. It also allows the UAV to maintain a level altitude while transitioning into cruise state.

2) GPS: A Here GPS is used to localize the UAV in its environment.

3) Telemetry: An RFD 900x telemetry is used to create a link between the ground systems and send appropriate messages in between the flight.

4) Sensors: These include everything ranging from imaging systems, airspeed sensors and

5) Raspberry Pi 3: The onboard computer's primary purpose is, to carry out the image recognition task among other necessary mission scripts.

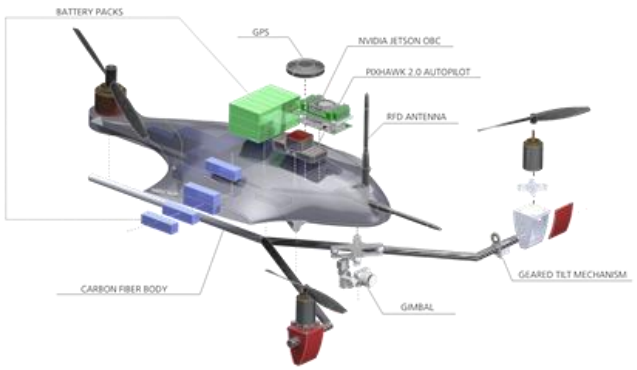


Fig 20. Exploded View of subsystems

A key feature is the form of the modular propulsion batteries. Multiple Lithium Ion cells can be used to form arrangements so as to fit into the body. The blended design offers a lot of payload volume which is exploited by stacking 9 parallel sets of 6 cells each connected in series. Some of the battery packs are strategically placed very close to the CoG which allows the user to swap the pack for additional payload or sensors.



Fig 21. Final RTF prototype

C. Flight Testing:

Table 6. TUNED FLIGHT PARAMETERS

Parameter	Original	Changed
ARSPD_FBW_MAX	30	28
ARSPD_FBW_MIN	10	16
PTCH2SRV_D	0.200000003	0.04
PTCH2SRV_P	2.5	1
Q_FRAME_CLASS	1	7
Q_TILT_YAW_ANGLE	0	30
Q_A_RAT_PIT_P	0.25	0.2
Q_A_RAT_YAW_P	0.1800000072	0.25
Q_A_THR_MIX_MAN	0.1000000015	0.5

The prototype was tested physically over a few flight tests to evaluate flight dynamics and tune the autopilot. Mission Planner was used for tuning along with ArduPlane software. The testing resulted in many unforeseen issues which were resolved after each iteration. Some key points of discussion are noted:

1) The low cost and quality of materials due to the tight budget made it very difficult to debug what the problem was and resulted in increased number of flight times.

2) Lack of prior experience and exposure in handling or operating a hybrid craft made every challenge as a new problem to be tackled.

3) Even post debugging the problem, finding the right solution was a tedious task due to the vast number of PID parameters to tune. A great help at this time was the shared experience of the community at the arduplane forums.

VIII. CONCLUSION

A VTOL unmanned aerial system capable of flying 100km and performing a relief mission is developed. Particular care was taken that no commercial tools or codes are used. Using completely opensource tools has its own sets of challenges and advantages. Some comparisons and discussion are therefore required to close the loop.

A. Discussion

Aircraft design is and always will be a multi-processing task where no linear approach exists and no defined starting point awaits. Even a piece of cardboard is airworthy enough to fly, but for it to take shape into an optimized craft with required capabilities, various parameters have to be tweaked

simultaneously with one affecting the other at every step in the design process. To overcome such complexities, communities exist online as well in laboratories, aiming to pool in the unquantifiable amount of knowledge of Aero Design into open source software and libraries:

1) Opensource tools are not without their limitations. A significant effort was required to get acquainted with all tools at every stage and produce meaningful results.

2) A key difference which tends to take up significant time is the lack of standardised coordinate systems, conventions of unit systems, and general understanding of some concepts between each softwrae. This is because opensource softwares, while guided by some core teams are the heart of it-collaborative. And this results in different conventions on each platform. Effort from the user's side is required to understand this point and change models/configurations accordingly

3) The above point becomes more important when there is significant lack of documentation, discussion forums and in general people using it. To overcome this , the clarity of fundamentals of types software architectures and the core field are important from the user's side.

4) The issue of 'accuracy' of results was verified collectively by reading up on literature and conducting some of the analyses on commercial software and calculating errors.

ACKNOWLEDGMENT

The UAV was fabricated at Team UAS-DTU laboratory, Delhi Technological University and the authors duly acknowledge the help of the team members for the same.

REFERENCES

[1] Emilio Botero and Juan J. Alonso – 'Conceptual Design and Optimization of SmallTransitioning UAVs using SUAVE', AIAA (5-9 June 2017).

[2] Ugur Ozdemir, Yucel Orkut Aktas, Aslihan Vuruskan, Yasin Derel, · Ahmed Farabi Tarhan, Karaca Demirbag, Ahmet Erdem, Ganime Duygu Kalaycioglu, Ibrahim Ozkol, Gokhan Inalhan – 'Design of a Commercial Hybrid VTOL UAV System', Journal of Intelligent Robot Systems, 2014.

[3] Matthew Clarke, Jordan Smart, Emilio Botero, Walter Maier and Juan J. Alonso – 'Strategies for Posing a Well-Defined Problem forUrban Air Mobility Vehicles', AIAA SciTech (7-11 January 2019).

[4] Okoi, Martins (December 1, 2017). "FreeCAD – A 3D Modeling and Design Software for Linux". FOSSMint

[5] Hrvoje Jasak, 'OpenFOAM: Open source CFD in research and industry' in International Journal of Naval Architecture and Ocean Engineering, Elsevier – Dec 2009.

[6] Herbst, Staudenmaier, Sebastian, Adrian (September 25–30, 2016). "Using OpenVSP in a Conceptual Aircraft Design Environment in MATLAB" in Congress of the International Council of the Aeronautical Sciences.

[7] Mohammed Saadarey -- Aircraft design: A Systems Engineering Approach.

[8] Ashraf M. Kamal, Alex Ramirez-Serrano – 'Design Methodology for hybrid (VTOL+ Fixed wing) Unmanned Aerial Vehilces, Aeronautics and Aerospace Open Access Journal, 2018.

[9] Daniel P. Raymer -- 'Aircraft Design : A Conceptual Approach', AIAA Education Series, 2018.

[10] Rishabh Dagur, Vikrant Singh, Shabir Grover, Nikhil Sethi – 'Design of Flying Wing UAV and Effect of Winglets on its Performance', IJETAE, March 2018.

2020 IEEE International Conference for Innovation in Technology (INOCON)
Bengaluru, India. Nov 6-8, 2020

Lyapunov Stability Theory based Non Linear Controller Design for a Standalone PV System

Aman Sharma
Electrical Engineering Department
Delhi Technological University
Delhi, India
aman_mt2k18@dtu.ac.in

Narendra Kumar
Electrical Engineering Department
Delhi Technological University
Delhi, India
narendrakumar@dtu.ac.in

***Abstract*: This paper presents the comparative analysis between different nonlinear controllers based on Lyapunov Stability Theory. Due to the nonlinear characteristics of photovoltaic systems, conventional Hill-Climbing methods like Perturbate and Observe and Incremental Conductance do not show reliable tracking of Maximum Power under various climatic uncertainties. Therefore, these methods requires nonlinear tools to meet the control objectives and design specifications. Out of various nonlinear controlling techniques, the one presented in this paper is the Sliding Mode Approach for Maximum Power Point Tracking (MPPT). In perspective of Lyapunov Stability Theory, sliding mode approach uses a switching manifold. In this approach, the system trajectories are made to follow the sliding surface and to remain there forever to ensure the stability of equilibrium points. Two types of Sliding Mode controllers have been simulated namely Conventional - Sliding Mode Controller (CSMC) and Terminal - Sliding Mode Controller (TSMC). Their results are analyzed and compared scientifically under varying climatic conditions. All the control algorithms are developed in MATLAB/Simulink. They are compared based on various performance parameters, including duty cycle ratio, ideal and PV output power and time taken for error convergence.**

Keywords* – *Nonlinear Controller, Lyapunov Theory, Sliding Mode Controller, TSMC, CSMC, MPPT.

I. INTRODUCTION

Photovoltaic effect involves the creation of a terminal potential difference across the junctions of two materials having different properties whenever radiation of suitable frequency strikes the surface. The process of photovoltaic effect can be summarized as:

1. A junction formed (due to charge carriers) when the photons of sufficient energy is absorbed by the materials.

2. These charge carriers are separated across the junction.

3. Finally the charge carriers are collected at the terminals which give rise to a potential difference.

A solar cell can be represented as an electrical equivalent one-diode model, presented in Fig. 1.

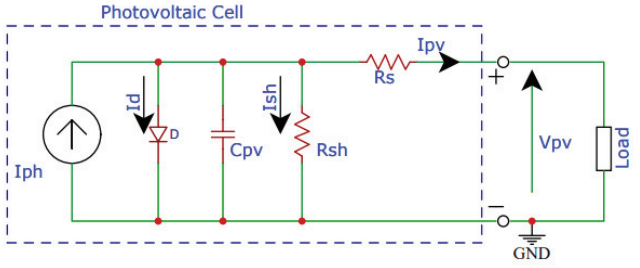


Fig. 1. Equivalent One-Diode Circuit of Solar Cell

This circuit could be employed as an individual Solar cell as in [1]. Fig.1 shows a current source (Iph), series resistance (Rs) and diode (D) which is the resistance inside each cell. The output current is the difference between Photogenerated current (Iph) and diode current (I_D) given by

$$I_{ph} = I_{PV} - I_D - I_{sh} \quad (1)$$
$$= I_{PV} - Io\{exp\left[\frac{e(V_{PV}+I_{PV}Rs)}{\eta KT}\right] - 1\} - \frac{V_{PV}+I_{PV}Rs}{Rsh}$$

By neglecting the two resistances, the net current equation can be modified as:

$$I_{PV} = I_{ph} - I_D = I_{ph} - Io\{exp\left[\frac{e(V_{PV}+I_{PV}Rs)}{\eta KT}\right] - 1\} \quad (2)$$

The PV panel is not like an ordinary classical source. The output voltage is coupled and is affected by output current, expressed by a nonlinear I-V relationship. Output voltage is coupled with the current by a logarithmic relation given by

$$V_{PV} = V_{Thermal} * ln\left[I_{ph} - I/\mathrm{I_{saturation}}\right] \quad (3)$$

Where, $V_{Thermal}$ and $I_{saturation}$ are the thermal equivalent voltage and dark saturation current respectively.

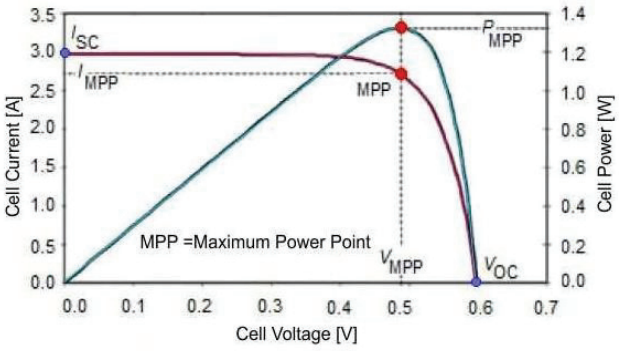


Fig. 2. IV and PV Characteristics of Solar Module

Conventional IV and PV characteristics of a solar cell is deduced in Fig. 2. where MPP corresponds to Maximum Power Point. Further Isc and Voc are greater than I_{PP} and V_{PP} respectively due to internal losses within the cell. Where,

I_{sc} Short circuit current
V_{oc} Open circuit voltage
I_{PP} Peak-Point current
V_{PP} Peak-Point voltage

Two important parameters as presented in [2], that needs to be considered while evaluating the performance efficiency of the photovoltaic systems are

1. Solar Irradiance
2. Temperature

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:04:06 UTC from IEEE Xplore. Restrictions apply.

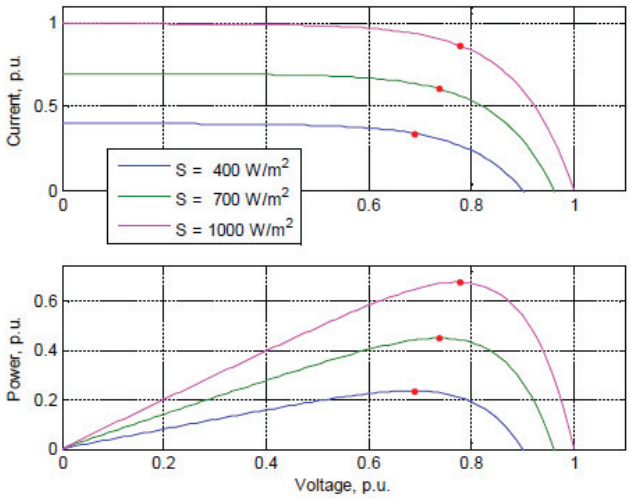


Fig. 3. Variations in IV and PV curves under Varying Irradiance

Variations in I-V and P-V curves under conditions of varying irradiance can be seen in Fig. 3. From the plot it can be observed that as the irradiance changes, the change in the short circuit current is more as compared to open circuit, this is due to fact that short circuit current depends on the photogenerated current given by

$$I_{ph} = \left[(I_{sc} + K_i(T - T_{ref})\right]\frac{G}{1000} \quad (4)$$

Where,

T_{ref} Reference Temperature

G Irradiance

K_i Short circuit coefficient of Temperature

and the Photogenerated current further depends on the amount of irradiance falling on the PV Module.

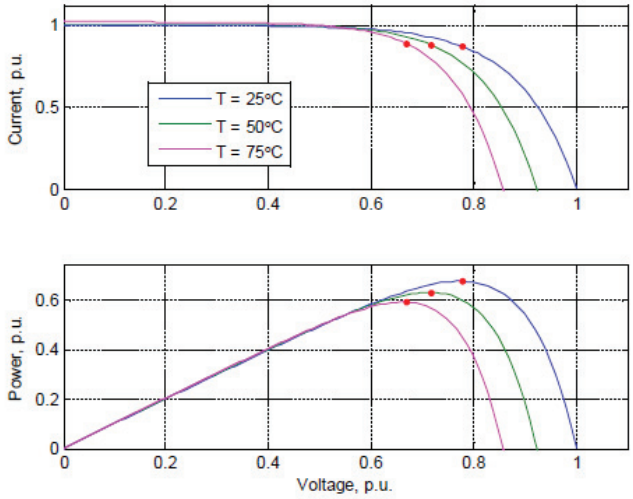


Fig. 4. Variations in I-V and P-V curves under Varying Temperature

Variations in IV and PV curves under varying Temperature conditions is presented in Fig. 4. From the plot it can be concluded that as the temperature varies, changes in open circuit voltage is more as compared to the short-circuit current.

II. LYAPUNOV STABILITY THEORY

A. Non Linear Design Tools

Equation (3) represents a logarithmic relationship between output voltage and current of PV system. Since the solution for nonlinear differential equations are generally difficult to obtain and frequency domain transformation does not apply so tedious efforts have to be made to develop appropriate tools for finding its solution as given in [3].

According to this theory, the equilibrium points are said to be stable if all the system solutions beginning at neighborhood points stays nearby, otherwise they are unstable. Those points are asymptotically stable if all solutions beginning at neighborhood points not only stays nearby, also be likely to move towards the equilibrium point as time approaches infinity. It uses the idea for constructing a scalar energy-like function also known as Lyapunov Function of the system state and then choosing an appropriate control law to make this function decrease i.e. its first derivative should be negative

$$\dot{V} \leq 0 \quad (5)$$

There are several nonlinear design tools to meet control objectives and design specifications. The ones chosen for analysis are:

1. Conventional - Sliding Mode Control (CSMC)
2. Terminal - Sliding Mode Control (TSMC)

B. System Design and Modelling

Boost type converter is connected to the PV System with a resistive load as shown in the Fig. 5.

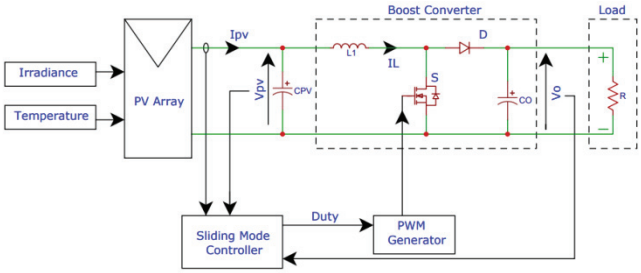


Fig. 5. PV System with Boost Converter Schematic

Mathematical modelling of the system can be carried out using state space averaging technique as in [4]. The system can be expressed in 2 set of state space equations depending upon the position of the switch S.

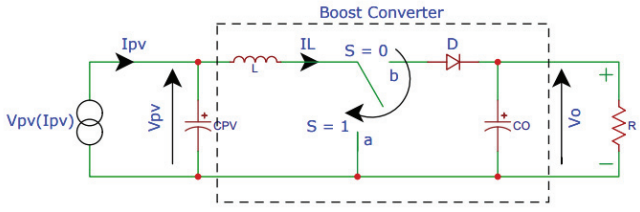


Fig. 6. Boost Converter Operation

When the switch is at position 'b' i.e. S = 0 (OFF) the system differential equations can be written as:

$$\frac{dI_L}{dt} = \frac{V_{PV}}{L} - \frac{V_o}{L} \quad (6)$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:04:06 UTC from IEEE Xplore. Restrictions apply.

$$\frac{dV_o}{dt} = \frac{I_L}{C_o} - \frac{V_o}{C_o R} \tag{7}$$

When the switch is at position 'a' i.e. S = 1 (ON) the system differential equation is written as:

$$\frac{dI_L}{dt} = \frac{V_{PV}}{L} \tag{8}$$

$$\frac{dV_o}{dt} = -\frac{V_o}{C_o R} \tag{9}$$

Based on concept of Pulse - Width Modulation (PWM), ratio of switch position at point 'a' to the time period is known as duty ratio.

If we consider "d" as the duty ratio for the control signal "u" provided to switching Mosfet, dynamics of the system can be rewritten as follow:

$$\frac{dI_L}{dt} = \frac{V_{PV}}{L} - (1-d)\frac{V_o}{L} \tag{10}$$

$$\frac{dV_o}{dt} = \frac{I_L}{C_o}(1-d) - \frac{V_o}{C_o R} \tag{11}$$

where 0 < d < 1

TABLE I. SYSTEM SPECIFICATIONS

Converter and PV Module Parameters			
Parameters	***Value***	***Parameters***	***Value***
Cpv	2000 μF	Maximum Power(MP)	305.226 W
L	10 mH	Cells per Module (n_s)	96
Co	2000 μF	V_{OC}	64.2 V
R	70 Ω	I_{SC}	5.96 A
Fsw (Switching Freq.)	50 KHz	V_{mp}	54.7 V
Vd (Diode Vol. Drop)	0.82 V	I_{mp}	5.58 A
A	0.94497	K_i	0.061 / ℃

III. SLIDING MODE METHODOLOGY FOR MPPT

Sliding Mode Control is a nonlinear control tool where the trajectories of the system are forced to follow or reach a sliding manifold also known as sliding (switching) surface in a finite duration of time and trajectories are made to stay on these switching manifolds for all future time. It is a variable structure control method that can switch form one continuous structure to another depending on the present position in the state space as presented in[5].

It has two modes of operation:

1. Approaching Mode – The system states converges to predefined switching manifold in a finite time duration.
2. Sliding Mode – System states are confined on the sliding surface and are forced toward origin.

A. Conventinal - Sliding Mode Control (CSMC)

First step involves selecting a switching manifold that will guarantee the system trajectories to hit the sliding surface as in [6]-[7] and produces maximum output power repeatedly.

We have selected the sliding surface corresponding to the point where Maximum power occurs. So the switching manifold can be mathematically expressed as

$$\frac{\partial P_{pv}}{\partial I_{pv}} = 0 \tag{12}$$

now,

$$\frac{\partial P_{pv}}{\partial I_{pv}} = \frac{\partial (I_{pv}^2 * R_{pv})}{\partial I_{pv}} = I_{pv}\left(2R_{pv} + I_{pv}\frac{\partial R_{pv}}{\partial I_{pv}}\right) = 0 \tag{13}$$

Where the term $R_{pv} = V_{pv}/I_{pv}$ refers to equivalent load connected to the photovoltaic system. So, the sliding surface is expressed as :

$$\sigma = 2R_{pv} + I_L \frac{\partial R_{pv}}{\partial I_L} \tag{14}$$

have assumed that $I_{pv} \cong I_L$

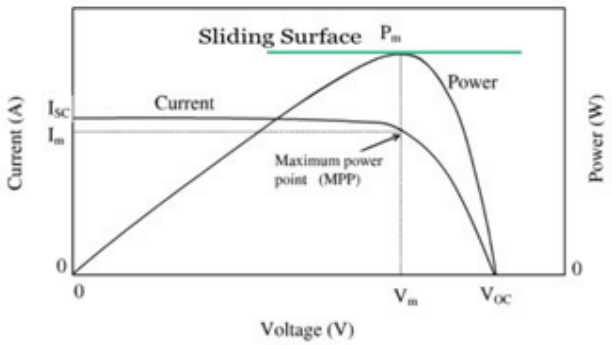


Fig. 7. Selecting switching surface correspoinding to MPP

The updation in duty cycle provided to the control the operation as depicted in Fig. 8. could be preferred as:

$$d_{update} = \begin{cases} d + \Delta d, & \sigma > 0 \\ d - \Delta d, & \sigma < 0 \end{cases} \tag{15}$$

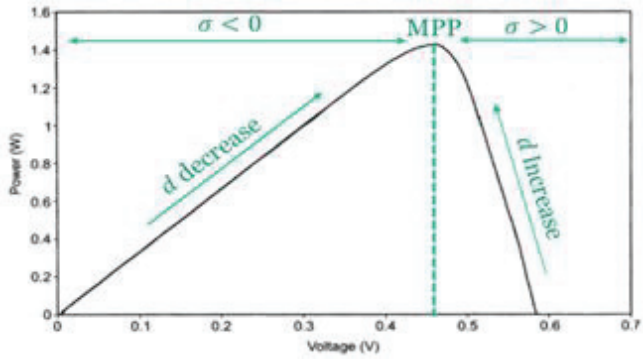


Fig. 8. Updation in duty cycle

Equivalent control is determined from following condition:

$$\dot{\sigma} = \frac{d\sigma}{dt} = 0 \tag{16}$$

$$= \frac{d\sigma}{dI_{pv}}\frac{dI_{pv}}{dt} = 0$$

$$= \frac{d\sigma}{dt} * \left[\frac{V_{pv}}{L} - (1-d)\frac{V_o}{L}\right] = 0$$

The trivial solution for the above condition can be given by:

$$d_{equ} = 1 - \frac{V_{pv}}{V_o} \tag{17}$$

The real control signal can be given as:

$$d = \begin{cases} 1, & d_{equ} + Ksign(\sigma) \geq 1 \\ d_{equ} + Ksign(\sigma), & 0 < d_{equ} + Ksign(\sigma) < 1 \\ 0, & d_{equ} + Ksign(\sigma) \leq 0 \end{cases}$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:04:06 UTC from IEEE Xplore. Restrictions apply.

K is a positive constant.

Evaluating system under the following varying irradiance conditions shown in Fig. 9.

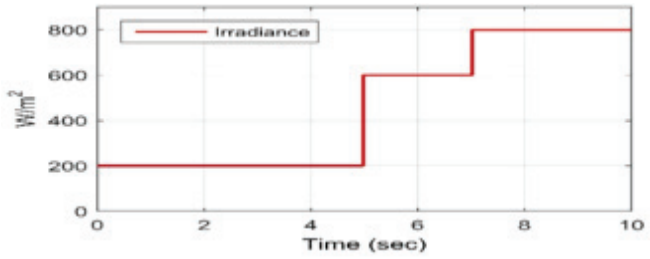


Fig. 9. Irradiance Curve

It is evident from plot shown in Fig. 10. that duty cycle changes with the irradiance, suggesting that MPP tracking algorithm is correct.

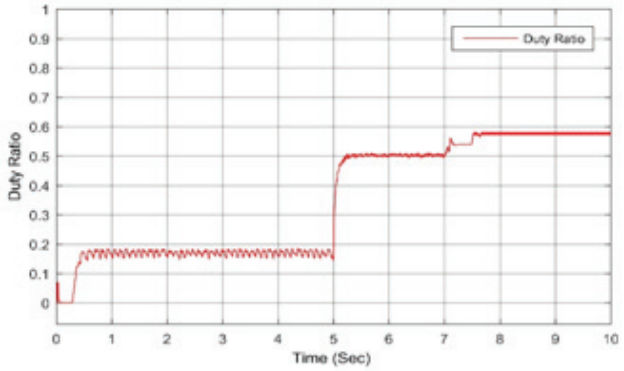


Fig. 10. Variation of Duty Cycle with the Irradiance

Because of the delay in the switching devices the CSMC suffers from the chattering phenomena as given in [8]. It occurs when the system operating trajectories heading towards the switching manifold trying to slide on the surface. When the trajectories hits the surface, ideally it should glide on the surface but due to delay in switching devices it crosses the surface into the region s < 0. When the control switches, the trajectories reverses its direction and moves towards the switching manifold. The repetition of this process causes the 'zig-zag' motion known as chattering shown in Fig. 11.

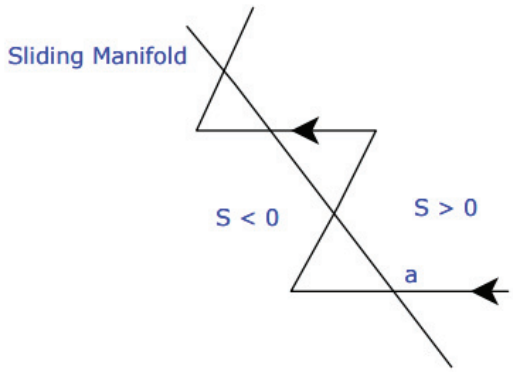


Fig. 11. Chattering Phenomena

The plot for the Ideal Power and PV Output Power under the varying irradiance is shown in Fig. 12

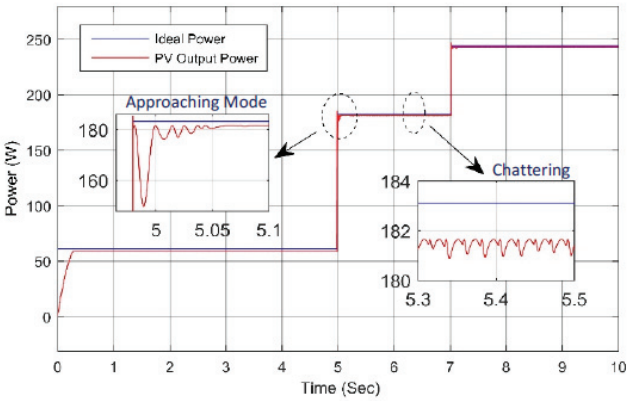


Fig. 12. Ideal and PV Output Power

It evident form the subplot that approaching mode of sliding mode control is achieved in finite duration of time however, it suffers from overshoots and undershoots in reaching the switching surface.

Further Analyzing the Approaching Mode shown in Fig.13.

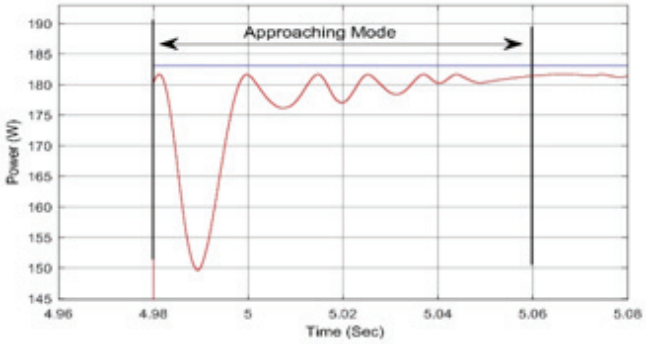


Fig. 13. Analyzing the Approaching Mode Subplot

TABLE II. APPROACHING MODE PARAMETERS

Parameters	*Measurements*
Overshoot	3.646%
Undershoot	-3.646%
Approaching Time	83.116ms

Main function of MPPT algorithms is to obtain the Maximum Output power from Photovoltaic Systems under all irradiance conditions and this is achieved by using a DC/DC converter, to make Output power of Converter same as that Output power of PV System shown in Fig. 14.

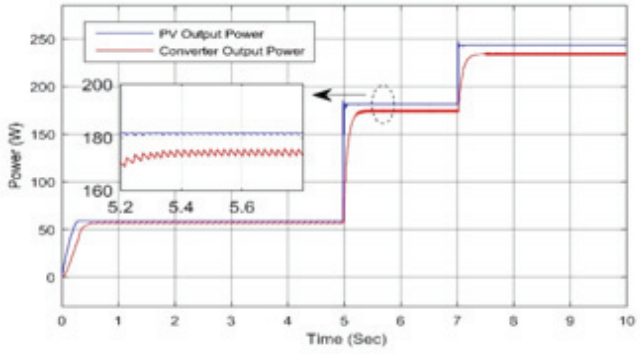


Fig. 14. PV Module and Converter Output Power

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:04:06 UTC from IEEE Xplore. Restrictions apply.

we can be observed that due to the chattering the output power of the converter suffers from oscillations leading more heat losses in power and switching circuits, may lead to high frequency dynamics and it is difficult to comment on the converter efficiency.

B. Terminal - Sliding Mode Control (TSMC)

TSMC takes into consideration both the voltage and current tracking errors and a sliding surface is such designed to converge these errors to zero in a fixed time duration as presented in [9]-[10]. Convergence of these errors to 0 will guarantee for the existence of controller action. TSMC more robustly drives the system trajectories towards the switching manifold described by $\sigma(t)$.

$$e_1 = V_{PV} - V_{PV}^* \tag{18}$$

$$e_2 = I_L - I_L^* \tag{19}$$

Where V_{PV}^* and I_L^* are reference voltage and current, respectively.

Steps involved are:

1. Developing an MPV(Maximum Power Voltage) Algorithm that will provide the reference voltage as per irradiance and temperature conditions.
2. Calculating the reference current expressed as $I_L^* = I_{PV} - C_{PV}V_{PV}^*$. Here considered both I_{PV} & I_L to be different.
3. Both the tracking errors e_1 and e_2 are then given to the controller, where based on the appropriate switching manifold these errors are converged to zero in fixed time duration.

The schematic for the terminal - sliding mode controller can be expressed as shown in Fig. 15.

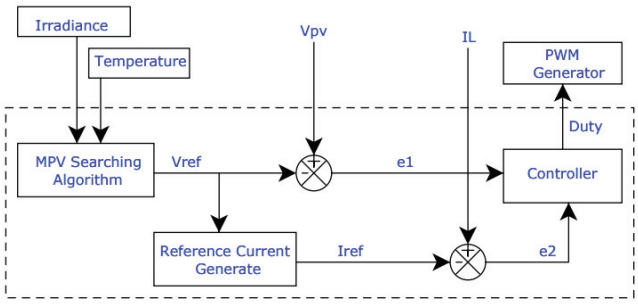


Fig. 15. Terminal Sliding Mode Controller

A numerical method based on linear dependency among the open circuit voltage and reference voltage (optimal voltage) can be used to find the operating point of system subsequent to MPP under varying irradiance and temperature conditions as given in [11]. The equation is given as

$$V_{PV}^* = K_v V_{oc} \tag{20}$$

Where V_{oc} and K_v are the open circuit voltage of the Photovoltaic System and voltage proportionality constant respectively. The value of K_v depends on the material and structure of PV Module, generally lying between 0.77 to 0.83.

On ignoring the shunt current under open circuit conditions (2) is rewritten as:

$$I_{ph} - Io\left\{\exp\left[\frac{V_{oc}}{n_s V_t}\right] - 1\right\} = 0 \tag{21}$$

Open circuit voltage is expressed as

$$V_{oc} = n_s V_t \, ln\left\{\frac{I_{ph}+I_d}{I_d}\right\} \tag{22}$$

$$V_t = \frac{AK_bT}{e} = \text{Temperature Equivalent Voltage} \tag{23}$$

$$I_o = I_{or}\left(\frac{T}{T_{ref}}\right)^3 \exp\{\frac{qE_g}{n_s K_b T}\left(\frac{1}{T_{ref}} - \frac{1}{T}\right)\} = \text{Diode Current}$$

Where,

K_b = Boltzmann Constant

A = Ideality Factor

E_g = Band gap Energy

K_i = Short circuit coefficient of Temperature.

T_{ref} = Reference Temperature

The plot for reference voltage and current is shown in Fig. 16.

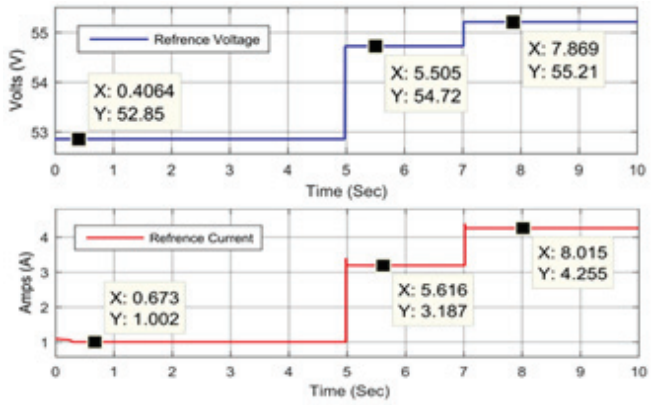


Fig. 16. Refrence Voltage and Refrence Current

TABLE III. % ERROR IN MPV VALUES

Irradiance Value (W/m^2)	*MPV Values from PV Module*	*MPV Values from Algorithm*	*% Error*
200	52.61 V	52.85 V	0.45
600	54.35 V	54.72 V	0.68
800	54.45 V	55.21 V	1.02

The aim is to design an appropriate sliding or switching manifold $\sigma(t)$ that will drive these errors to 0 in finite time duration. Terminal Sliding Mode Controller will cause all the system trajectories to approach towards the manifold and further stays or glide over the surface forever by the action of control law i.e. varying the duty cycle ratio of converter as given in [12]-[13].

Sliding surface for providing the control input is chosen as:

$$\sigma(\mathrm{t}) = \frac{1}{\beta}e_2^r - e_1 \tag{24}$$

Where $\beta > 0$, is a positive constant

$r = {}^{p}/_{q}$ with $1 < r < 2$ and p and q are odd integers.

The error dynamics can be described by taking the first derivate of tracking errors:

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:04:06 UTC from IEEE Xplore. Restrictions apply.

$$e_1^{\cdot} = V_{PV}^{\cdot} - V_{PV}^{*\cdot} = \frac{1}{C_{PV}}\left(I_{pv} - I_l\right) - V_{PV}^{*\cdot} = -\frac{e_2}{C_{PV}} \quad (25)$$

$$e_2^{\cdot} = I_L^{\cdot} - I_L^{*\cdot} = \frac{V_{pv}}{L} - (1-d)\frac{V_o}{L} - I_L^{*\cdot} \quad (26)$$

The sliding controller is satisfied when $\sigma(\mathrm{t}) = 0$ then

$$e_1 = \frac{1}{\beta} e_2^r \quad (27)$$

$$e_2 = e_1^{1/r} \beta^{1/r} \quad (28)$$

Consequently the $e_1^{\cdot}$ becomes

$$e_1^{\cdot} = -\frac{\beta^{1/r}}{C_{PV}} e_1^{1/r} \quad (29)$$

This represents the dynamics of voltage tracking error, for satisfying the sliding condition it is required that e_1 should converge to zero. Which will guarantee for the convergence of tracking error e_2 to zero shown in Fig. 17.

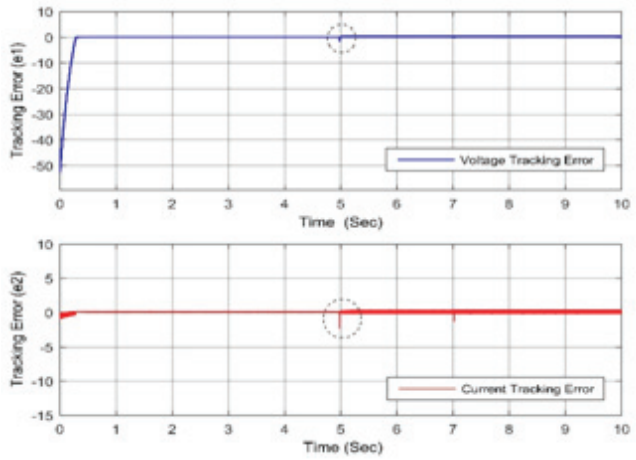


Fig. 17. Convergence of Tracking Errors e1 and e2

The final step involves developing an appropriate control law based on the switching manifold that will cause the tracking errors e_1 and e_2 to converge zero. The control law is basically providing the duty ratio to the switching Mosfet for guiding the system trajectories towards switching manifold for satisfying $\sigma(\mathrm{t}) = 0$.

Considering (10), (19) and (29), the control law can be deduced as:

$$\mathrm{d} = \frac{-L}{V_o}\left[\frac{1}{L}\left(V_{PV} - V_o\right) + \frac{\beta e_2^{2-r}}{r C_{PV}} - I_L^{*\cdot} + K.sign(\sigma)\right] \quad (30)$$

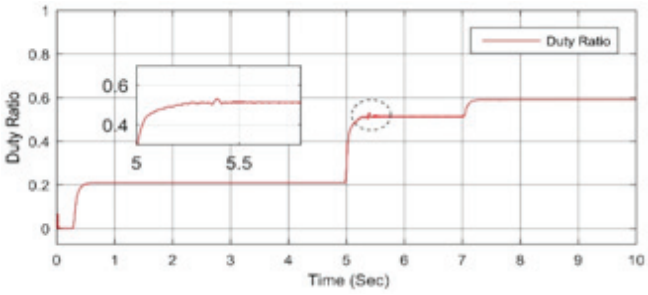


Fig. 18. Variation of Duty Cycle with the Irradiance

Fig. 18 shows the duty cycle variation under the varying irradiance conditions as per control action law in order to converge the errors to zero. As compared to duty ratio obtained from CSMC, the chattering is reduced considerably.

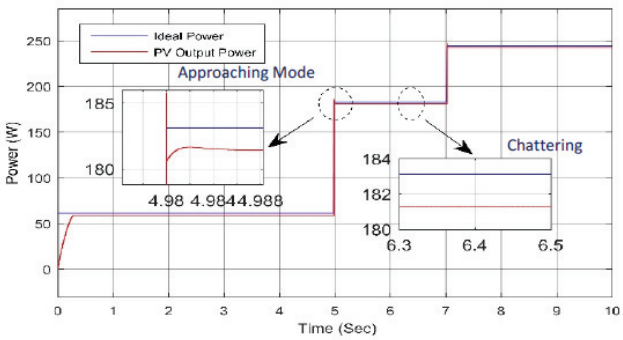


Fig. 19. Ideal and PV Output Power

From Fig. 19. we analyze that oscillations in approaching mode is reduced, thereby reducing overshoots and undershoots. Further in the sub-plots it is evident that chattering phenomena is reduced considerably as compared to CSMC as presented in [14].

IV. CONCLUSION

Two different types of Sliding Mode nonlinear controller CSMC and TSMC were developed and studied in MATLAB/Simulink also their comparative assessment have been carried out under dynamic climatic conditions. Based upon the results obtained, it is concluded that TSMC suffers from the chattering phenomena which results in the oscillations in the output power of the converter. Due to this phenomena the converter efficiency could not be determined. After that modifications have been introduced in the switching surface in TSMC which resulted into the reduced chattering and increased converter efficiency around 96%. Further by integrating (29) time taken for the error convergence can be found out. However, in terms of design complexity TSMC is more complex but accurate as compared to CSMC.

REFERENCES

[1] Bellia, H., Youcef, R., & Fatima, M. (2014). A detailed modeling of photovoltaic module using MATLAB. NRIAG Journal of Astronomy and Geophysics, 3(1), 53–61

[2] Jafari Fesharaki V, Dehghani M, Jafari Fesharaki J. The effect of temperature on photovoltaic cell efficiency. In Proceedings of the 1st International Conference on Emerging Trends in Energy Conservation – ETEC, 2011.

[3] K.-H. Chao C.-J. Li "An intelligent maximum power point tracking method based on extension theory for PV systems" Expert Syst. Appl. vol. 37 no. 2 pp. 1050-1055, Mar. 2010.

[4] Sun, J., Gratstollen, H., 1992. Averaged modeling of switching power converters: reformulation and theoretical basis. Power Electronics Specialists Conference, 1165–1172.

[5] Battista, H. De, Mantz, R.J., 2002. Variable structure control of a photovoltaic energy converter. IEE Proc. Control Theor. Appl. 149,303–310.

[6] Rekioua, D., Achour, A. Y., & Rekiouaa, T. (2013). Tracking power photovoltaic system with sliding mode control strategy. Energy Procedia, 36, 219–230.

[7] Cabal, C., Martinez-Salamero, L., Seguier, L., Alonso, C. and Guinjoan, F. (2014) 'Maximum power point tracking based on sliding mode control for output series connected converters in photovoltaic systems' , IET Power Electronics, Vol. 7, No. 4, pp.914–923

[8] Parra-Vega, V., Hirzinger, G., 2004. Chattering-free sliding mode control for a class of nonlinear mechanical systems. International Journal of Robust and Nonlinear Control 11 (12), 1161–1178.

[9] Chiu, C-S., Ouyang, Y-L. and Ku, C-Y. (2012) 'Terminal sliding mode control for maximum power point tracking of photovoltaic power generation systems', Solar Energy, Vol. 86, No. 10, pp.2986–2995

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:04:06 UTC from IEEE Xplore. Restrictions apply.

[10] Man, Z., Yu, X.H., 1997. Terminal sliding mode control of MIMO linear systems. IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications 44 (11), 1065–1070.

[11] Jin, Y., Chang, P.H., Jin, M. and Gweon, D.G. (2013) ‘Stability guaranteed time-delay control of manipulators using nonlinear damping and terminal sliding mode’, IEEE Transactions on Industrial Electronics, Vol. 60, No. 8, pp.3304–3317.

[12] Chu, C.-C., & Chen, C.-L. (2009). Robust maximum power point tracking method for photovoltaic cells: A sliding mode control approach. Solar Energy, 83(8), 1370–1378.

[13] Mobayen, S. (2015) ‘Fast terminal sliding mode controller design for nonlinear second-order systems with time-varying uncertainties’, Complexity, Vol. 21, No. 2, pp.239–244.

[14] Zhankui, S. and Sun, K. (2013) ‘Nonlinear and chaos control of a micro-electro-mechanical system by using second-order fast terminal sliding mode control’, Communications in Nonlinear Science and Numerical Simulation, Vol. 18, No. 9, pp.2540–2548.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:04:06 UTC from IEEE Xplore. Restrictions apply.

Materials Research Express

PAPER • OPEN ACCESS

Mechanical and wear performance of Al/SiC surface composite prepared through friction stir processing

To cite this article: Ravi Butola *et al* 2021 *Mater. Res. Express* **8** 016520

View the article online for updates and enhancements.



IOP ebooks™

Bringing together innovative digital publishing with leading authors from the global scientific community.

Start exploring the collection–download the first chapter of every title for free.

This content was downloaded from IP address 14.139.251.100 on 18/01/2021 at 09:57

Materials Research Express

PAPER

Mechanical and wear performance of Al/SiC surface composite prepared through friction stir processing

Ravi Butola[1,*], **Lakshay Tyagi**[1], **Ranganath M Singari**[1], **Qasim Murtaza**[1], **Harish Kumar**[2] and **Dhiraj Nayak**[3,*]

[1] Department of Mechanical Engineering, Delhi Technological University, Delhi 110042, India
[2] Department of Mechanical Engineering, National Institute of Technology Delhi, Delhi 110040, India
[3] Department of Mechanical and Aerospace Engineering, Tribhuvan University Institute of Engineering, Pulchowk Campus, Nepal
[*] Authors to whom any correspondence should be addressed.

E-mail: ravibutola33855@gmail.com and dhirajnayak444@gmail.com

Keywords: friction stir processing, coefficient of friction, microstructure, surface composite, wear morphology

OPEN ACCESS

RECEIVED
29 October 2020

REVISED
18 December 2020

ACCEPTED FOR PUBLICATION
5 January 2021

PUBLISHED
13 January 2021

Original content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence.

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

Abstract

In the present research work, AA7075 composite reinforced with silicon carbide particles has been fabricated using Friction stir processing (FSP). The silicon carbide particles having a size of 40 μm were placed in grooves of length 160 mm, width 2 mm, depth 3.5 mm, that were generated on the AA7075 plate. The square pin tool is utilized for fabricating the composite at two different rotational speed i.e. 700 and 1000 rpm. Effect of processing, particle addition and tool rotational speed is analyzed on mechanical and wear properties of the material. On friction stir processing the microhardness value and elongation of the material increased. Reinforcement addition contributed to decrease in ductility and tensile strength while on the contrary microhardness and wear resistance of the material improved. Tool rotational speed showed a direct relation with the tested mechanical and wear properties. Adhesive wear was the prominent wear mechanism and Fe layer formation was observed on the worn surface, contributing to increased wear resistance. These fabricated composites can find vast application in industries like automotive, defence and aerospace.

1. Introduction

Composites can be defined as a mixture of two or more different materials having different physical and chemical properties [1, 2]. The fabricated composite shows different properties than its constituent materials [3]. Materials like aluminium, copper require enhanced mechanical and wear properties to widen the scope of their application. In industries, aluminium and its alloys find vast application due to their high strength to weight ratio [4, 5]. The composite materials are fabricated in order to attain better properties than their parent material. Metal matrices are reinforced with a variety of reinforcements like TiC, SiC, Al_2O_3, and B_4C in order to improve performance [6, 7]. Nowadays aluminium metal matrix composites (AMMC) find wide application in aerospace, marine industries and they are also used as cylinder liners, brake rotors, pistons [8, 9]. AMMCs are acquiring great interest of researchers also, and various aluminium composites have been fabricated and analysed by researchers. Zawawi *et al* [10] analysed the wear rate and coefficient of friction (CoF) on Al2024 + Al_2O_3-SiO_2/PAG composite nano lubricant at a constant load of 10 kg and at different speeds. The CoF and wear rate of composite nano lubricant were reduced compared to the pure lubricant. It was found that the minimum value of CoF was achieved at a 0.06% vol concentration of Al_2O_3–SiO_2/PAG composite. Naveed *et al* [11] analysed the dry sliding wear behaviour of Al6061 reinforced with SiC with graphite particulates. Liquid metallurgy route was used for fabricating the composite, taking graphite content up to 4 wt% and taking constant silicon carbide content of 7 wt%. Due to the low density of Gr, it is not possible to reinforce composite with more than 4 wt% of SiC. Wear resistance increased on addition of SiC and it further increased with increasing content of Gr till 4 wt%. Heat treatment had an intense effect on wear properties; of all processes, ice quenching showed the best improvement in wear resistance for all samples.

© 2021 The Author(s). Published by IOP Publishing Ltd

Caliman [12] analysed the wear properties of aluminium composite having short carbon fibre as reinforcement. It was observed that due to the development of lubrication film, CoF and wear rate decreased with time and reached a static value. Raghavendra *et al* [13], by electro-deposition, fabricated a Ni-Al_2O_3 coating on an aluminium substrate. They produced a design of experiment using central composite design, taking process variables like coating density, particles loading and bath temperature as varying parameters. For all developed composites, they observed the surface roughness before and after the coating. They found that with the presence of Al_2O_3, wear resistance increased and better wear resistance was achieved. Shankar *et al* [14] fabricated AA8011 + Gr composite by two-step stir casting process at varying weight% of reinforcement and analysed the fabricated composites for its wear properties. They found that with increasing load, the wear of material increased. With the incorporation of reinforcement, wear resistance increased, and it further increased on increasing the graphite content in the composite. Razzaq *et al* [15] reinforced AA6063 with varying wt% of fly ash particulate with compo-casting technique. They investigated the effect of applied load and fly content on tribological properties of the fabricated composites. They found that on increasing the applied load, the wear increased and it decreased with the increasing content of particulate. Nagaral *et al* [16] studied the dry sliding behaviour of Al6061 hybrid metal matrix composite reinforced with silicon carbide and graphite, fabricated using vortex method. It was observed that with increasing load wear loss increased. With the introduction of reinforcement, wear decreased as most of the load on the matrix was bearded by hard ceramic particles. Wear loss increased with the increasing sliding speed, though the increase was not constant. The microstructures revealed that the distribution of reinforcement throughout the matrix was uniform. Mittal *et al* [17] analysed the abrasive wear of aluminium composite reinforced with SiC, red mud and Al_2O_3. Stir casting process was used for fabricating the composite. It was observed that wear resistance was improved with the addition of reinforcement of Al_2O_3, red mid and silicon carbide. With increasing concentration of SiC, Al_2O_3 an increase in wear resistance was observed whereas for red mud wear resistance increased till 7.5 wt% and on further addition, it decreased.

Baradeswawaran *et al* [18] used stir casting method to fabricate Al7075 composite reinforced with B_4C particles. They analysed the fabricated composites for mechanical and tribological properties. Increased hardness was observed in the fabricated composites due to the presence of hard ceramic particles. They found that due to the formation of mechanical mixed layers, wear loss and CoF decreased as metal-to-metal contact was avoided. With the introduction of reinforcement, CoF and wear loss decreased and it further decreased with the increasing B_4C content and achieved its minimum value at 10 vol% B_4C. Rahman *et al* [19] analysed the aluminium composite reinforced with SiC for its mechanical and wear properties. Stir casting process was used for fabricating the composite with varying concentration of silicon carbide i.e. 0, 5, 10, 20 wt%. It was found that hardness, tensile strength and wear resistance were maximum for the composite with 20 wt% of SiC. With the increasing content of SiC, increased porosity was observed. Non-homogeneous and clustering distribution on reinforcement was observed in the generated microstructures. Pradhan *et al* [20] reinforced 5 wt% silicon carbide with LM6 base by stir casting process and under corrosion observed its wear behaviour. Wear test was carried out on different normal load and sliding speed, for a fixed time period under different environments i.e. deionised, dilute acid and dry. Due to acidic environment being most corrosive of all, maximum wear loss occurred here; wear loss increased with increasing load in all three conditions. Through SEM, adhesive and abrasive wear traces are observed on the wear surface. Tirth *et al* [21] developed aluminium hybrid aluminium matrix composite via stir casting and reported the effect of reinforcement and ageing on the wear behaviour of the material. He concluded that with the increasing sliding distance wear loss increases, wear rate increases with the load and decreases with the addition of reinforcement. Also, aging contributed to decrease in wear rate and volume loss in unreinforced alloy. Bansal *et al* [22] observed the mechanical and tribological properties of Al 359 matrix reinforced with silicon carbide and silicon carbide/graphite particles. The composites are fabricated using stir casting process. It was found that hardness of composite reinforced with SiC/Gr was more than that of composite reinforced with SiC. With the introduction of reinforcement wear resistance of the alloy increased and composite with SiC/Gr reinforcement showed the superior wear resistance of all.

Friction stir processing (FSP) is used for surface modification and manufacturing of composite, by localised plastic deformation [23]. Enhanced tribological, mechanical and corrosion properties of material can be achieved by using this technique [24]. This process is the succedent of Friction stir welding (FSW). Yuvaraj *et al* [25] fabricated AA5083 composites reinforced with boron carbide, titanium carbide and with a combination of both by friction stir processing. On evaluating for mechanical and wear properties, they found that the composite reinforced with boron carbide showed the highest tensile strength and microhardness, whereas superior wear properties were shown by the hybrid composite. They utilised SEM for analysing the worn out surfaces. Dhayalan et [26] fabricated AA6063 composites reinforced silicon carbide, graphite and a hybrid of both via friction stir processing. FSP was carried out using a circular pin profile tool made of high carbon high chromium. On analysing its hardness it was found that on the incorporation of reinforcements, the hardness was enhanced and it showed a maximum value for the composite reinforced with silicon carbide. Microstructural changes were governed by scanning electron microscopy (SEM). Dolatkhah *et al* [27] reinforced

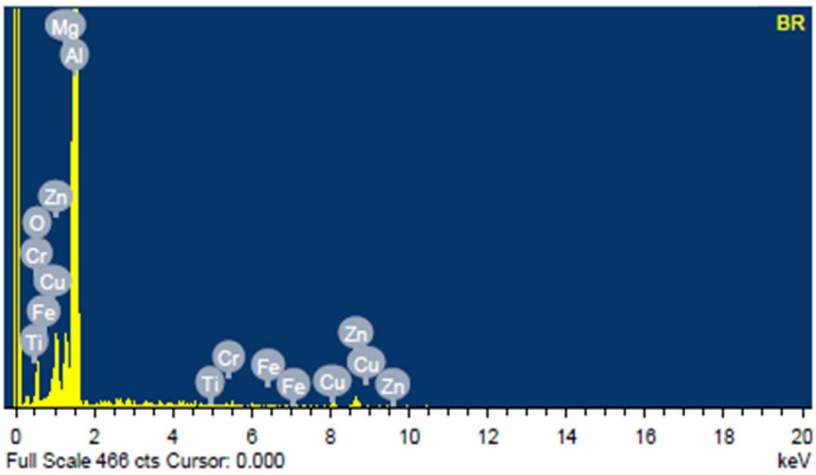


Figure 1. EDS of 7075-T6 aluminium alloy.

silicon carbide with aluminium 5052 sheets via friction stir processing at different process parameters. They studied the influence of particle size, tool rotation direction between passed, number of passes, tool rotational and traverse speed on wear properties and microhardness. It was found that for proper dispersion of reinforcement, optimum values of tool rotational and traverse speed should be achieved. Saadatmand *et al* [28], using friction stir processing, fabricated Al6061/SiC composite. They studied the effect of mass function SiC distribution on wear and mechanical properties of the fabricated composites. Hardness increased and elongation decreased on increasing the mass fraction of SiC. The optimum functional graded composite obtained for better wear resistance and tensile strength was the composite having 10 wt% of SiC. Due to clustering, the tensile strength and wear resistance decreased on further addition of SiC. Butola *et al* [29] by FSP, along with self-assembled monolayer (SAM), fabricated AA7075-T6 + B_4C composite. They found that wear and frictional resistance behaviour enhanced with the addition of reinforcement particulate. Butola *et al* [30] In their study analysed the effect of different process parameters i.e. tool rotational speed, tool pin profile and reinforcement type, using the ANOVA for better microhardness value.

The primary objective of the present research work is to create a layer of SiC particles mixed with a viscous liquid, Di-ethyl ether and fabricate aluminium composite via friction stir processing technique. This is done to analyse the effect of the processing, tool rotational speed and addition of reinforcement, on the mechanical and wear properties of the material.

2. Experimental procedure

2.1. Materials

7075-T6 aluminium alloy is taken as matrix material because of its excellent strength to weight ratio. Plates having dimensions 200 mm × 80 mm × 6 mm are taken and groove of length 160 mm width 2 mm, depth 3.5 mm are generated on the plates using a milling cutter. This alloy of aluminium has good resistance, high strength and toughness, and improved mechanical properties. It is mostly utilized in the aerospace industry for structural parts. Figure 1 shows EDS of the base alloy. Silicon Carbide (SiC) powder having a particle size of 40 μm is used as a reinforcement for the study. SiC is taken as reinforcement material because it exhibits enhanced mechanical properties and good wear resistance [31]. SiC find wide application in semiconductors, turbines, pumping equipment, ball valve parts etc. Prior to FSP, SiC particles are mixed with Di-ethyl ether and a viscous paste is generated. Then this paste is filled in the grooves that have been made on the aluminium plates.

2.2. Fabrication of surface composite

FSP is used for fabricating the surface composite in this work. FSP is carried out on a semi-automatic machine setup. Clamping of workpiece, vertical and traverse motion of tool are all governed by a hydraulic system. Before processing the composite, capping pass is done with a pinless tool. This is done to protect the viscous solution from escaping during the actual processing. Then FSP is carried out with a non-consumable square pin tool made of H13 tool steel. The tool has a total length of 117 mm, pin length 3.5 mm and the tool is made to have a flat shoulder profile with a shoulder diameter of 19.95 mm. The tilt angle is 2°, kept constant for all fabricated

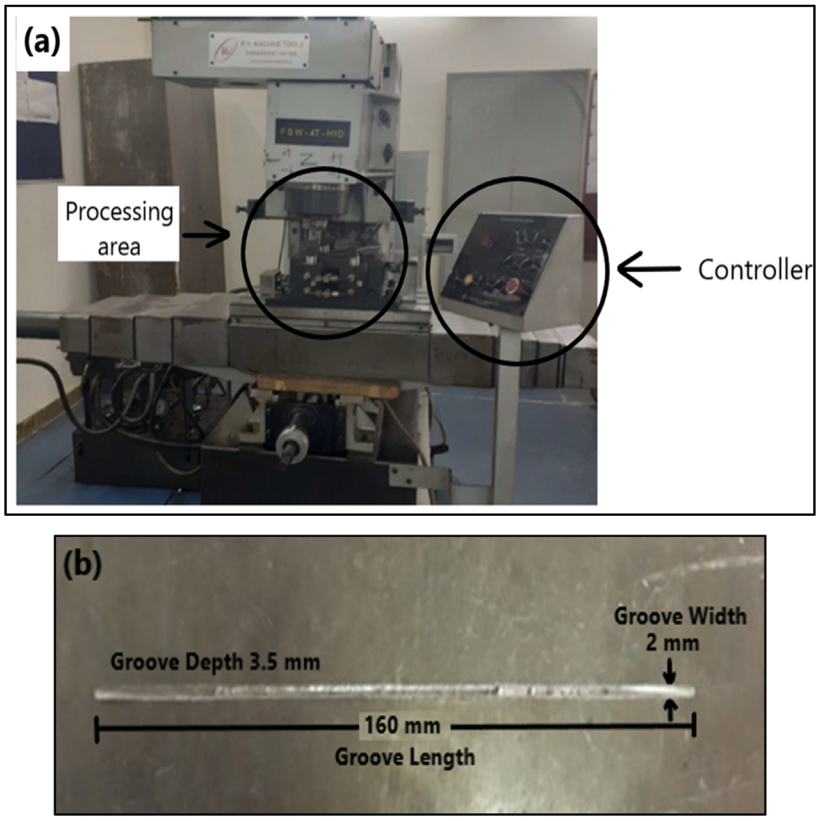


Figure 2. (a) Friction stir processing set up (b) Grooved on the base metal.

Table 1. Composition of processed composites.

Sample	Composition	Tool rpm
Sample 1	AA7075	700
Sample 2	AA7075 + SiC (2 wt%)	700
Sample 3	AA7075 + SiC (2 wt%)	1000

composites. Figure 2 shows the FSP set up used for fabricating. Tool rotational speed i.e. 700, 1000 rpm, is selected based on the literature review [30]. The tool is rotated at selected rpm and inserted in the clamped workpiece gently. For some time, it is made to rotate to produce heat and plastic deformation and then it is made to travel in the traverse direction. Table 1 shows all processed composites. Three different composites are fabricated i.e. sample 1, sample 2 and sample 3. Sample 1 is FSPed of base metal without reinforcement, sample 2 is AA7075 + SiC (2 wt%) FSPed at 700 rpm and sample 3 is AA7075 + SiC (2 wt%) FSPed at 1000 rpm. For this work, the projection of tool pin is mentioned in equation (1), area of groove in equation (2) and Fraction volume of reinforcement in equation (3) respectively.

$$Projection\ of\ tool\ pin\ =\ pin\ length\ *\ square\ base\ diagonal \tag{1}$$

$$Groove\ area\ =\ Depth\ *\ Width\ of\ groove \tag{2}$$

$$Volume\ fraction\ of\ particle\ =\ \frac{Area\ of\ groove}{Projection\ of\ tool\ pin} \tag{3}$$

2.3. Microhardness

Fisherscope HM2000S is utilised to generate Vickers microhardness value of the prepared samples. In this process, an intender is made to press on the surface with predefined force, for a limited period of time. Diagonals of the produced indentation are measured and quantified giving the microhardness of the sample. Four readings

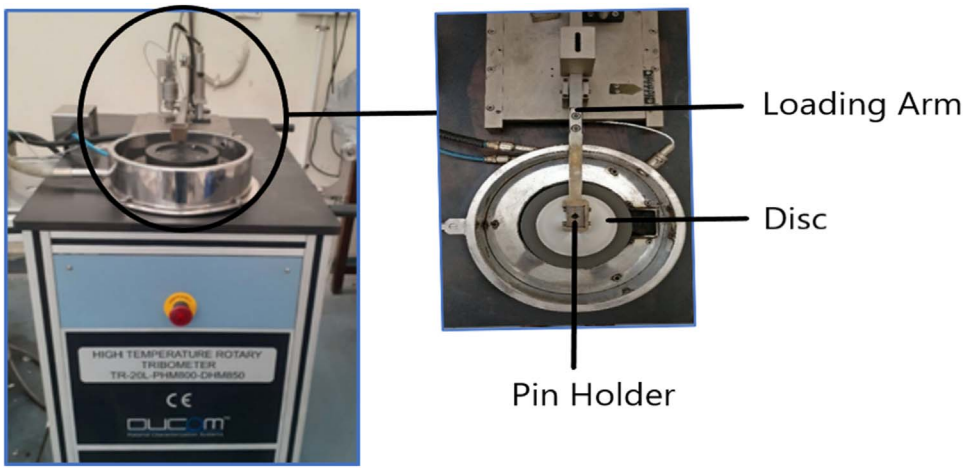


Figure 3. High temperature rotatory tribometer set up.

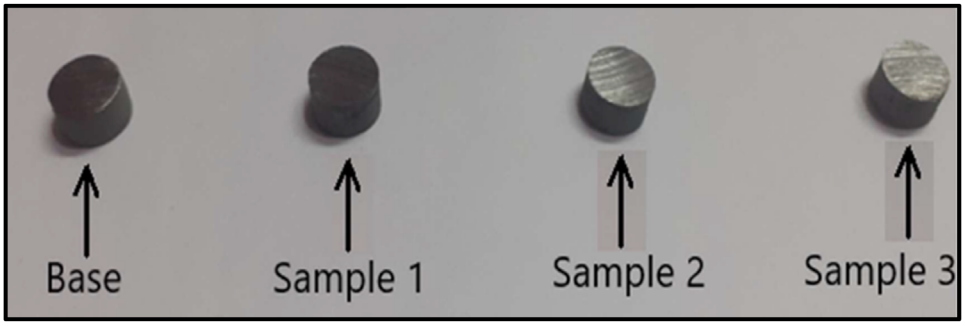


Figure 4. Cylindrical pins used for wear testing.

are taken for each sample and average of the four is considered as the microhardness value of the sample. The indenter is in dwell for 20 s with a force of 0.3 N.

2.4. Tensile testing

The tensile test of the base alloy and FSPed samples was analysed using a computer-controlled universal testing machine (Tinius Olsen H50KS) at a constant crosshead speed of 1 mm/min at room temperature. Dog bone-shaped samples prepared according to ASTM: E8/E8M-011 standard with a length and thickness of 80 mm and 3.5 mm respectively, were used for tensile testing. The samples are clamped from both sides and applied with a force till the breaking of the sample is achieved. The samples i.e. Base, sample 1, sample 2, sample 3, were recorded for their ultimate tensile strength (UTS) and elongation in order to analyse the effect of processing, tool rotational speed and reinforcement on the material.

2.5. Wear testing

Pin on disc wear tests is performed on all three fabricated composites i.e. sample 1, sample 2, sample 3 and also on the base metal. Wear testing is performed using DUCOM manufactured High Temperature Rotary Tribometer shown in figure 3. In order to carry out the experiment cylindrical shaped samples of length 6 mm and diameter 10 mm are carved out from processed region, shown in figure 4. Electric discharge machining (EDM) process is used for carving out the cylindrical samples. A total of 12 pins are used for wear test and all pins were weighed before the wear test. 3 from each sample 1, sample 2, sample 3 and base metal. Pins are slid along the disc made of EN24 steel having a hardness of 56 HRC. Wear is tested at different applied loads of 20 N, 30 N, 40 N at corresponding track diameter of 30 mm, 40 mm, 50 mm. Pins are slidfor a distance of 950 m in a time period of 16 min. After the completion of test, pins are weighed again and the difference in weights before and after the test is then analysed.

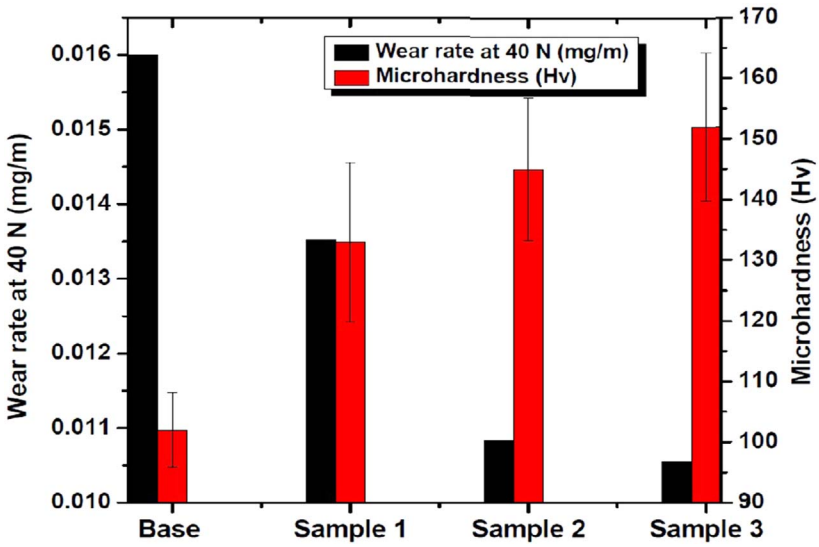


Figure 5. Microhardness and wear rate at 40 N of all tested samples.

3. Results and discussion

3.1. Microhardness

Figure 5 shows the recorded microhardness for all samples. It can be seen that minimum microhardness recorded was for base metal, having a value of 102 Hv. Recorded microhardness for sample 1 is 132 Hv, this enhancement in microhardness might be because of grain refinement due to processing. Yuvaraj *et al* [25] reported similar type of results in their study. Microhardness of sample 2 is 145 Hv, which is more than that of sample 1; this increase can be attributed to well-dispersed SiC particles in the matrix [32]. Sample 3 showed a microhardness value of 152 Hv. On comparing the microhardness value of sample 2 and sample 3 it can be concluded that on increasing the tool rpm the hardness of the material is enhanced. This increase in hardness might be attributed to the grain boundary refinement because at higher rpm finer grains are produced [33]. Decrease in wear rate of sample can be attributed to the increase in hardness of samples i.e. the harder the sample, less it will wear out during the process.

3.2. Tensile testing

The effect of reinforcement and processing on ultimate tensile strength (UTS) can be observed in figure 6. All FSPed samples showed decreased UTS value when compared to that of base metal. There are a variety of microstructural properties on which tensile properties depends i.e. dislocation density, grain size and interaction reinforcement and the base metal. This decrease in UTS might be due to softening of the matrix due to FSP [34]. The reinforced specimen displayed less UTS value compared to that of FSPed material without reinforcement. Similar type of results was reported by Yuvaraj *et al* [35] and Butola *et al* [29] in their studies on aluminium composites. The UTS of material processed at 1000 rpm is more than that of the material produced at 700 rpm. This might be attributed to the finer grains produced at high tool rpm [36]. It is well observed from figure 6, that on processing the material without reinforcement, the elongation increases. The grain refinement due to processing increases the plastic deformation of the material, adding to elongation of the material [34]. Due to the hindrance in the mobility of dislocation with the addition of SiC, the elongation percentage decreases and the recorded elongation was little more for the material processed at 1000 rpm than the sample processed at 700 rpm, because of smaller grains produced at higher rpm.

3.3. Microstructural analysis

Microstructure images shown in figure 7 are generated using Olympus GX41 inverted metallurgical optical microscope at 50 μm. The conversion scale of 2 pixels per microns is used for the measurement of the microstructures. Microstructure holds great importance when analysing the processing effect of techniques like FSP. On evaluating figures 7(a) and (b), it can be concluded that there is extreme plastic deformation and grain refinement in base metal when processed using FSP i.e. sample 1. This confirms that improved properties are because of grain refinement. It can be observed that grains in figure 7(d) are finer compared to the grains in figure 7(c), this confirms that on processing the material at higher tool rpm more fine grains are produced.

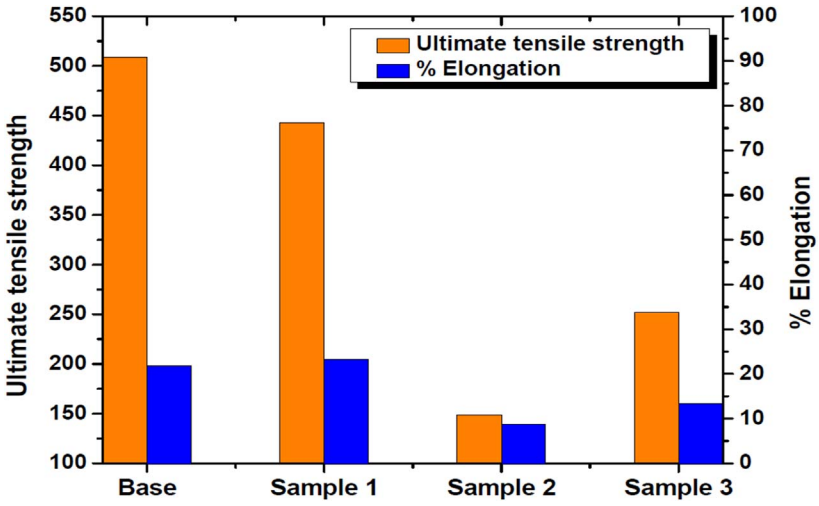


Figure 6. Ultimate tensile strength and elongation percentage of all tested samples.

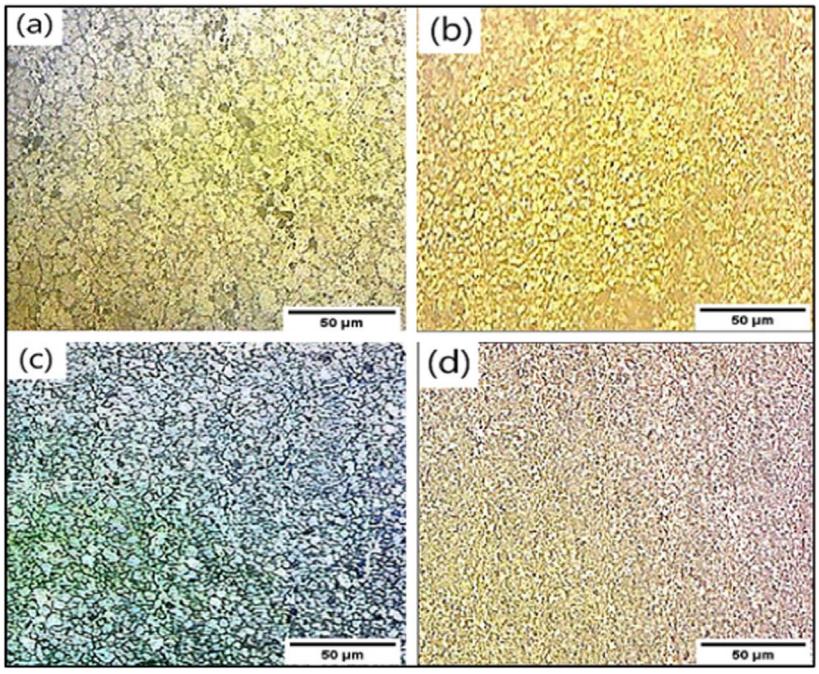


Figure 7. Microstructures of samples at 50 μm (a) Base (b) Sample 1 (c) Sample 2 (d) Sample 3.

3.4. Weight loss

Weight loss of all tested samples at different applied load is shown in figure 8. For all loads it can be observed that maximum weight loss is shown by the base metal followed by FSPed base metal, then by AA7075 + SiC composite fabricated at 700 tool rpm and minimum weight loss is shown by the AA7075 + SiC composite fabricated at 1000 tool rpm. From observing weight loss for base and sample 1 it is evident that on processing the weight loss decreases. This might be because of the grain refinement due to FSP [26]. Sample 1 shows less weight loss than sample 2 for all applied load. This shows that on incorporating the matrix with reinforcement wear resistance increases. This is because the load from aluminium alloy is transferred to the hard-ceramic particles introduced in the matrix [37]. Sample 3 shows less weight loss than sample 2, which shows that on fabricating composite at higher tool rotational speed, weight loss decreases. This is because less cavities and defects are generated at higher tool rotational speed and proper reinforcement distribution is there at high tool rpm [38].

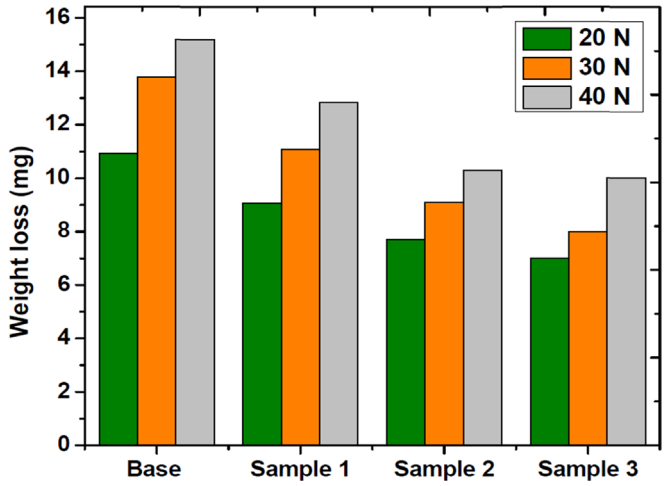


Figure 8. Weight loss of all tested samples.

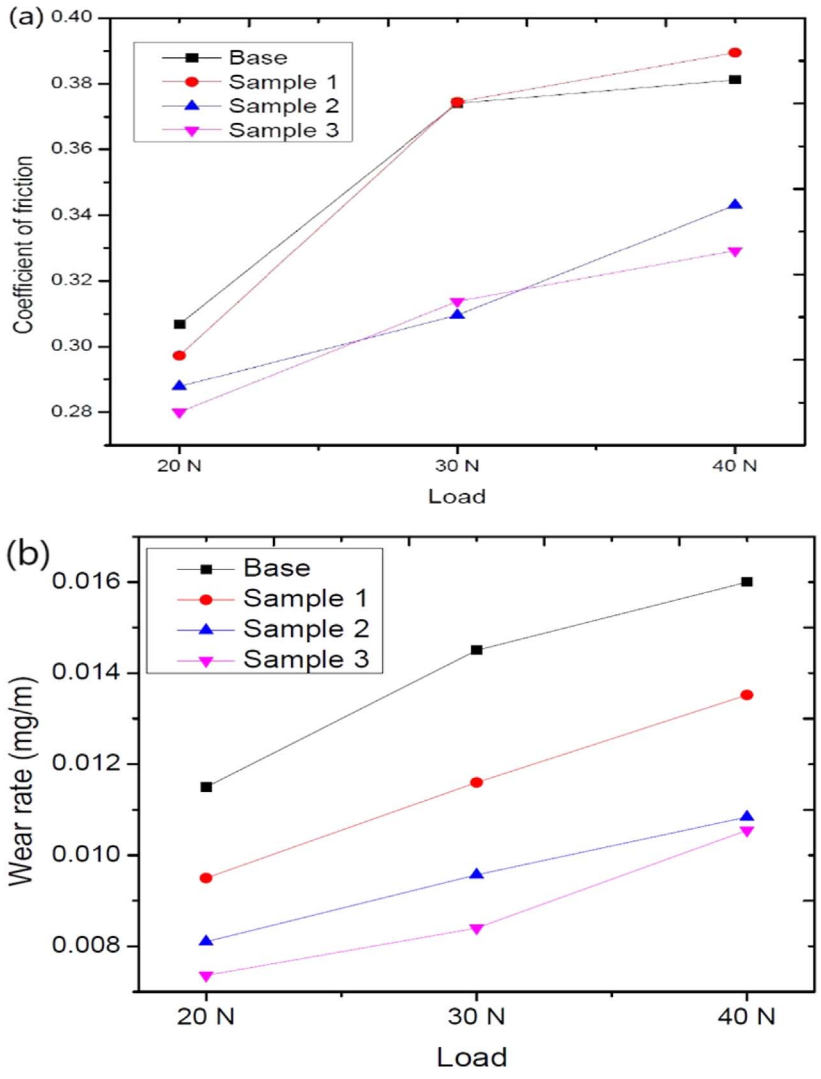


Figure 9. Effect of load on (a) coefficient of friction (b) wear rate.

3.5. Effect of load

Coefficient of friction variation with load can be observed in figure 9 (a), at lower load it is more marked. It can be observed that CoF increases with increasing load, though the increase in non-reinforced samples is more

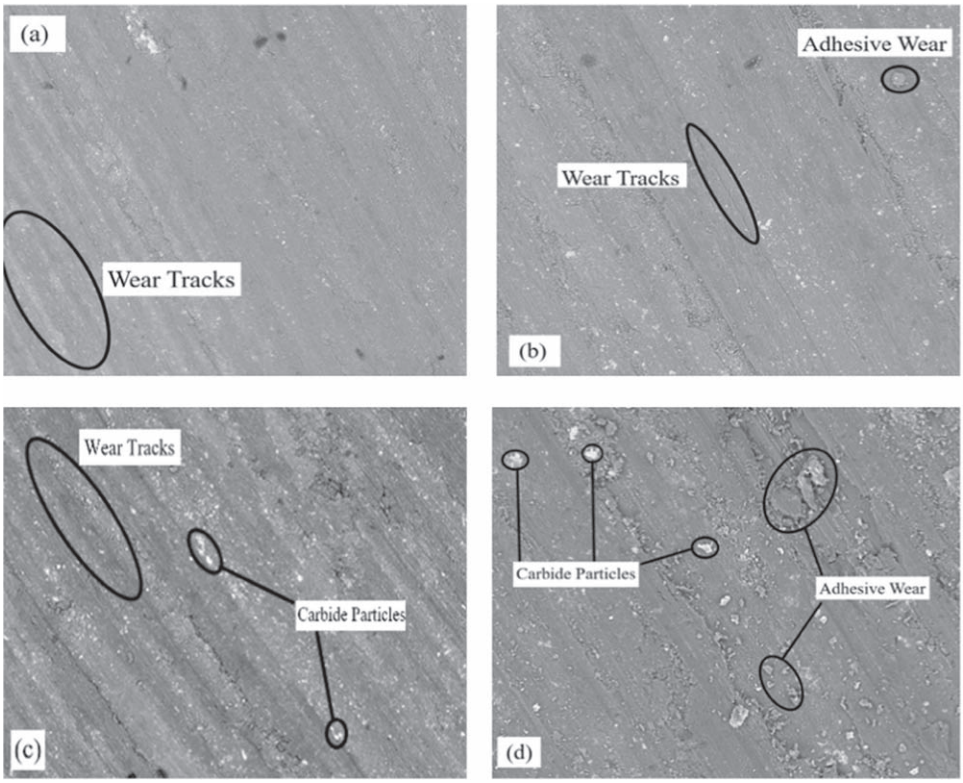


Figure 10. SEM image of worn surface at 40 N of (a) Base metal (b) Sample 1 (c) Sample 2 (d) Sample 3.

compared to the reinforced samples. Coefficient of friction is reduced in reinforced samples because of the formation of the mechanical matrix layer which performs as a lubricant. Similar type of results was reported by Tyagi *et al* in their study [32]. Variation of wear rate with load can be observed in figure 7 (b). It can be seen that wear rate increases with increasing load and it is more pronounced at larger loads. As at higher loads, more pressure is between the mating surfaces, so more wear rate is observed; at low loads, less pressure is there, less wear loss is observed [39]. Wear rate decreased on processing the samples, it decreased with the introduction of reinforcement in the matrix and it further decreased on fabricating the reinforced samples at higher tool rpm.

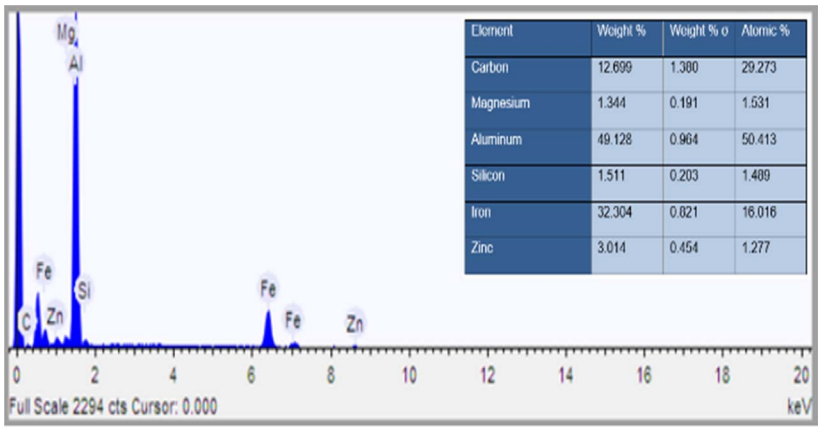
3.6. Wear morphology

SEM image of the worn surface of all tested samples at 40 N load is illustrated in figure 10. Wear tracks can be attributed to the sliding of the sample against the disc, and white particles in the SEM images illustrate the presence of the carbide particles. The SEM images indicate that adhesive wear is the primary wear mechanism. In adhesive wear, a part of the surface ploughs and sticks to a different section of the surface. The part sticks to the surface because of the large heat generated due to friction. Presence of different elements i.e. Fe, C, Si, Zn, Mg and Al can be seen in the sample, figure 11. The EDS indicates the high content of Fe-this might be because of the element transfer that takes place between the pin and the disc, during early phase of wear test. This also indicates the Fe layer formation, as it acts as a solid lubricant so it might be responsible for the decreased value of wear and coefficient of friction. Alam *et al* [40] reported similar kind of result in their study.

4. Conclusions

AA7075 matrix composites reinforced with silicon carbide having a particle size of 40 μm is successfully developed using friction stir processing (FSP). Effect of reinforcement, processing and tool rotational speed are analyzed on wear performance of the material. Major conclusions of the study are –

- Weight loss decreases on processing the alloy due to refinement of the grains.
- On addition of reinforcement, weight loss decreases because load from the alloy matrix is transferred to the hard-distributed silicon carbide particles.



Element	Weight %	Weight % σ	Atomic %
Carbon	12.699	1.380	29.273
Magnesium	1.344	0.191	1.531
Aluminum	49.128	0.964	50.413
Silicon	1.511	0.203	1.489
Iron	32.304	0.821	16.016
Zinc	3.014	0.454	1.277

Figure 11. EDS analysis of FSPed worn surface of processed sample.

- At high tool rotational speed, fabricated defects-free surface composite and uniform distribution of reinforcement is there, which contribute to increased wear resistance.
- Presence of C, Fe, Mg, Zn, Al, Si is confirmed in the worn out sample tested by EDS and it indicates the formation of Fe layer.
- SEM images of the worn out samples illustrates that adhesive wear is the prominent wear mechanism.
- Microhardness of the composite fabricated at 1000 tool rpm was 49% more than that of the base alloy.
- Elongation % of the FSPed without reinforcement particles increased on processing due to grain refinement as compare to the FSPed with reinforcement; it decreases the FSPed sample with the addition of reinforcement due to hardness increases.
- Microstructure confirms that on processing, the grains refined and grains were even finer on processing the material at higher tool rpm, which justified the increased hardness of the material produced at higher tool rpm.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

ORCID iDs

Ravi Butola https://orcid.org/0000-0002-6581-4131
Lakshay Tyagi https://orcid.org/0000-0001-5739-9807
Ranganath M Singari https://orcid.org/0000-0003-4981-3857
Qasim Murtaza https://orcid.org/0000-0001-8784-2304
Harish Kumar https://orcid.org/0000-0002-7555-795X
Dhiraj Nayak https://orcid.org/0000-0002-5233-7647

References

[1] Ashok Kumar M, Honnaiaha C and Ajit Prasad S 2018 Influence of extrusion process on mechanical and tribological properties of aluminium A356-Al_2O_3 stir cast MMC *Materials Today: Proceedings* **5** 26918–24
[2] Butola R, Tyagi L, Kem L, Ranganath M S and Murtaza Q 2020 Mechanical and wear properties of aluminium alloy composites: a review *Manufacturing Engineering. Lecture Notes on Multidisciplinary Industrial Engineering.* ed V Sharma, U Dixit, K Sørby, A Bhardwaj and R Trehan (Singapore: Springer) (https://doi.org/10.1007/978-981-15-4619-8_28)
[3] Tirth V, Elkotb M, Badruddin I, Mahmoud E, Islam S and Kamel B 2020 Effect of pressure on ageing response of (SiC + Al_2O_3)/6063 composites *Journal Of Materials Research And Technology* **9** 11834–48
[4] Tirth V and Arabi A 2020 Effect of liquid forging pressure on solubility and freezing coefficients of cast aluminum 2124, 2218 and 6063 alloys *Arch. Metall. Mater.* **65** 357–66
[5] Mahmoud E, Algahtani A and Tirth V 2020 Study on microstructure characterisation of three different surface coating techniques on 6082-t6 aluminum alloy *Met. Mater. Int.*

[6] Tirth V, Ray S and Kapoor M 2009 Effect of squeeze pressure on aging and mechanical properties of AA2218-5 Wt Pct Al_2O_3 (TiO_2) composites *Metallurgical And Materials Transactions A* **40** 1246–54
[7] Mahmoud E, Tirth V, Algahtani A and Khan S 2020 Microstructural characterization of different metal matrix composite claddings reinforced by TiC through YAG laser cladding *Mater. Res. Express* **7** 066407
[8] Coyal A, Yuvaraj N, Butola R and Tyagi L 2020 An experimental analysis of tensile, hardness and wear properties of aluminium metal matrix composite through stir casting process *SN Applied Sciences* **2** 892
[9] Gurusamy P, Prabu S and Paskaramoorthy R 2015 Influence of processing temperatures on mechanical properties and microstructure of squeeze cast aluminum alloy composites *Mater Manuf Proces* **30** 367–73
[10] Zawawi N, Azmi W, Redhwan A and Sharif M 2017 Coefficient of friction and wear rate effects of different composite nanolubricant concentrations on Aluminium 2024 plate *IOP Conf. Ser.: Mater. Sci. Eng.* **257** 012065
[11] Naveed M and Khan A 2016 Dry sliding wear of heat treated hybrid metal matrix composites *IOP Conf. Ser.: Mater. Sci. Eng.* **149** 012084
[12] Caliman R 2016 Aspects regarding wearing behaviour in case of aluminium composite materials reinforced with carbon fibers *IOP Conf. Ser.: Mater. Sci. Eng.* **145** 072013
[13] Raghavendra C, Basavarajappa S and Sogalad I 2018 Study on influence of Surface roughness of Ni-Al_2O_3 nano composite coating and evaluation of wear characteristics *IOP Conf. Ser.: Mater. Sci. Eng.* **310** 012112
[14] Latha Shankar B, Anil K and Patil R 2016 A study on 3-body abrasive wear behaviour of aluminium 8011/graphite metal matrix composite *IOP Conf. Ser.: Mater. Sci. Eng.* **149** 012099
[15] Razzaq A, Majid D, Manan N, Ishak M and Basheer U 2018 Effect of fly ash content and applied load on wear behaviour of aa6063 aluminium alloy *IOP Conf. Ser.: Mater. Sci. Eng.* **429** 012038
[16] Nagaral M, Auradi V, Parashivamurthy K, Shivananda B and Kori S 2018 Dry sliding wear behaviour of aluminium 6061-sic-graphite particulates reinforced hybrid composites *IOP Conf. Ser.: Mater. Sci. Eng.* **310** 012156
[17] Mittal P 2018 Wear behaviour of aluminium 7075 based composites reinforced with SiC, red mud and Al_2O_3 *International Journal For Research In Applied Science And Engineering Technology* **6** 2802–16
[18] Baradeswaran A and Elaya Perumal A 2013 Influence of B4C on the tribological and mechanical properties of Al 7075–B4C composites *Composites Part B: Engineering* **54** 146–52
[19] Rahman M H and Rashed H M 2014 Characterization of silicon carbide reinforced aluminum matrix composites *Procedia Engineering* **90** 103–9
[20] Pradhan S, Barman T K, Sahoo P and Sutradhar G 2016 Wear behavior of Al-SiC metal matrix composite under various corrosive environments *IOP Conf. Ser.: Mater. Sci. Eng.* **149** 012088
[21] Tirth V 2017 Dry sliding wear behavior of 2218 al-alloy-Al_2O_3(TiO_2) hybrid composites *J. Tribol.* **140** 9
[22] Bansal S and Saini J S 2015 Mechanical and wear properties of SiC/graphite reinforced AL359 alloy-based metal matrix composite *Def. Sci. J.* **65** 330
[23] Butola R, Murtaza Q and Singari R 2020 An experimental and simulation validation of residual stress measurement for manufacturing of friction stir processing tool *Indian Journal Of Engineering & Materials Sciences* **27** 826–36
[24] Butola R, Singari R and Murtaza Q 2020 Mechanical and wear behaviour of friction stir processed surface composite through self-assembled monolayer technique *Surface Topography: Metrology And Properties* **8** 045007
[25] Yuvaraj N, Aravindan S and Vipin 2016 Wear characteristics of Al5083 surface hybrid nano-composites by friction stir processing *Trans. Indian Inst. Met.* **70** 1111–29
[26] Dhayalan R, Kalaiselvan K and Sathiskumar R 2014 Characterization of AA6063/SiC-Gr surface composites produced by FSP technique *Procedia Engineering* **97** 625–31
[27] Dolatkhah A, Golbabaei P, Besharati Givi M and Molaiekiya F 2012 Investigating effects of process parameters on microstructural and mechanical properties of Al5052/SiC metal matrix composite fabricated via friction stir processing *Mater. Des.* **37** 458–64
[28] Saadatmand M and Aghazadeh Mohandesi J 2015 Optimization of mechanical and wear properties of functionally graded Al6061/SiC nanocomposites produced by friction stir processing (FSP) *Acta Metallurgica Sinica (English Letters)* **28** 584–90
[29] Butola R, Murtaza Q and Singari R 2020 Formation of self-assembled monolayer and characterization of AA7075-T6/B4C nano-ceramic surface composite using friction stir processing *Surface Topography: Metrology And Properties* **8** 025030
[30] Butola R, Singari R and Murtaza Q 2019 Fabrication and optimization of AA7075 matrix surface composites using Taguchi technique via friction stir processing (FSP) *Engineering Research Express* **1** 025015
[31] Bhushan R, Kumar S and Das S 2012 Fabrication and characterization of 7075 Al alloy reinforced with SiC particulates *The International Journal Of Advanced Manufacturing Technology* **65** 611–24
[32] Tyagi L, Butola R and Jha A 2020 Mechanical and tribological properties of AA7075-T6 metal matrix composite reinforced with ceramic particles and aloevera ash via Friction stir processing *Mater. Res. Express* **7** 066526
[33] Khojastehnezhad V, Pourasl H and Vatankhah Barenji R 2017 Effect of tool pin profile on the microstructure and mechanical properties of friction stir processed Al6061/Al_2O_3—TiB2 surface hybrid composite layer *Proc. Inst. Mech. Eng. Part L J. Mater. Des. Appl.* **233** 900–12
[34] Barmouz M, Asadi P, Besharati Givi M and Taherishargh M 2011 Investigation of mechanical properties of Cu/SiC composite fabricated by FSP: effect of SiC particles' size and volume fraction *Mater. Sci. Eng. A* **528** 1740–9
[35] Yuvaraj N, Aravindan S and Vipin 2015 Fabrication of Al5083/B4C surface composite by friction stir processing and its tribological characterization *Journal Of Materials Research And Technology* **4** 398–410
[36] Zhou D, Wang J, Lu Y and Bai Z 2020 Microstructure evolution and mechanical properties of friction stir processed TiC/7085Al nanocomposites *Mater. Res. Express* **7** 036515
[37] Yuvaraj N 2018 Improving the wear properties of aluminum 6082 alloy by surface compositing with zro2 ceramic particles via friction stir processing *Material Science Research India* **15** 68–74
[38] Raaft M, Mahmoud T, Zakaria H and Khalifa T 2011 Microstructural, mechanical and wear behavior of A390/graphite and A390/Al_2O_3 surface composites fabricated using FSP *Mater. Sci. Eng. A* **528** 5741–6
[39] Radhika N and Raghu R 2017 Investigation on mechanical properties and analysis of dry sliding wear behavior of Al LM13/AlN metal matrix composite based on taguchi's technique *J. Tribol.* **139** 10
[40] Alam M, Arif S and Ansari A 2018 Wear behaviour and morphology of stir cast aluminium/SiC nanocomposites *Mater. Res. Express* **5** 045008

2020 IEEE International Conference for Innovation in Technology (INOCON)
Bengaluru, India. Nov 6-8, 2020

MPPT Based Grid Connected Photovoltaic System Using Flyback Converter

MD. Modassir Masoom
Department of Electrical Engineering,
Delhi Technological University
New Delhi,India
Email- masoom_mt2k18@dtu.ac.in

DR. Narendra Kumar
Department of Electrical Engineering,
Delhi Technological University
New Delhi,India
Email- narendrakumar@dtu.ac.in

Ashish Rajeshwar Kulkarni
Department of Electrical Engineering,
Delhi Technological University,
New Delhi, India
Email- ashishkulkarni@dtu.ac.in

***Abstract*—This paper presents the collecting of solar energy by using the flyback converter method. This method is applied for the inverter to increase output voltage of Photovoltaic module. Since the conventional inverters have large size and low efficiency. So, as to overcome this drawback a flyback converter is designed, connected in parallel to increase the efficiency and also to reduce the switching devices. This method employs flyback converter as a boost converter which increases the output voltage of the PV source. The output DC power from the solar panel is converted to AC form with higher efficiency and low switching losses. Gate pulses to flyback switching devices are given in accordance with maximum power point tracking algorithm. Thereby, low power from the photovoltaic system is increased that also reduces the size of converter. This method is more efficient than the conventional method. The proposed system is analyzed and studied in MATLAB/Simulink.**

Keywords—Solar energy, Maximum power point tracking (MPPT), Flyback converter, Grid

I. INTRODUCTION

Now a days it has been seen that number researchers doing research on renewable energy sources like solar, wind etc. Since there is growing concern regarding the demand of renewable sources so the researchers are putting their maximum potentials to achieve more and more benefits from the most non-conventional renewable resources (i.e. solar energy).

Now it can be seen that everywhere in the country as well as in the world people are focusing on renewable energy resources and the most abundant renewable energy resource is solar energy as of data from the central electricity authority the total energy generated by solar in India on May 2020 was ~5606.88MW and on May 2019 it was ~4342.59MW so the total increment in the energy is ~1246.29MW after one year [1].

Since population of the country is increasing day by day so the demand of the energy as well. The population of country in 2014 was 1.2 billion and the required electricity demand was 776TWh in 2012 and it is expected in 2030 the population of India would be 1.5 billion and the electricity demand would be 2499TWh [2]. So, keeping this in mind research are being done to increase the power demand. So, to increase the power demand number of factors has to be keep in mind like climatic condition etc.

As from the data of 2005-2015 the solar power installed capacity has increased only 3.7MW to 4076MW respectively [2]. So, the Government of India had made an initiative to implement the solar mission to increase the solar power demand and it has also developed a government schemes like 25 solar parks, ultra-mega solar power project, canal top solar projects and 100 thousand solar pumps for farmers at different stages. The striving solar expansion program seeks to increase the solar capacity to 100GW by 2022 [2].

So, to meet the above demand researchers are doing research to increase the power demand from the available sources. In this paper maximum power tracker is used (MPPT) to track the maximum power from the PV array and a flyback converter to boost the input supply. Since flyback converter is used instead of normal boost converter because it contains a transformer which isolates the primary and secondary of the transformer and it can operate over wide range of input voltage variations and can offer single or multiple isolated output voltage and any ripples in the secondary would not transfer to the primary due to isolation since it is energy efficient and low cost in comparison to other switched mode power supply(SMPS).

The main attainments of this paper are isolation between photovoltaic module and grid, high efficiency and less loss.

II. PROPOSED TOPOLOGY

This paper consists of three flyback converter connected in parallel to boost the output voltage of the PV module. The output from the PV module is fed to the flyback through switch which is connected in series with the transformer as shown in Fig. (1)[3]. The magnetic circuit design of transformer is done like an inductor which stores the energy. The output voltage from the secondary of the flyback transformer consists of voltage rectifier and a filter circuit to get the desired output. It is different from the normal transformer because in normal transformer under load, primary and secondary windings conducts simultaneously due to which ampere turns of primary and secondary windings are nearly balanced but this is not in the case of flyback transformer but they are like two magnetically coupled inductor[4], [5].

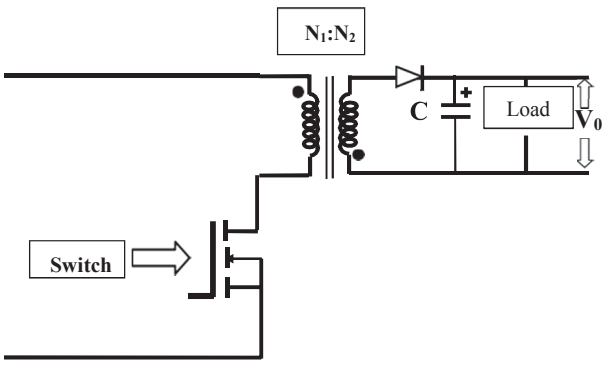


Fig. 1. Flyback Converter

978-1-7281-9744-9/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:02:19 UTC from IEEE Xplore. Restrictions apply.

The block diagram of the above proposed system is as shown below:

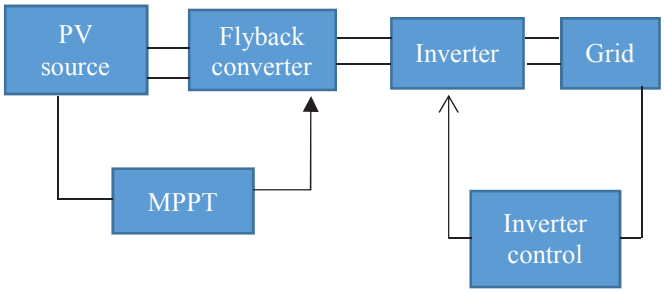


Fig. 2. Block diagram of proposed system

A. Photovoltaic module specification

It consists of 4 series connected modules per string and 10 parallel strings of SunPower T5-SPR-305(305 watt) module the characteristics curve at different irradiance and temperature is as shown below in Fig. (3).

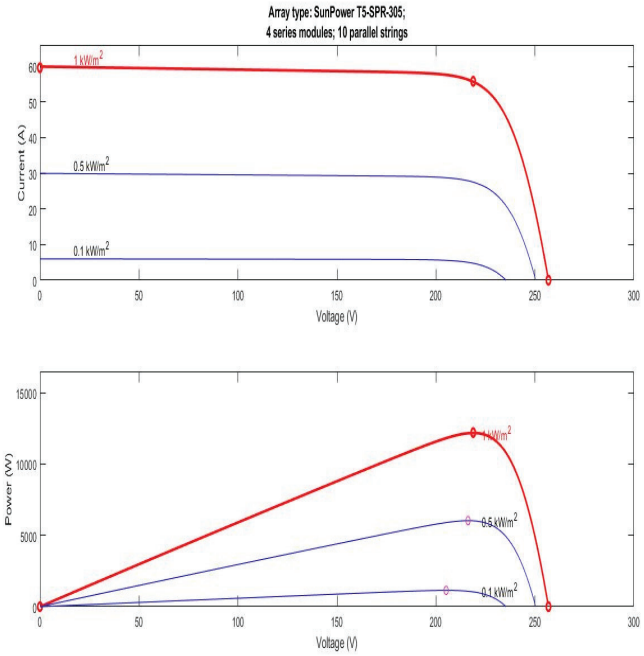


Fig. 3. Characteristic curve of sunpower-T5-SPR-305 at different irradiance and temperature.

An equivalent circuit diagram of single solar cell is shown below in Fig. (4) [6].

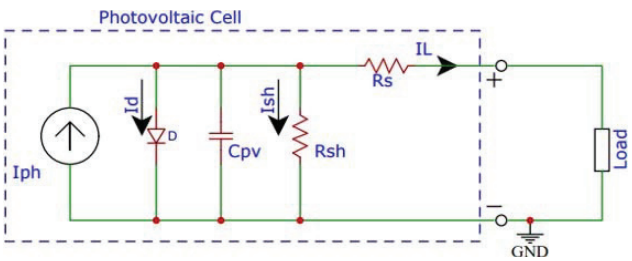


Fig. 4. Equivalent circuit diagram of solar cell

The output current that is load current can be calculated from the above equivalent circuit as shown in the Fig. 4.

$$I_L = I_{ph} - I_D - I_{sh} \quad (1)$$

Where, I_L is Load current in ampere(A), I_{ph} is Photocurrent, I_D is Forward diode current and I_{sh} is Shunt current.

B. Forward diode current determination

Forward diode current can be determined by Schottky diode method as shown below,

$$I_D = I_0 [exp\ (qv_0/kT) - 1] \quad (2)$$

Where, I_0 is Reverse saturation current of diode, q is charge of an electron 1.602 x 10^{-19} C, V_0 is Forward voltage of diode, k is Boltzmann constant 1.381 x 10^{-13} and T is Temperature of the PV cell in ^{0}C.

From the Fig. 4.

$$V_0 = V + I_L R_S \quad (3)$$

$$I_{sh} = (V + I_L R_s)/R_{sh} \quad (4)$$

Now, from the equation (3) and (4) putting in the equation (1) it will give final equation of the load current,

$$I_L = I_{ph} - I_0 \{exp\ [q(V + I_L R_S)/kT] - 1\} - ((V + I_L R_s)/R_{sh}) \quad (5)$$

C. Maximum power point tracker of proposed system

Perturbation and observation method are one of the modest online approaches which, has been considered by a number of researchers. P&O can be executed by applying perturbations to the reference voltage or the reference current signal of the solar panel. A curve showing tracking of MPP by P&O , which is also known as the 'hill climbing method', is depicted in Fig. (5) [7].

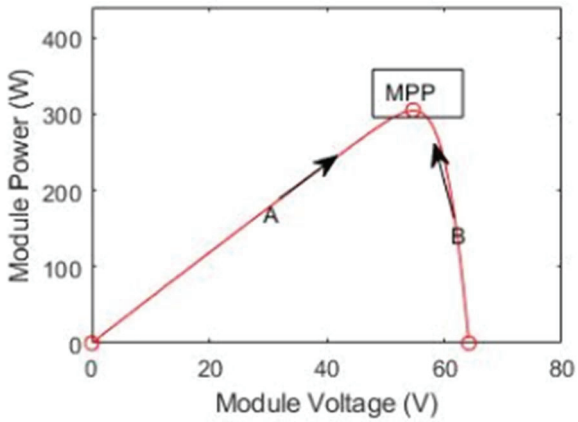


Fig. 5. Curve showing tracking of MPP by P&O

Since due to day night condition and the weather conditions or when the seasons changes the irradiance from the sun and the temperature changes accordingly due to which there is variations in output of the PV panel. So, to

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:02:19 UTC from IEEE Xplore. Restrictions apply.

track maximum power from the PV panel P&O Maximum power point tracker is used. [8].

The output of the MPPT is given as the switching signal to the flyback converter and when the switch is on the magnetizing current starts flowing through the primary of the transformer and the formula is given as below [9],

$$I_m = \frac{V_{ab}}{L_m} \tag{6}$$

Where,

I_m = Magnetizing current of flyback transformer.

V_{ab} = output voltage from PV module.

L_m = Magnetizing inductance of flyback transformer.

And the output power of PV module is analyzed as,

$$P_{pv} = V_{ab}I_{pv} = \frac{n_{cell}V_{ab}^2D_{peak}^2}{4L_mf_s} \tag{7}$$

Where,

n_{cell} = number of inserted cells.

f_s = Switching frequency.

D_{peak} = Peak value of duty ratio.

III. CONTROL STRATEGY

In this paper a voltage controller is designed since at the output of the inverter some noise and ripples were present so, to decrease the error two controllers were designed

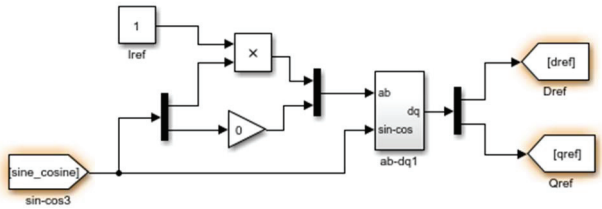


Fig. 6. Simulink model of first controller

In the first controller Fig. (6) reference current has been taken to produce the Direct axes reference (d_{ref}) and Quadrature axes reference (q_{ref}) so that it can be used in the second controller for error detection.

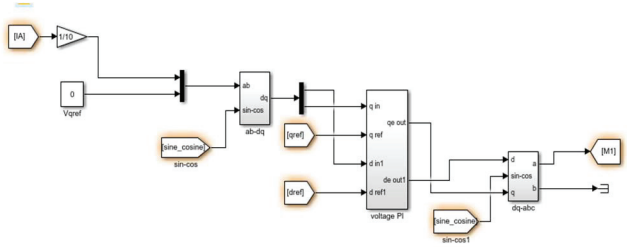


Fig. 7. Simulink model of second controller

In the second controller the output voltage and reference voltage from the first controller is passed through the proportional plus integral controller (PI) to reduce the error and the final error signal is given to Pulse width modulation (PWM) generator from which four pulses are created and these signals is given to the switch of the inverter.

IV. MATLAB SIMULINK MODEL

This is the complete Simulink model of the system fig. 8. in which output of the PV module is fed to the flyback converter through decoupling capacitor.

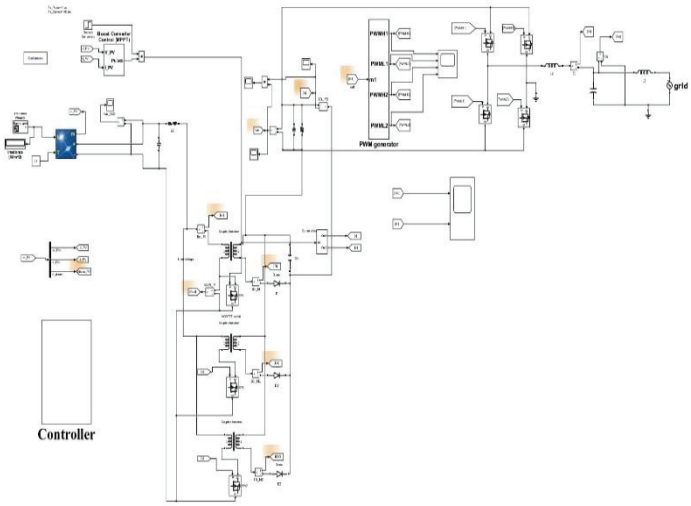


Fig. 8. Simulink model of the system

There are three switches SW1, SW2, SW3 to the three interleaved flyback converter connected in parallel. Thus, the duty ratio of MPPT controls the pulse generator and the pulse generated by it is given to the switch SW1 of the flyback. The Voltage generated by the three flyback converter is boosted and pulsating DC So, to get AC it is fed to the inverter circuit with four MOSFET switch M1, M2, M3 and M4 and these four switches gets the signal from PWM generator which is controlled by the controller error signal.

When all the four switches conduct at same instants the bridge inverter inverts the pulsating DC to sinusoidal voltage output. Since, output from the inverter had some harmonics so to smooth the output voltage it is passed through the low pass filter and then it is finally given to the grid. The filter and grid specification are given in the below table 1.

TABLE 1

Serial no.	Element	Rating
1	Filter inductor	20mH
2	Filter capacitor	30μF
3	Grid	230√2 V

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:02:19 UTC from IEEE Xplore. Restrictions apply.

V. SIMULATION RESULT

In the given fig. 9 below it shows the fast Fourier transformation analysis of the output voltage in which the total harmonic distortion of the curve is 1.60% for 3 cycles and the fundamental frequency is set to 50Hz.

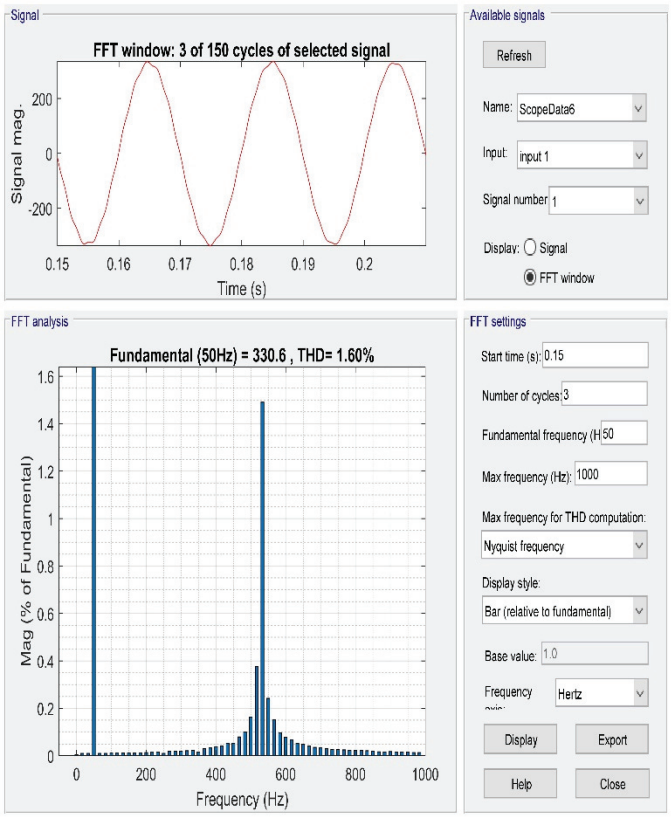


Fig. 9. FFT analysis of output Voltage

In the below fig. 10 it is the output voltage across the grid and from time 0 to 0.075 it is in the transient state due to which it is showing some distortions in the output and after the transient state output gets settled whose peak value is 332 V.

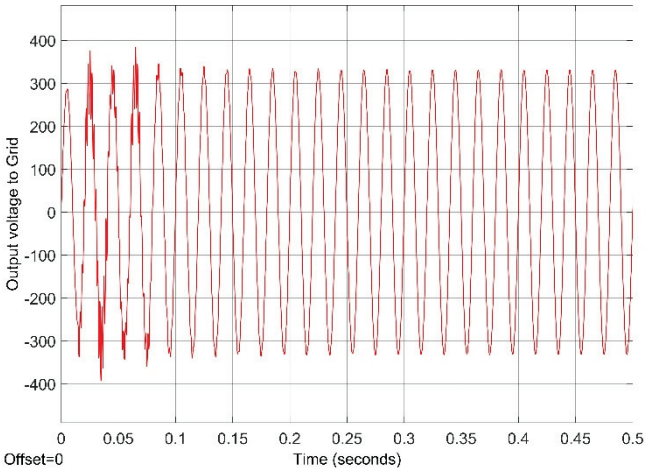


Fig. 10. Output voltage across the grid

In the below fig. 11. it is fast Fourier transformation analysis of the output current in which the total harmonic distortion of the curve is 1.26% for 3 cycles and the fundamental frequency for this also is set to 50Hz.

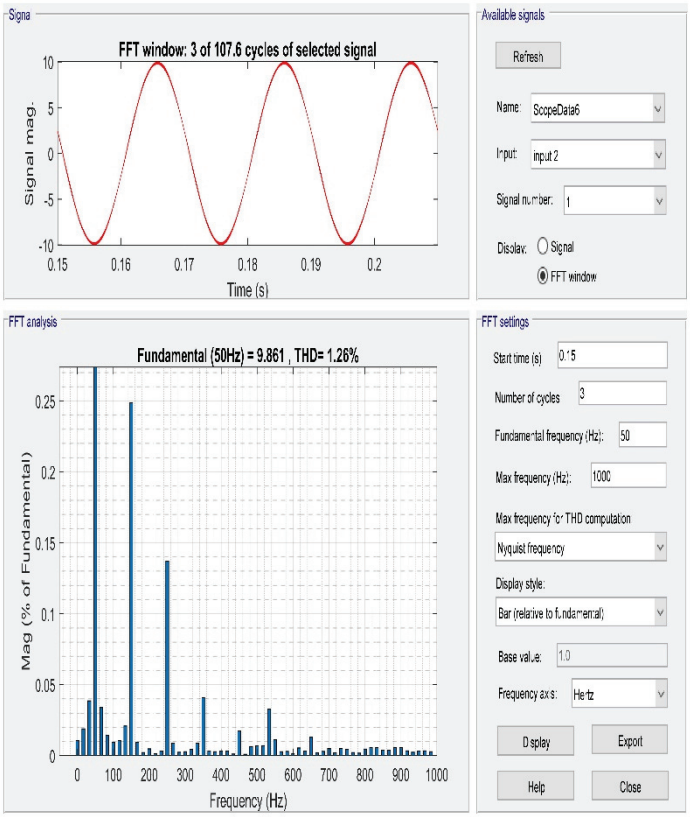


Fig. 11. FFT analysis of output Current

In the below fig. 12 it is the output current through the grid here also it can be seen that from time 0 to 0.075 there is some distortion likewise in the voltage output this is due to the transient state and after which it gets settled and the peak value of the output is 10 A.

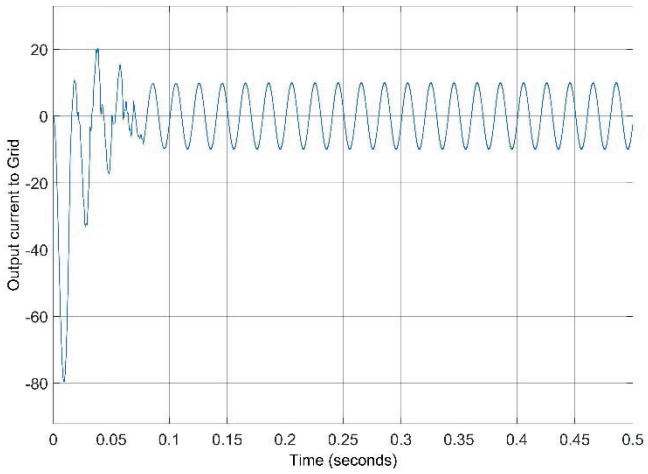


Fig. 12. Output current through the grid

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:02:19 UTC from IEEE Xplore. Restrictions apply.

VI. CONCLUSION

This paper explains the new method of inverter control by incorporating three interleaved flyback converter and MPPT to track the maximum power from the PV panel. The MPPT tracking algorithm operate under all climatic condition. Since by using flyback converter the output voltage increases effectively and the number of switching devices decreases due to which the switching loss is very less than the conventional method and also it acts as an isolation between the PV module and the grid. The THD value of the output voltage waveform is 1.60% and for the current waveform it is 1.26% which is less than 5% within standard limits. Hence it can be concluded that the use of flyback converter is more efficient as compared to conventional method employing converter.

REFERENCES

[1] C. E. Authority, "Government of India Ministry of Power," 2019.

[2] M. of power Government of India,Central Eectricity Authority, "All India Installed Capacity (in Mw) of Power Stations Installed Capacity (in Mw) of Power Utilities in the States / Uts Located in," *Cent. Eectricity Authority, Minist. power*, vol. 4, pp. 1–7, 2020, doi: http://www.cea.nic.in/reports/monthly/installedcapacity 2016 / installed_capacity-03.pdf.

[3] D. C. Converters, "Lesson Fly-Back Type Switched Mode Power Supply," pp. 1–15.

[4] S. P. Patil, R. R. Patil, A. M. Mulla, and S. D. Patil, "Performance analysis of interleaved flyback inverter topology for single phase high power application using MATLAB Simulink," *Proc. IEEE Int. Conf. Circuit, Power Comput. Technol. ICCPCT 2017*, 2017, doi: 10.1109/ICCPCT.2017.8074192.

[5] L. M. K. Johny and K. M. Shafeeque, "PV fed flyback DC-AC inverter with MPPT control," *2014 Annu. Int. Conf. Emerg. Res. Areas Magn. Mach. Drives, AICERA/iCMMD 2014 - Proc.*, 2014, doi: 10.1109/AICERA.2014.6908213.

[6] O. Singh and S. K. Rajput, "Mathematical modelling and simulation of solar photovoltaic array system," *Int. Conf. Res. Adv. Integr. Navig. Syst. RAINS 2016*, 2016, doi: 10.1109/RAINS.2016.7764395.

[7] A. Sharma, "Modelling and Comparison of Different MPPT Algorithms for a Standalone PV System," no. Icesc, pp. 989–995, 2020.

[8] A. Rai, B. Awasthi, A. Dixit, and C. K. Dwivedi, "Modeling of solar photovoltaic module and study parameter variation effect using MATLAB/Simulink," *ICCCCM 2016 - 2nd IEEE Int. Conf. Control Comput. Commun. Mater.*, vol. 250, no. Iccccm, 2017, doi: 10.1109/ICCCCM.2016.7918239.

[9] W. Sinvert and M.- Drives, "Connected Photovoltaic Systems," no. 1, 2000.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:02:19 UTC from IEEE Xplore. Restrictions apply.

AIP Conference Proceedings

Numerical simulation of n-TiO2/p-CIGS solar cell

Cite as: AIP Conference Proceedings **2265**, 030675 (2020); https://doi.org/10.1063/5.0017203
Published Online: 05 November 2020

Gagandeep, Mukhtiyar Singh, Ramesh Kumar, and Sonia Kumari

View Online Export Citation

Meet the Next Generation of Quantum Analyzers
And Join the Launch Event on November 17th

Register now
Zurich Instruments

AIP Publishing

© 2020 Author(s).

Numerical Simulation of n-TiO2/p-CIGS Solar Cell

Gagandeep[1, a)], Mukhtiyar Singh[2], Ramesh Kumar[1, b)] and Sonia Kumari[1]

[1]*Department of Physics, Guru Jambheshwar University of Science & Technology, Hisar, India*
[2]*Department of Applied Physics, Delhi Technological University, Delhi, India*

[a)]gagankundu22@gmail.com
Corresponding author: [b)] rameshkumar.bibiyan@gmail.com

Abstract. The mostly used buffer layer in CIGS solar cell is CdS. This buffer layer is toxic. Therefore, the p-type Cu (In, Ga)Se_2 (CIGS) absorber layer is simulated with non toxic and wide band gap TiO_2 material as n-type buffer layer in AFORS-HET software under the illumination condition of AM 1.5 G. To see the effect of absorber layer thickness on the performance parameters of the cell, it is varied from 200-1900 nm. The highest efficiency of the cell is obtained 16.03% for optimized thickness of absorber layer i.e., at 1900 nm. The temperature influences the performance of the cell and its performance is maximum at temperature 300 K. The external quantum efficiency behavior with wavelength is observed. The reported results in this study show that TiO_2 material is better replacement of CdS material.

Key Words: Solar Cell, Efficiency, TiO2, AFORS-HET, CIGS

INTRODUCTION

The chalcopyrite Cu(In, Ga)Se_2 (CIGS) material based solar cells is very promising thin film solar cell and play very important role in the enhancement of photovoltaic technology [1-3]. The CIGS solar cell has excellent outdoor stability. The CIGS based solar cells are attracting considerable space application interest because of its stability against particle irradiation. Its stability is proved by electron and proton irradiation test of CdTe and CIGS solar cells which is superior to Ⅲ-Ⅴ or silicon solar cells. The laboratory scale power conversion efficiency (PCE) of this cell is attaining 20.4% on a flexible polymer substrate [4] and 22.6% on a soda lime glass (SLG) substrate [5]. The CdS plays the role of buffer layer in CIGS solar cell and resulting the highest above mentioned device efficiency. The CdS layer on CIGS is formed by chemical bath deposition (CBD) method. The CdS layer have relatively low band gap energy (2.4-2.5 eV). Hence, the finest performance of the cell is limited by the parasitic absorption in short wavelength region [6]. The CdS buffer layer is also toxic in nature.

Therefore, different buffer layer with a wider band gap is proposed [7- 9]. W. Hsu et. al [10] experimentally studied that amorphous TiO_2 act as buffer layer on CIGS cell and improved the performance of the device

Keeping the observation in mind, we simulated the TiO_2 as buffer layer in CIGS solar cell and obtained the PCE is 16.03%. Our theoretical results reported herein are improved the performance of the device.

SIMULATION DETAILS

Here, we simulate a polycrystalline CIGS based solar cell by n-TiO_2 /p-CIGS heterojunction structure with ITO and Mo are front and back contact respectively. In this paper, we use AFORS-HET (Automat for simulation of heterostructure) [11-13] to analyze the working of CIGS based thin film solar cell. The simulated structure is shown in figure 1.

DAE Solid State Physics Symposium 2019
AIP Conf. Proc. 2265, 030675-1–030675-4; https://doi.org/10.1063/5.0017203
Published by AIP Publishing. 978-0-7354-2025-0/$30.00

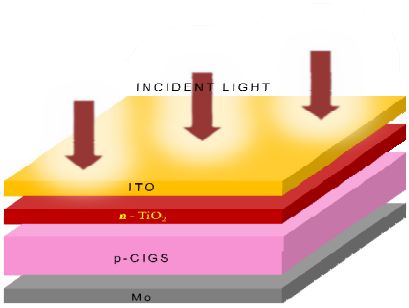


FIGURE 1. Schematic structure of simulated CIGS solar cell

The thickness of the CIGS layer is varied from 200-1900 nm. The variations in the values of the performance parameters of the cell such as short current density (J_{sc}), open circuit voltage (V_{oc}), fill factor (FF) and PCE (η) are noticed with absorber layer thickness. The impact of temperature on the performance parameters of the device is observed. The J-V (current density-voltage) characteristics are investigated for this structure with optimized absorber layer thickness. The quantum efficiency spectra with wavelength are also plotted. The layer parameters used in this work are taken from different theoretical and experimental studies [14-15] and mentioned in Table 1.

TABLE1. Layer parameters used in simulation

Parameter	**n-TiO_2**	**p-CIGS**
Layer thickness (nm)	90	200-1900
Dielectric constant	100	10
Electron affinity (eV)	4	4.2
Band gap (eV)	3.2	1.4
Optical band gap (eV)	3.2	1.4
Valence band density (cm^{-3})	1×10^{21}	2.2×10^{18}
Conduction band density (cm^{-3})	2×10^{20}	1.8×10^{19}
Hole mobility ($cm^2\ V^{-1}\ s^{-1}$)	0.006	100
Electron mobility ($cm^2\ V^{-1}\ s^{-1}$)	0.006	25
Thermal velocity of electron(cm/s)	1×10^{6}	1×10^{7}
Thermal velocity of hole(cm/s)	4×10^{6}	1×10^{7}
Doping concentration(cm^{-3})	5×10^{19}	2×10^{13}
Layer density(g cm^{-3})	2.328	2.328

RESULTS AND DISCUSSIONS

The thickness of CIGS layer is varied from 200-1900 nm and performance of the cell is observed. The cell parameters variation with absorber layer thickness is displayed in fig.2. The value of cell parameters such as J_{sc}, V_{oc}, FF and η are changed from 23.06 mA/cm^2 to 28.01 mA/cm^2, 606.4 mV to 697.7 mV, 81.95% to 82% and 11.46% to 16.03% respectively. It is explained based on the fact that larger thickness of CIGS layer absorbs more radiation and produce more carrier i.e. electron-hole pair and probably increase the performance parameters i.e. V_{oc}, J_{sc}, and η of the device. The FF of the cell is increased up to 82.24% at 1200 nm absorber thickness but decreases up to 82% at 1900 nm absorber thickness. The optimum thickness of CIGS layer is around 1900 nm from which the highest

efficiency of 16.03% is obtained for the n-TiO_2/p-CIGS cell. The results reported herein are accorded with the general trends of the device i.e. increased layer thickness improves the efficiency of the cell [16].

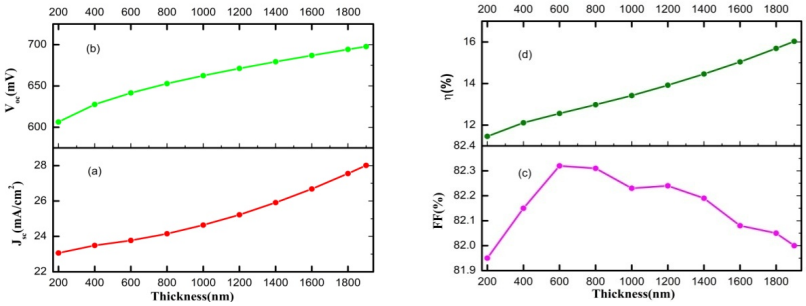


FIGURE2. Variation of solar cell parameters (a) J_{sc}, (b) V_{oc}, (c) FF and (d) Efficiency with absorber layer thickness

The variation of efficiency, FF, J_{sc} and V_{oc} with operating temperature is shown in figure 3. The performance of the cell, are also influenced by the operating temperature. The cell parameters variation is discussed in the temperature range 300-600 K under a constant illumination of 1000 W/m^2. The characteristics parameters of the cell namely V_{oc}, FF and η are strongly influenced by the operating temperature and get reduced with increase in temperature of the cell but J_{sc} remains constant. From the observed results we analyzed that the η of the cell is decreased from 16.03% to 0.15% in the range 300-600 K. It may be probably happened due to that, as we increase the temperature of the cell then there is increased in the reverse saturation current density J_o which is caused by the increase in the intrinsic carrier concentration and decrease in the band gap of the material. The conductivity of the cell is also affected by the temperature. At optimized value of absorber layer thickness J-V characteristics of the cell is also studied and obtain the η of 16.03% with FF of 82%. We also calculate the V_{oc} = 697.7 mV and J_{sc} = 28.01 mA/cm^2 respectively and results are shown in fig. 4(a).

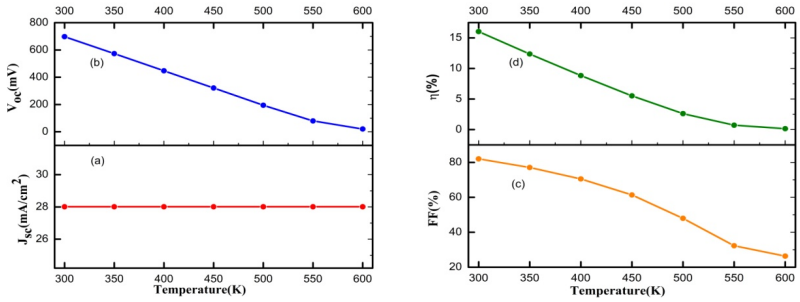


FIGURE3. Variation of solar cell parameters (a) J_{sc}, (b) V_{oc}, (c) FF and (d) Efficiency with operating temperature

The radiation absorbed by the cell is given by the spectra of EQE (external quantum efficiency). The QE spectra provides an information about the absorption of light by the CIGS device as compared to incident one. The EQE provides the detailed study about all photons whether they are absorbed, reflected or transmitted and provide us a maximum value in the wavelength range of 600-900 nm. All the absorbed photon by the cell are counted by the IQE (internal quantum efficiency) and gives maximum in 400-900 nm of wavelength but IQE1 becomes almost unity in this range of wavelength and counts only that photons which are absorbed by the electrical layers. The results of quantum efficiency are shown in fig 4 (b)

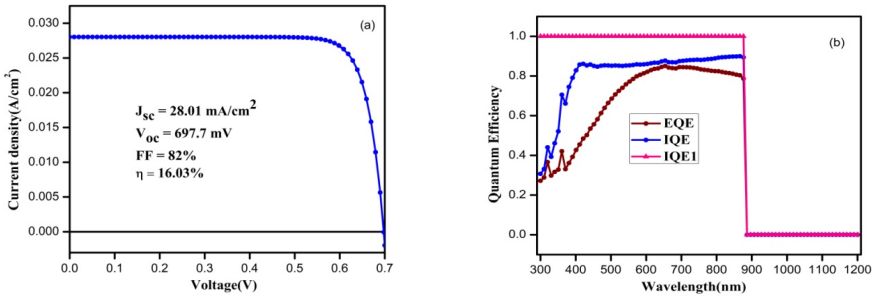


FIGURE4. (a) J-V characteristics (b) Quantum efficiency with wavelength

CONCLUSIONS

Theoretically, we simulated the TiO_2 layer as buffer layer in CIGS solar cell. The thickness of the absorber layer influences the performance of the device. The maximum value of power conversion efficiency is 16.03% at 1900 nm absorber thickness. The cell performance is influenced by the temperature and have attained maximum value at 300 K. The J-V curve and variation of external quantum efficiency of the solar cell with wavelength are also plotted. The obtained results in this study boost the performance of TiO_2 as buffer layer.

ACKNOWLEDGMENT

One of the authors (**Ramesh Kumar**) is thankful to the **University Grants Commission, INDIA** for giving the financial support through START-UP GRANT under FRPS (No. F.30-364/2017 (BSR), Dated 15/12/2017). The authors highly acknowledge the AFORS-HET software developer for providing free access of the software.

REFERENCES

1. J. Löckinger, S. Nishiwaki, T. P. Weiss, B. Bissig et al., Solar Energy Materials and Solar Cells, 174, 397 (2018).
2. C. Zhang, H. Zhu, X. Liang, D. Zhou, Y. Guo, X. Niu, Z. Li, J. Chen, Y. Mai, Vacuum, 141, 89 (2017).
3. Gagandeep, M. Singh and R. Kumar, AIP Conference Proceedings, 2093, 020018 (2019).
4. A. Chirila et.al., Nat. Mater. 12, 1107 (2013).
5. P. Jackson et.al., Phys. Status Solidi-R 10, 583(2016).
6. N. Naghavi et.al., Prog. Photovolt.18,411 (2010).
7. S. Spiering et.al.,Solar Energ.Mater. Sol.C 144, 544 (2016).
8. J. Lindahl et. al., IEEE J. Photovolt.3, 1100 (2013).
9. T. M.Friedlmeir et. al., IEEE J. Photovolt 5, 1487(2015).
10. W. Hsu et.al., Sci. Rep. 5, 16028 (2015).
11. R.Varache et. al., Sol. Energ. Mat. Sol. C.141, 14 (2015).
12. R. Stangl, J. Haschke and C. Leendertz,Solar Energy, Intech (R.D. Rugescu (Ed.), Croatia, 2009), pp. 432.
13. R.Stangl, M. Kriegel and M. Schmidt, Proceeding of the WCPEC-4, Fourth World Conference on Photovoltaic Energy Conversion (Hawaii, USA), PP. 1350 (2006).
14. M. Mostefaoui et. al., Energy Procedia, 74, 736 (2015).
15. S. Dabbabi, T. B. Nasr, N. K. Turki, Results in Physics, 7, 4020 (2017).
16. S. Tobbeche and H. Amar, JNTM, 04, 89 (2014).

iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** XI **Month of publication:** November 2020

DOI: https://doi.org/10.22214/ijraset.2020.32050

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue XI Nov 2020- Available at www.ijraset.com

Numerical Simulation of Single and Double Slope Solar Still for Different Performance Affecting Variables

Rohit Kumar[1], Rahul Kumar Meena[2]
[1, 2]*Department of Mechanical Engineering, Delhi Technological University, Delhi (India)*

Abstract: ***Two-phase, three dimensional, a single slope and double slope single basin still were simulated using ANSYS FLUENT v19.2. Simulations were made done for evaporation and condensation process at Delhi's climate conditions (27.0238° N, 74.2179° E). Temperature inside the single slope solar still is maximum between 12:00 to 13:00 hrs while the double slope still has a low temperature at the same time, compared to the single still. The maximum and minimum temperature of the water-vapor mixture inside the single slope still were calculated 54.9445°C and 19.4401 °C, whereas, in double slope, the maximum and minimum temperature of the water-vapor mixture inside the still was 52.21 °C and 12.6988°C. The maximum water production rate in single slope solar still is 0.60729 kg/m^{2}hr, while the maximum water production rate in the double slope is 0.54371 kg/m^{2}hr. Condensation depends on the temperature difference between the glass surface and outer environment, more condensation is noticed in the single slope solar still. Inner water temperature is responsible for more evaporation and higher temperature, i.e., more than 4°C approximately is observed in single slope solar still as compared to the double slope.***
Keywords: Solar Still; Computational Fluid Dynamics; Water temperature; Solar heat flux; Water Productivity

I. INTRODUCTION

Freshwater is a basic need for the survival of humans beings and other organisms. But, there is a limited supply of fresh water on the surface of the earth. Earth contains 71% water on it and only 2.5% of water is available as potable water, that too as ice and groundwater, leaving possibly 1.04% of freshwater available for human consumption [1]. Also, 7.6 billion of the world population depends upon this mere percentage of water to drink, cook, irrigate crops, and feed livestock. The issue is that the percentage of freshwater reduces day by day due to pollution and indiscriminate uses. Due to inadequate availability of water, the health issues are prevailing among humans and other living beings. Globally, at least two billion people use a drinking water source contaminated with feces. Contaminated water can transmit diseases such as diarrhea, cholera, dysentery, typhoid, and polio. Consumption of this contaminated water is estimated to cause 485000 diarrhoeal deaths each year [2].

There are so many technologies involved in the treatment of contaminated water using different methods by using membrane process (reverse osmosis and non-filtration) and thermal process (multi-stage flash distillation, multi-effect distillation, and electro-dialysis). But solar desalination is the best method for removing impurities from water to provide potable water. It is an eco-friendly, low-cost method and also provides drinkable water in remote areas. Solar distillation uses the evaporation and condensation process in which solar radiation incidents on the glass surface and then transmits and absorbed by an absorber surface, increasing the basin's temperature and converting water into vapor.

Many research works are being carried out to enhance the different parameters of a single and double solar still with experiments and simulation on different software. There is much software for modeling as MATLAB, ANSYS, COMSOL, FORTRAN, TRNSYS, CATIA, etc. The experimental work consumes more time for setup and experimentation, but simulation takes less time to give results. Setoodeh et al. [3] developed a three-dimensional, two-phase model for the evaporation and condensation process using a computational fluid dynamics simulation tool to simulate single slope solar still for calculating water temperature and production and also analyzed heat transfer coefficient. The simulation work was done using ANSYS 11. As a result, it was predicted that computational fluid dynamics is a powerful tool for design, parameter analysis, and removing difficulties in solar still construction. Singh et al. [4] studied experiment and simulation work to enhance productivity with varying depth of water and checked for its thermal efficiency. The modeling was done using ANSYS v14.0 for the climate condition of Jaipur. Solar still has more productivity at low water depth. Panchal et al. [5] worked on the experimental and computational fluid dynamics model of the single-slope solar still. The Geometry was created using ANSYS Workbench 10 and unstructured tetrahedron mesh was used to predict the results. The simulation results of distillate output, water temperature, and heat transfer coefficients are predicted using ANSYS CFX 10.

©IJRASET: All Rights are Reserved

Maheshwari et al. [6] researched a single basin double slope solar still using ANSYS CFX 14.0 to check the overall production. The overall production is maximum in March and November. Gokilavani et al. [7] did experimental and modeling work of single slope solar still. The geometry of the model was created using ANSYS Workbench 14.5 and then simulated in ANSYS CFX. The experimental work was compared with the simulated study and verified the water temperature distribution over the glass surface for 27/Nov/2013 and 28/Nov/2013.

The result shows the maximum temperature achieved inside the glass on 27/Nov/2013 is higher as compared to 28/Nov/2013. Mahendren et al. [8] did a simulation work for double slope solar still to enhance its efficiency for getting better productivity. The mathematical model and coding were done by using the MATLAB program to calculate the various heat fluxes in the still, determine the hourly output of the still, and find the still's efficiency. Complete numerical analysis was done and various characteristics graphs were plotted in MATLAB. Simulink toolbox was used to show simulation. ASCII text files, which are M-files, were used to calculate convective, evaporative, and radiative heat transfer rates. Nayak et al [9] did thermal modeling and experiment on double slope solar still.

Modified double slope solar still was designed by applying a transparent acrylic sheet and the north wall of the solar still was imposed by an opaque fiber-reinforced plastic. The geometry of the model was created with the help of CATIA V5R18 software. The simulation result and experimental result was almost closer to each other.

The experiment was conducted for Allahabad (25.23°N, 79.09° E) to check the productivity of solar still. The result shows the modified double slope solar still has 16 kg of purified water from 25 kg of brackish water, which was formed twice than conventional solar still.

Tripathi and Tiwari [10] developed a simulation model for thermal analysis of single-slope solar still by taking a parameter of the solar fraction. The geometry of the model was created in AUTOCAD 2000. The simulation work was done for weather conditions of New Delhi (27.0238° N, 74.2179° E). A MATLAB program was used to calculate convective and evaporative heat transfer coefficients and also estimated solar fraction. The following formula for the solar fraction (Fn) for a particular wall still was used in modeling.

$$F_n = \frac{solar\ radiation\ on\ the\ wall\ of\ the\ still\ for\ a\ given\ time}{solar\ radiation\ on\ the\ wall\ and\ the\ floor\ of\ the\ still\ for\ the\ same\ time}$$

Chaibi [11] presented a simulation model of double slope solar still integrated with a greenhouse roof to check the influence of solar irradiation, optical material properties, and process parameters on freshwater. A numerical simulation was performed for the climate condition of Tunisia.

For the calculation of solar irradiation, simulation code TRNSYS (Transient System Simulation) was used, which helps in hourly calculations of radiation values for an inclined glass surface. And for solving heat transfer equations, An engineering equation solver (EES) was used in the model.

The effect of solar radiation and visual material properties are considered in the system. Solar radiation proportionally affects the efficiency of the rooftop- incorporated framework. The thermal and optical performance could be improved by maximization of the solar energy yield. Ali Saeed et al. [12] performed a numerical investigation using Nano-Phase Change Material (NPCM) to optimize the performance of Single Slope Solar still. Nanoparticles (Al_2O_3) spread in paraffin Wax (PCM) beneath the basin of solar still. The numerical solution was done using COMSOL 5.3 software. Results were compared with previous outcomes and showed significant agreement.

The addition of 1 kg of PCM constitutes the optimum amount of improvement. Thus, using 3 Vol.% concentration of Al_2O_3 Nano-Particles dispersed in 1 kg of Paraffin Wax(PCM) gives the perspective of improvement in the traditional single slope solar still productivity by about 20% daily. Ileri et al. [13] performed an experimental work for solar stills to analyze the glass cover thickness's effect on the productivity yield. In these solar stills, three glasses are of 3mm, 5mm, and 6 mm thickness while the fourth cover was made up of plastic.

Thermal modeling was developed for solar still and for getting the solution of the thermal equations, a programming software FORTRAN-77 was used. A numerical simulation was performed for Ankara's climate condition (40°N,33° E). The simulation results were compared with the experimental data.

The Newton-Raphson method was used for solving the mathematical model to find the roots of radiative heat transfer coefficients for glass and water temperature. Efficiency for 3 mm thickness glass was found to be 26.22 % more, as compared to 5mm and 6 mm. Hamadou et al. [14] worked on single slope solar still having a copper heating plate above heat transfer fluid to enhance the thermal effect. A MATLAB program was developed based on the command ode 15s to solve nonlinear differential explicit equations in matrix form.

©IJRASET: All Rights are Reserved

This work was enabled to optimize the inlet temperature, wind speed, heat transfer fluid, seawater rate, basin depth, and relative humidity but this technique was used only for the rate and yield of distilled water. The simulation results showed that wind speed has a major effect on distilled water production than relative humidity.

Kumar and Kumar [15] did simulation works to check the temperature difference in single and double slope solar still for the climate condition of New Delhi.

The geometry of both stills was created in ANSYS workbench and then simulated using ANSYS V19.0. As a result, it was observed single slope solar still has a higher temperature compared to double slope solar still. Zerrouki et al. [16] did a numerical simulation of capillary film solar still coupled with conventional solar still to obtain daily distillate production for Algeria's climate condition.

Mathematical modeling was done by using FORTRAN-90 to solve the non-linear equation of both systems using Ranga-Kutta-Fehlberg integration method.

As a result, capillary film solar still coupled with conventional solar still has more distillate yield than simple conventional solar still. Maalem et al. [17] used COMSOL multiphysics to simulate trapezoidal single slope solar still to check the temperature, velocity, and concentration field. Energy balance equations were solved by using the finite element method. The result shows that simulation work has good agreement with experimental work.

Thakur et al. [18] performed modeling work using computational fluid dynamics tools and experimental work for single slope solar still to check production at different depths. The geometry was created and then imported for meshing in ANSYS meshing with a minimum grid size of 4 mm and 10 mm to achieve better results. The meshing geometry has several nodes 632088 and the total number of elements 553048. The computational work was done for optimizing the different water depths of solar still i.e., 0.01, 0.02, and 0.03 m meshing type of hex-dominant was used. Result obtained through computational fluid dynamics work has a good agreement with the experimental data. The optimum depth for good productivity was 0.01m. Bhaisare et al. [19] developed a computational fluid dynamics model of double slope single basin solar still for the Gorewada water purification plant, Nagpur. The geometry was created in ANSYS workbench and then imported to ANSYS meshing.

The simulation result was obtained by using ANSYS 16.0v. Maximum heat flux decreased concerning time, i.e., from 9.1054e+005 to 2.0565e+005 in 1 sec and predicted maximum temperature remains constant and minimum temperature varies with time. Badusha and Arjunan [20] developed a two-phase three-dimensional model with the help of ANSYS. Still was made for evaporation and condensing process in solar still by using computational fluid dynamics techniques. There was good agreement between the simulation result and the experimental result with certain errors. The result predicted by ANSYS CFX shows that it is a potent tool for design, parameter analysis, and difficulty removal in solar still construction. Aybar and Assefi [21] performed an investigation for single-slope solar still to determine the effects of water depth in the basin and angle of inclination of the glass cover over the basin. The solar still is modeled mathematically using the 4^{th} order Runge Kutta (RK) method in FORTRAN. Optimum depth is found to be the least depth of water and the best inclination angle is the Latitude of the place of installation of the setup. The productivity of the system under local climatic conditions was found at 5.3 kg/m^2day. The written code results are compared with the experimental outcomes and the relative error is well within the permissible limits of 3.37%. The literature review shows no research work available to compare the water temperature, pressure, solar heat flux, and productivity of freshwater of single and double slope solar still and no further work available for New Delhi. Researchers have less attention doing simulation in ANSYS Fluent. This present study aims to create a three-dimensional computational fluid dynamics model of single slope and double slope to check the evaporation and condensation processes within solar stills at Delhi's climate conditions (27.0238° N, 74.2179° E). Both models were developed using ANSYS Workbench and simulation works were done with ANSYS FLUENT v19.2. All variables inside both solar still were estimated and examined at each stage. The results were compared with each other to check the differences. Water production of the single and double slope of both systems was also examined for the entire day of June 21, 2020.

II. METHODOLOGY

Two different two-phase 3D models were designed in the volume of fluid of the multiphase phase model. Both stills were also developed using evaporation and condensation processes at transient conditions, which means only surface evaporation of liquid occurred and their interface was considered for modeling. The performance of solar stills based on productivity as well as internal heat and mass transfer coefficient were studied. The internal heat and mass transfer coefficient in solar still based on convection, radiation, and evaporation. Hence, the convective heat transfer coefficient, radiative heat transfer coefficient, and evaporative heat transfer coefficient are for the heat transfer. An RNG k-ε turbulence model was applied with standard wall functions for both phases. The time and volume- average continuity, energy, and mass transfer equations were numerically solved in this work for each phase.

©IJRASET: All Rights are Reserved

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue XI Nov 2020- Available at www.ijraset.com

A. Governing Equations

Under steady-state conditions, model equations are dependent on continuity, momentum, energy, and mass transfer conservation concepts.

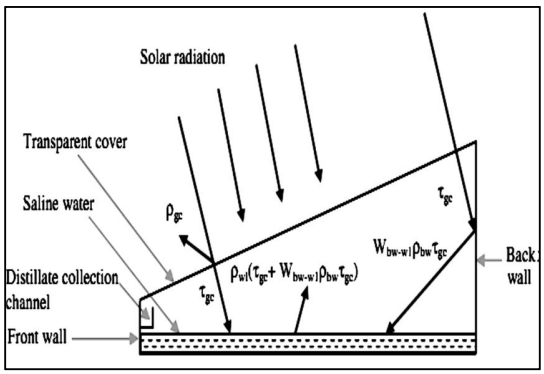


Fig 1. The single slope solar basin [4]

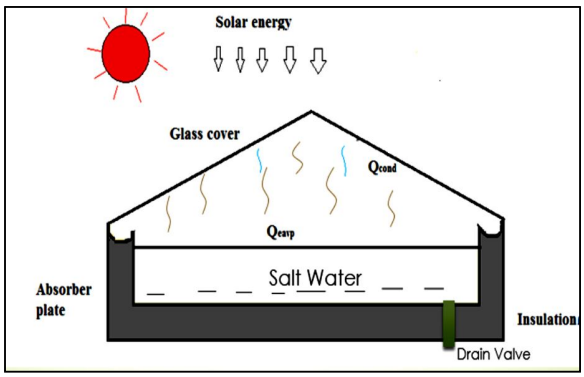


Fig 2. Double slope solar basin [6]

1) Energy Equations

The following assumption of energy balance equations have been made to write energy equation [4,5]:

1. There is no temperature gradient across the basin water and glass cover of solar still.
2. There is no leakage in solar still.
3. The water level is kept constant inside the basin.
4. The heat capacity of the glass surface and absorber, insulated material is neglected.

The energy equation of the mixture is given as:

$$\frac{\partial}{\partial t}\sum_{k=1}^{n}(a_k\rho_k E_k) + \nabla\cdot\sum_{k=1}^{n}\{a_k\bar{v}_k(\rho_k E_k + p)\} = \nabla\cdot\left(k_{eff}\nabla T\right) + S_E \quad (1)$$

where k_{eff} is the effective conductivity $(a_k(k_k + k_t)$,

a_k = the absorptivity

2) Continuity Equation

The continuity equation of the mixture is [4].

$$\frac{\partial}{dt}(\rho_m) + \nabla.(\rho_m\overline{v_m}) = 0 \quad (2)$$

where $\overline{v_m} = \frac{\sum_{k=1}^{n} a_k\rho_k\overline{v_k}}{\rho_m}$

$\overline{v_m}$ = the mass averaged velocity

ρ_m = Density of mixture

3) Momentum Equation

The momentum equation of all phases is achieved by combining each momentum equation [4]:

$$\frac{d}{dt}(\rho_m\bar{v}_m) + \nabla(\rho_m\bar{v}_m\bar{v}_m) = -\nabla p + \nabla[\mu_m(\nabla\bar{v}_m + \nabla v_m^T)] + \rho_m\bar{g} + \bar{F} + \nabla(\sum_{k=1}^{n} a_k\rho_k\bar{v}_{dr,k}\cdot\bar{v}_{dr,k})$$

(3)

ρ_m = Density of mixture

$\bar{F}$ = The gravitational body force and external body forces

μ_m= molecular viscosity of the mixture

4) Volume Conservation Equation

It is a constraint that shows the unity of volume fractions by adding [4].

$$r_G + r_L = 1 \quad (4)$$

r_G = volume fraction of gas.

r_L = volume fraction of liquid.

©IJRASET: All Rights are Reserved

5) Mass Transfer Equation

The enthalpy sources in a system for different phases are [4].

$$H_p = -m_p i_q j\left(h_p^i\right) \quad (5)$$

$$Hq = m_{p^i q} j\left(h_p^i + h_p^{f^i} - h_q^{f^j}\right) \quad (6)$$

$h_p^{f^i}$ = formation enthalpy of species i of phase p.

$h_q^{f^j}$ = formation enthalpy of species i of phase q.

h_p^i = enthalpy of species i of phase p.

H_p = total enthalpy of phase p

Hq = total enthalpy of phase q.

6) Pressure Constraint

The gas phase and liquid phase share the same pressure field [3]

$P_G = P_L = P$ (7)

Pg = Partial saturated vapor pressure at the gas temperature

7) Heat Transfer Equation

The two-resistance model was expressed for interphase heat transfer has been used. The heat transfer coefficient across the interface and the gas phase was modeled using the zero-equation model and the liquid phase has been expressed by [3]:

$$h_{total} = h_{r\omega} + h_{cw} + h_{e\omega} \quad (8)$$

$$h_{Cw} = 0.8840(Tw - Tg + \frac{(P_\omega - p_g)(T_w+273)}{268{\cdot}9\times10^3 - P_w})^{1/3}$$

where, $P_w = exp\left(25.317 - \frac{5144}{T_w+273}\right)$

where $Pg = \exp\left(25317 - \frac{5144}{Tg+273}\right)$

$$h_{e\omega} = 16.273 \times 10^{-3} h_{Cw} \cdot \frac{P_\omega - P_g}{T_w - T_g} \quad (9)$$

And,

$$h_{rw} = \varepsilon_{eff}\sigma[(TW + 273)^2 + (Tg + 273)^2(Tw + Tg + 546)]$$

where, $\varepsilon_{e_{ff}} = \frac{1}{\varepsilon_g} + \frac{1}{\varepsilon_\omega} - 1$

8) Water Production

Defined contact mass flux was used for mass transfer models. Some assumption of water production was used by Eq. (3). The rate of freshwater production is equivalent to the rate of vaporization of the water. So, the rate of vaporization of the water shows water production. Hence, the mass flux equation between the two different phases are:

$$\dot{m}_{ew} = \frac{\dot{q}_{ew}\cdot A_w \cdot t}{h_{fg}} \quad (10)$$

where,

$$h_{fg} = 2\cdot 4935 \times 10^6[1 - 9.4779 \times 10^{-4}T + 1.3132 \times 10^{-7}T^2 - 4.7974 \times 10^{-9}T^3]$$

$\dot{q}_{ew} = h_{ew}\left(T_w - T_g\right)$

$\dot{m}_{ew}$= distillate output

9) Solar Irradiance

$$\delta = 23\cdot 45\sin(360 * (n + 284) / 365)$$

$$\omega = (solar\ time - 12{:}00) \times 15$$

where,

$$Solar\ time = Standard\ time \pm 4\left(L_{st} - L_{L_0 ca}\right) + E$$

$Ig = I_b + I_d$ (11)

©IJRASET: All Rights are Reserved

where, $I_b = I_{bn}\cos\theta_z$ *and*, $Ig = I_{bn}\cos\theta_z + I_d$
δ = sun declination angle
n = day of the year, starting from 1st January
ω = Hour angle, degree.
Ig = Hourly global irradiance
I_b = Hourly diffuse radiation
I_d = Hourly diffuse radiation
I_{bn} = Beam radiation on the surface normal to the direction of the sun rays.

B. Computational Fluid Dynamics Model of Modeling

Computational Fluid Dynamics (Computational Fluid Dynamics) is a valuable tool that has been used to analyze and investigate the fluid flow pattern of moisturized air and temperature variation near the walls. For visualizing the flow filed, any location computational fluid dynamics can prepare the design, get the controlling step, and guide. In the study, computational fluid dynamics is utilized to create geometry, and the simulation result is obtained by using ANSYS 19.2v.

1) Geometry Creation: ANSYS is a beneficial engineering simulation software that is mostly used to solve engineering problems and make them more optimized. Figure 3 and Figure 4 show the 3D geometry of single slope and double slope solar still, which were modeled using Ansys Design modeler in Ansys 19.2.

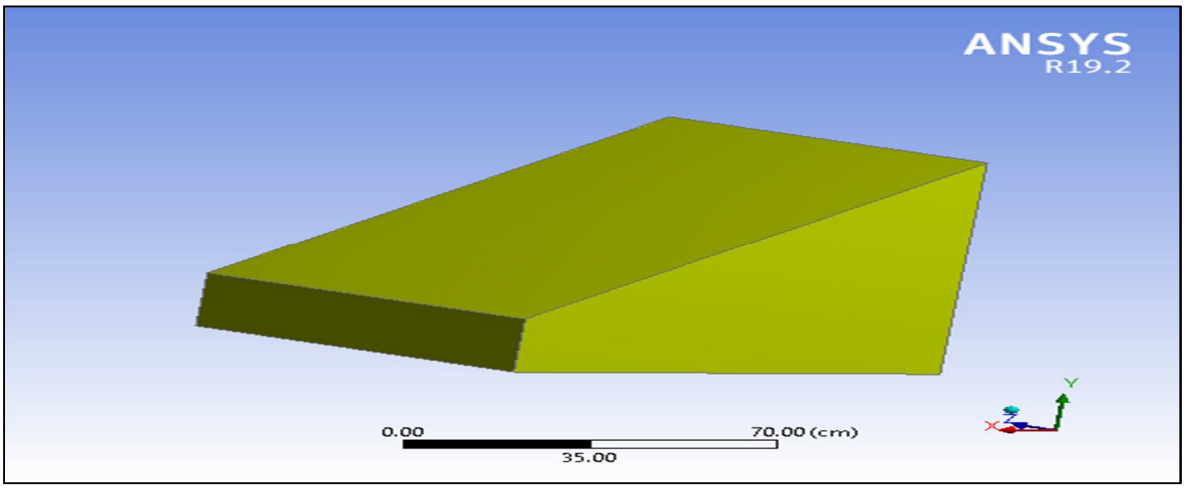


Fig. 3. Single Slope Single Solar Still.

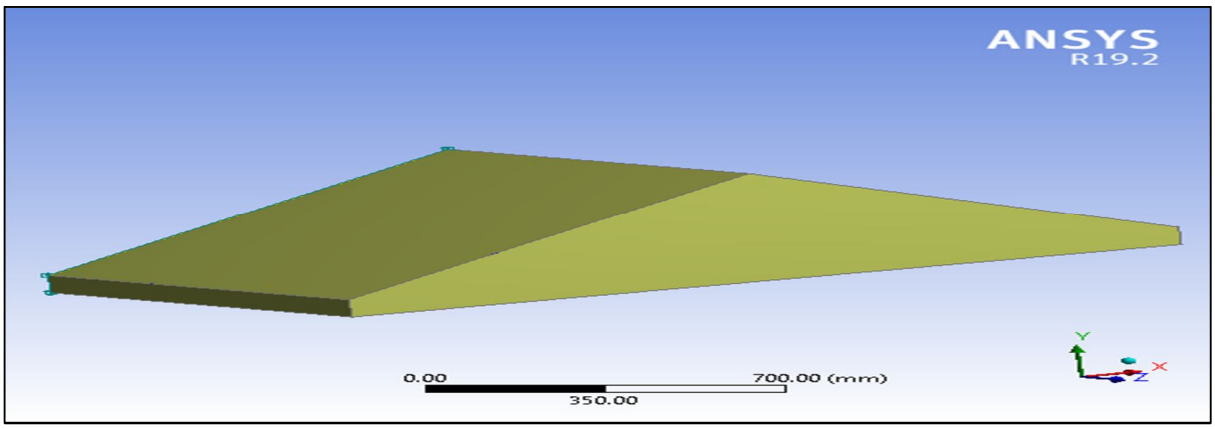


Fig. 4. Double Slope Single Solar Still.

Both stills consist of absorber parts, sidewalls on both sides, front wall, back wall, and collector (single collector in single slope and double collector in double slope) are modeled as per the dimensions.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue XI Nov 2020- Available at www.ijraset.com

2) *Meshing:* The physical model of both systems meshed by using ANSYS 19.2 workbench Meshing. The stills' results were checked at the different number of meshing elements as 40560, 54500, and 62430 for a single slope and 45320, 586200, and 64130 for a double slope. But, the results are predicted much closer at 54500 elements of single while 586200 elements. Figure 5 shows an unstructured mesh of single slope solar still, consisting of 545000 elements and 566610 nodes at a growth rate of 1.2 and element size 10mm using 3D hexahedral meshing.

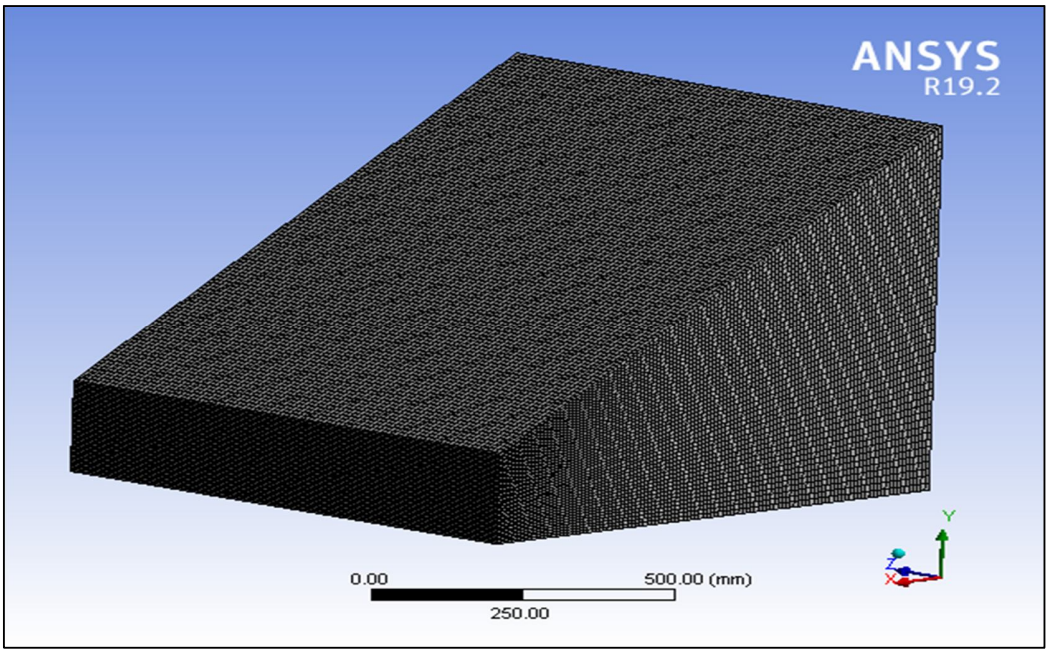


Fig. 5. The unstructured mesh of single slope.

Figure 6 shows an unstructured mesh of slope solar still, consisting of 586200 elements and 612464 nodes at a growth rate of 1.2 and element size 10 mm using 3D hexahedral meshing.

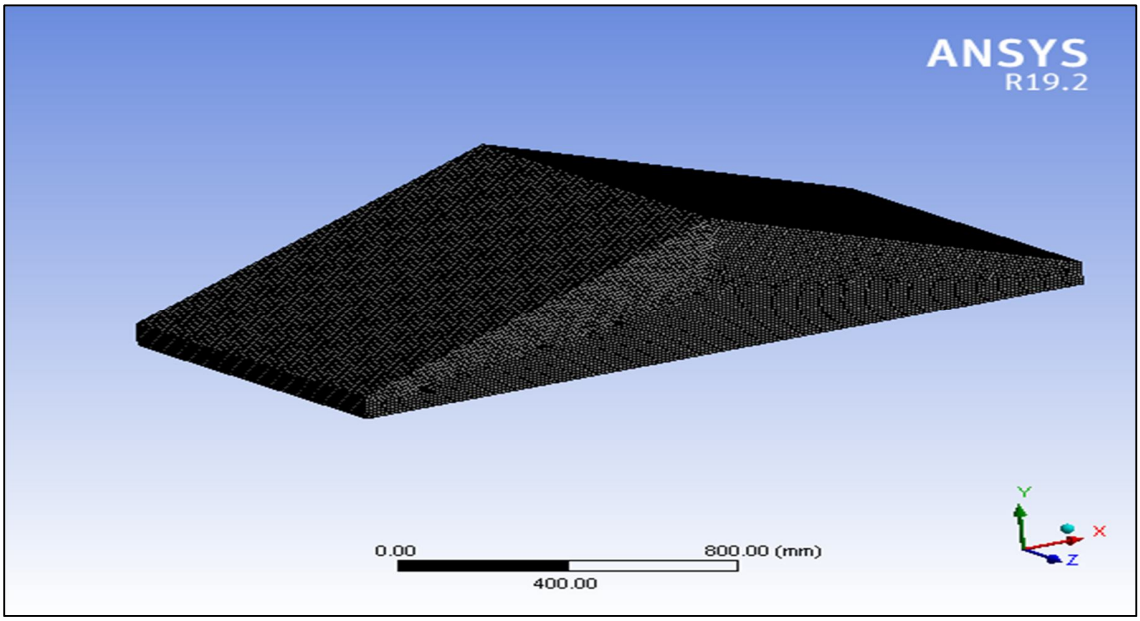


Fig. 6. The unstructured mesh of double slope still.

©IJRASET: All Rights are Reserved

3) *Boundary and Initial Conditions:* Meshing files of both solar stills are imported to ANSYS FLUENT. A two-phase model of both solar still was developed in the volume of fluid (VOF) framework for liquid (water) and a mixture of air and water vapor system at the transient condition with gravitational acceleration. Solar insolation incident on solar stills is the important factor inside stills, based on the absorptivity and transmissivity of the glass cover. First, it is incident upon glass cover and then absorbed by an absorber, which increases the water temperature. In the fluent, physics and boundary conditions are specified to solve the continuity equation, momentum equation, and energy equation at all the boundaries. The RNG k-ε viscous model was used, which predicts the spreading rate of flow and better water production performance. In the model, the near-wall condition was treated using 'standard wall function". Adhesion forces are taken for producing droplets inside the solar stills. No-slip wall boundary conditions were specified for both gas phase and liquid phase. Initial water level at the initiation of the stills' first stage was 2 cm, so water and air volume fraction were considered as 0.13 and 0.87. Computational fluid dynamics simulation had run time of 5 hours for single slope solar still and 6.30 hours for double slope solar still.

Table 1: Boundary conditions for the simulation process of single slope and double slope

Domain	Domain type	Location	Boundary type	Boundary details
Solid	Cell	Absorber plate	Wall	Stationary wall with heat flux
		Glass cover	Wall	Fixed wall Temperature
		Sidewalls	Wall	Adiabatic
		Front and back wall	Wall	Adiabatic

III. RESULT AND DISCUSSION

In this study, simulation results were calculated at fair weather conditions, and each stage of both stills using ANSYS FLUENT v19.2. In Simulation results, the temperature distribution, velocity, water volume fraction, solar heat flux, and water production of both solar stills are examined and predicted the temperature and water production of both stills for the entire day.

A. Simulation Results

ANSYS FLUENT v19.2 was used to carry out the results using 3 GHz CPU processors as parallel run with double precision. Fluent uses a second-order upwind solution method to transform the governing equations into a numerically solvable algebraic equation.

1) *Temperature Distribution:* Computational fluid dynamics predict both solar stills' water temperature and are compared with each other at every stage. In the volume rendering of Figures 7 and 8, the blue color shows the minimum temperature while the red color shows the maximum value of temperature. The maximum and minimum temperature predicted inside single slope solar still, are 327.9445 K and 292.4511 K, whereas in double slope solar stills maximum temperature is 323.6474 K and the minimum temperature is 294.0343K. There is an approximately 8.36% error in their maximum temperatures and an 8.139% error in minimum temperatures.

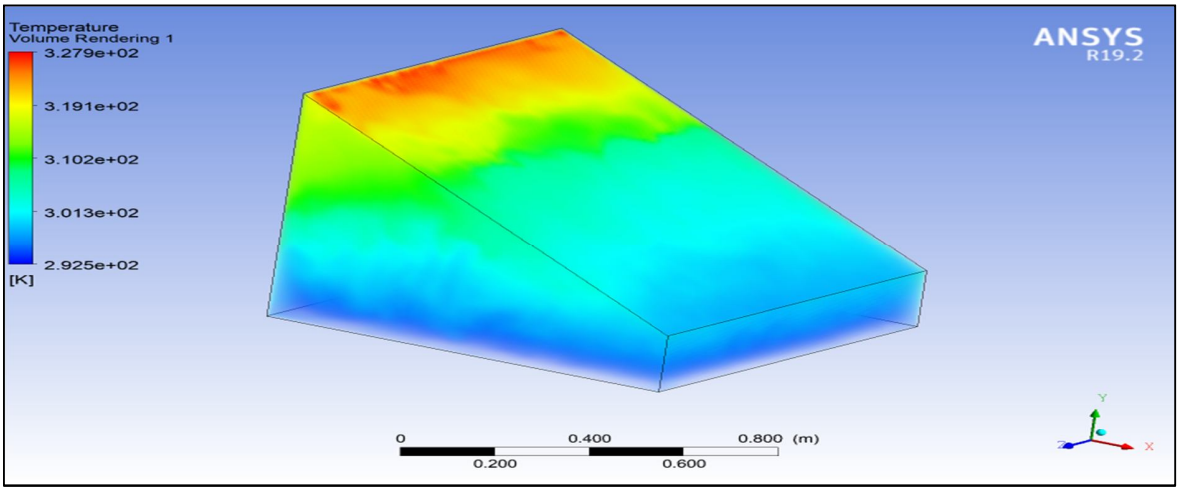


Fig. 7. Temperature Distribution inside Single slope

©IJRASET: All Rights are Reserved

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue XI Nov 2020- Available at www.ijraset.com

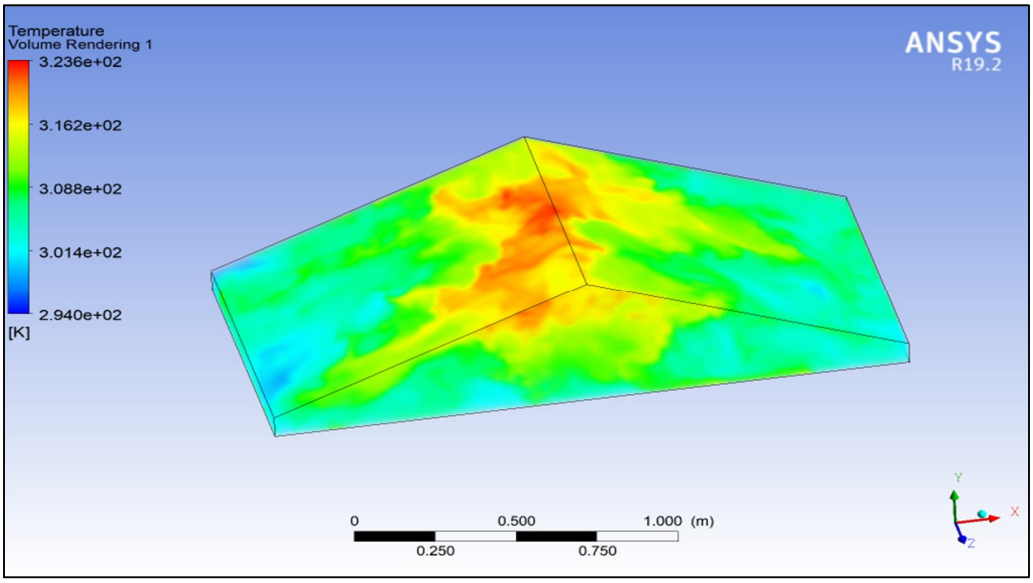


Fig. 8. Temperature Distribution inside double slope still

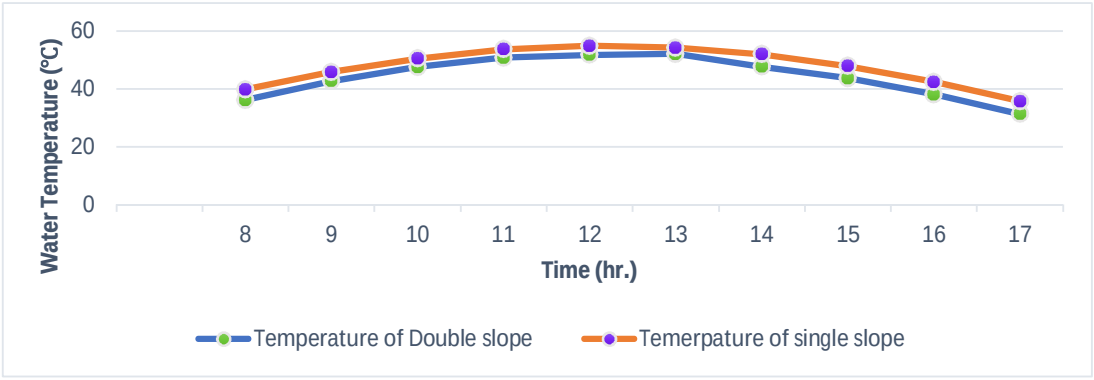


Fig. 9. The water temperature of both single and double slope solar still

The temperature of both solar stills is predicted from 08:00 to 17:00 hr. Figure 9 shows the temperatures of both solar stills at different hours. The average water temperature error from 08:00 to 17:00 hr was predicted at 7.932%, respectively. The maximum temperature was attained at double slope solar still at 12:00 noon while in single slope solar still at 13:00 hr.

Table 2: Simulation results for hourly variation of water and glass temperature, Solar intensity, and yield of single slope solar still

Time (hr)	Water Temperature (°C)	Glass Temperature (°C)	Solar Intensity (W/m^2)	m_{ew} ($kg/m^2.hr$)
08:00	39.9545	38.281	421.27	0.265
09:00	45.9567	43.8938	585.559	0.184
10:00	50.5961	48.9845	720.719	0.1595
11:00	53.7344	52.7698	798.367	0.16782
12:00	54.9445	53.562	845.284	0.315
13:00	54.2991	52.879	828.131	0.41762
14:00	52.0504	50.763	770.578	0.60729
15:00	47.988	46.8934	671.823	0.5812
16:00	42.4538	40.342	503.072	0.45152
17:00	35.7752	31.5839	325.553	0.364826

©IJRASET: All Rights are Reserved

Table 3: Simulation results for hourly variation of water and glass temperature, Solar intensity, and yield of Double slope solar still.

Time (hr)	Water Temperature (°C)	Glass Temperature (°C)	Solar Intensity (W/m^2)	m_{ew} ($kg/m^2.hr$)
08:00	36.2	33.561	408.742	0.20769
09:00	42.7615	41.6523	525.954	0.1527
10:00	47.6474	46.0343	680.782	0.13972
11:00	50.8486	48.0921	735.548	0.14873
12:00	51.8289	49.6988	810.132	0.2856
13:00	52.21	51.0652	790.095	0.39812
14:00	47.7914	45.0947	710.885	0.54371
15:00	43.7765	41.9073	598.516	0.50238
16:00	38.22	36.930	430.827	0.42621
17:00	31.3583	26.9563	289.362	0.3048201

2) *Water Volume Fraction:* As the water temperature increases, the inside area of both the stills starts warming up and water start changing into vapors. Figure 9 shows the condensed water droplets on the glass of a single slope, while Fig. 10 shows a side view of single slope solar still. The presence of water can be only seen on glass and near the base of still in both solar stills. The maximum and minimum volume fraction of water was obtained 0.0183968 and 1.8773×10^{-6}.

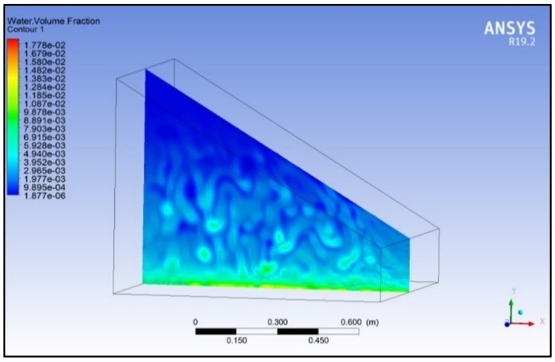


Fig. 10. Water volume Fraction contour inside single slope

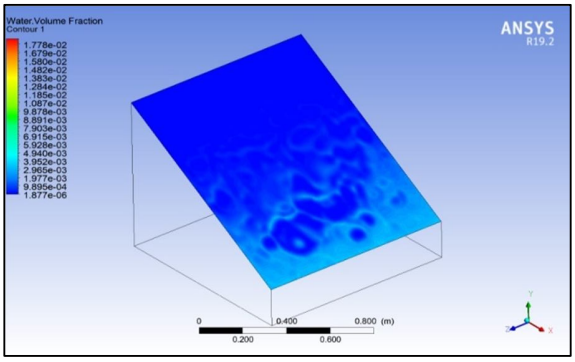


Fig. 10. Water volume Fraction contour on the glass of single slope solar still

The blue color shows less amount of the volume fraction of water, while the red color shows more water volume. The volume fraction of droplets is increasing downward on the glass. The maximum water volume fraction on the glass is observed 0.0309681 and a minimum of 3.22×10^{-5}. It is predicted that the water volume fraction of glass of double slope is higher compared to glass of single slope.

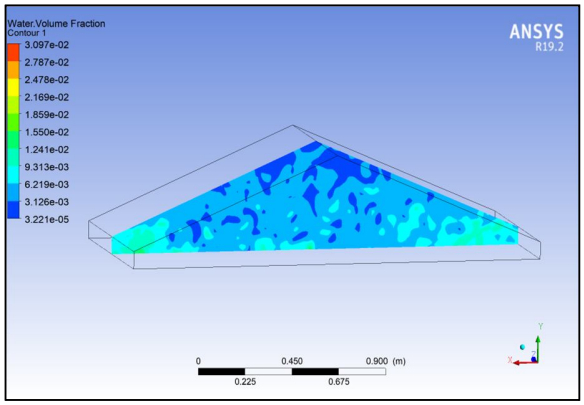


Fig. 11. Water volume Fraction contour inside double slope

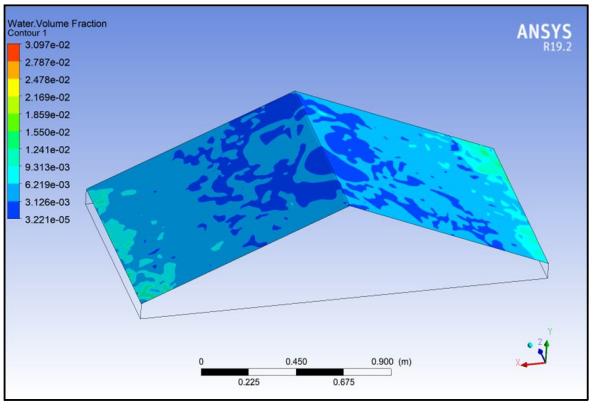


Fig. 12. Water volume Fraction contour on the glass of single slope solar still

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue XI Nov 2020- Available at www.ijraset.com

3) *Velocity Vector:* The vapor phase velocity vector of single slope solar still on a plane is shown in Figures 13 and 14. First, the water changed into vapors and started to move upward due to the buoyancy effect. The vapor phase follows a circular path from the bottom to the glass surface. Vapor condensate on the glass and changes into water droplets, which move downward on the glass surface. This process takes place in both solar stills, which shows the heat transfer mechanism. There is a difference in vapor velocity in both solar stills, which more near the bottom in single slope solar still while in double slope solar still. The maximum velocity of vapor in the single slope was observed 2.192 m/sec, while in double slope solar still, it is only 0.70289 m/sec—this difference of vapor velocity influences on water production.

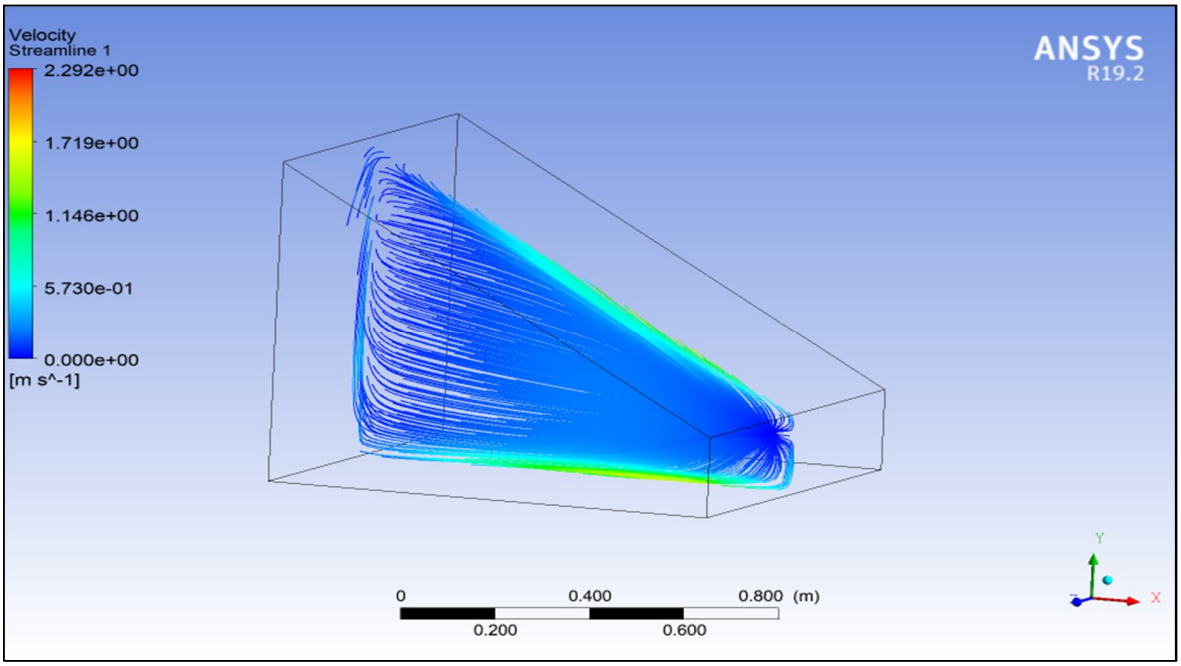


Fig. 13. Vapor phase velocity vector of single slope solar still

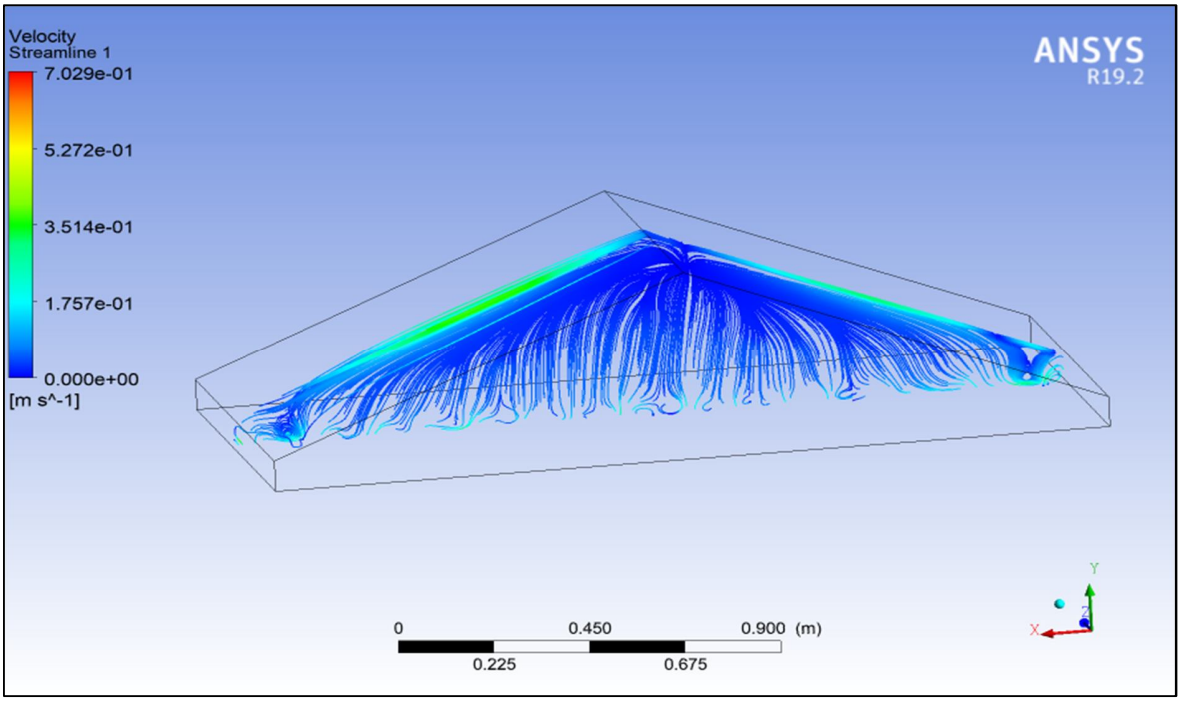


Fig. 14. Vapor phase velocity vector of single slope solar still

©IJRASET: All Rights are Reserved

4) *Solar Intensity:* When Solar energy incidents on the glass surface increase the inside temperature of the solar still and converts water into vapor. Fig. 15 shows the hourly variation of the solar intensity of both stills. The result predicted a single slope gets a higher intensity of solar energy than double slope solar still. There is approximately an error of 8.1846 % in the solar intensity of both stills. The maximum solar intensity absorbed by the single slope is 845.284 W/m^2 while in double slope is 810.132 W/m^2. The average solar intensity of the single-slope solar still is 647.0356 W/m^2. While in the double slope, the average solar intensity was 598.0843 W/m^2.

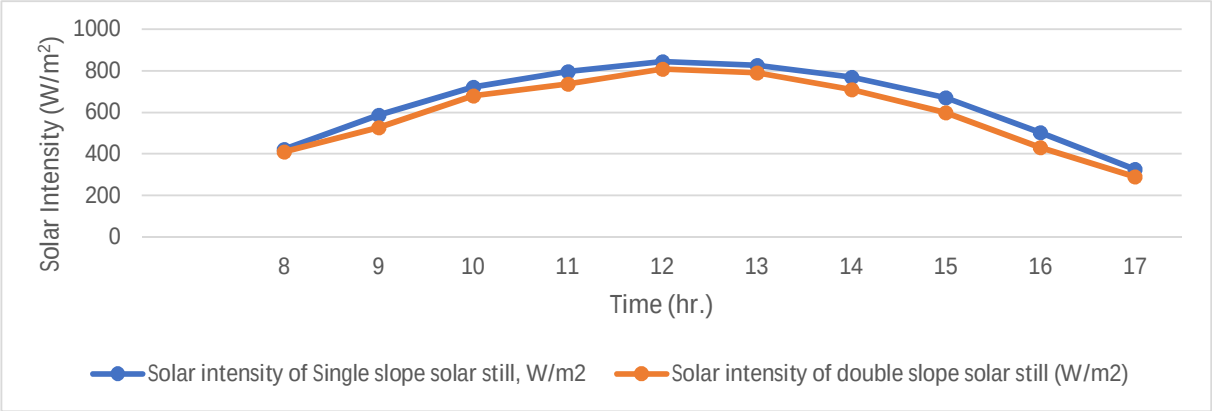


Fig. 15. Hourly variation of solar intensity of single and double slope solar still

5) *Water Production:* Fig. 16 shows the hourly variation of the mass yield of both stills. The simulation process was started at 08:00 hr and came to an end at 17:00 hr. As time passes, the inside area of both still warm-up and still space become saturated. The rate of production of freshwater increases from 11:00 to 14:00 hr in both still. Then, it starts to decrease. The maximum production of freshwater in single is predicted to be 0.60729 kg/m^2 hr at 14:00 hr while the double slope is of 0.54371 kg/m^2 hr. There is an approximately 11.69% difference in water production in both stills.

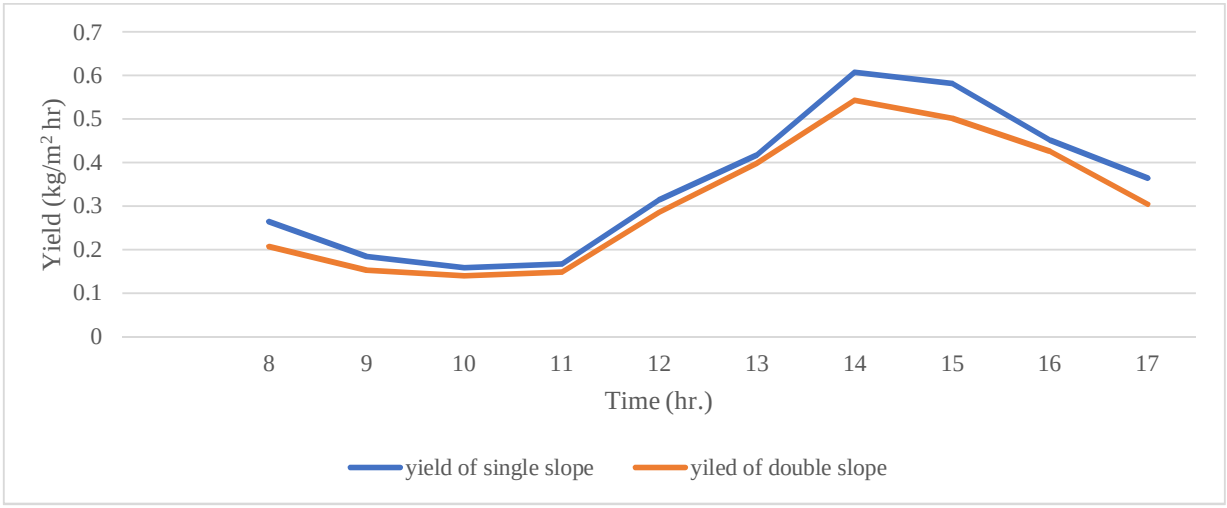


Fig. 16. Hourly variation mass yield of single and double slope solar still.

IV. CONCLUSIONS

The study aims to develop a computational fluid dynamics model of single slope and double slope solar still. Two different two-phase three-dimensional models using computational fluid dynamics have been developed. In both stills, evaporation and condensation processes occur. The models were conducted for water-mixture (air and water vapor) systems with the aim of ANSYS FLUENT V19.2 software. The computational fluid dynamics simulation was conducted for transient conditions using these models. Both systems' simulation data were obtained in 5 hours and 6.30 hr with 10 steps of 15 minutes period. In both systems, water absorbs solar radiation and evaporates and the vapor condenses on the glass. The main conclusions of the study are:

©IJRASET: All Rights are Reserved

A. The maximum temperature of air and water vapor mixture inside the single slope is 54.9445℃ and the minimum temperature is 19.4401 ℃ The maximum temperature of the mixture inside the double slope is 52.21 ℃ and the minimum temperature is 12.6988℃. There is an approximately 8.36 % error in their maximum temperatures and an 8.139 % error in minimum temperatures.

B. The maximum volume fraction is 0.0183968 and a minimum of 1.8773×10^{-6} in the single slope, while the maximum volume fraction inside the double slope is 0.0309681 and a minimum of 3.22×10^{-5}.

C. The maximum velocity of vapor in the single slope was observed 2.192 m/sec while in double slope solar still, it is only 0.70289 m/sec.

D. The maximum production of freshwater in single is predicted to be 0.60729 kg/m^2 hr at 14:00 hr while the double slope is of 0.54371 kg/m^2 hr. There is an approximately 11.69% difference in water production in both stills.

Computational fluid dynamics results show that computational fluid dynamics is a powerful tool for design, parameter analysis, and difficulties removal in solar still construction. In the future, further work can be done by changing different design parameters of solar still.

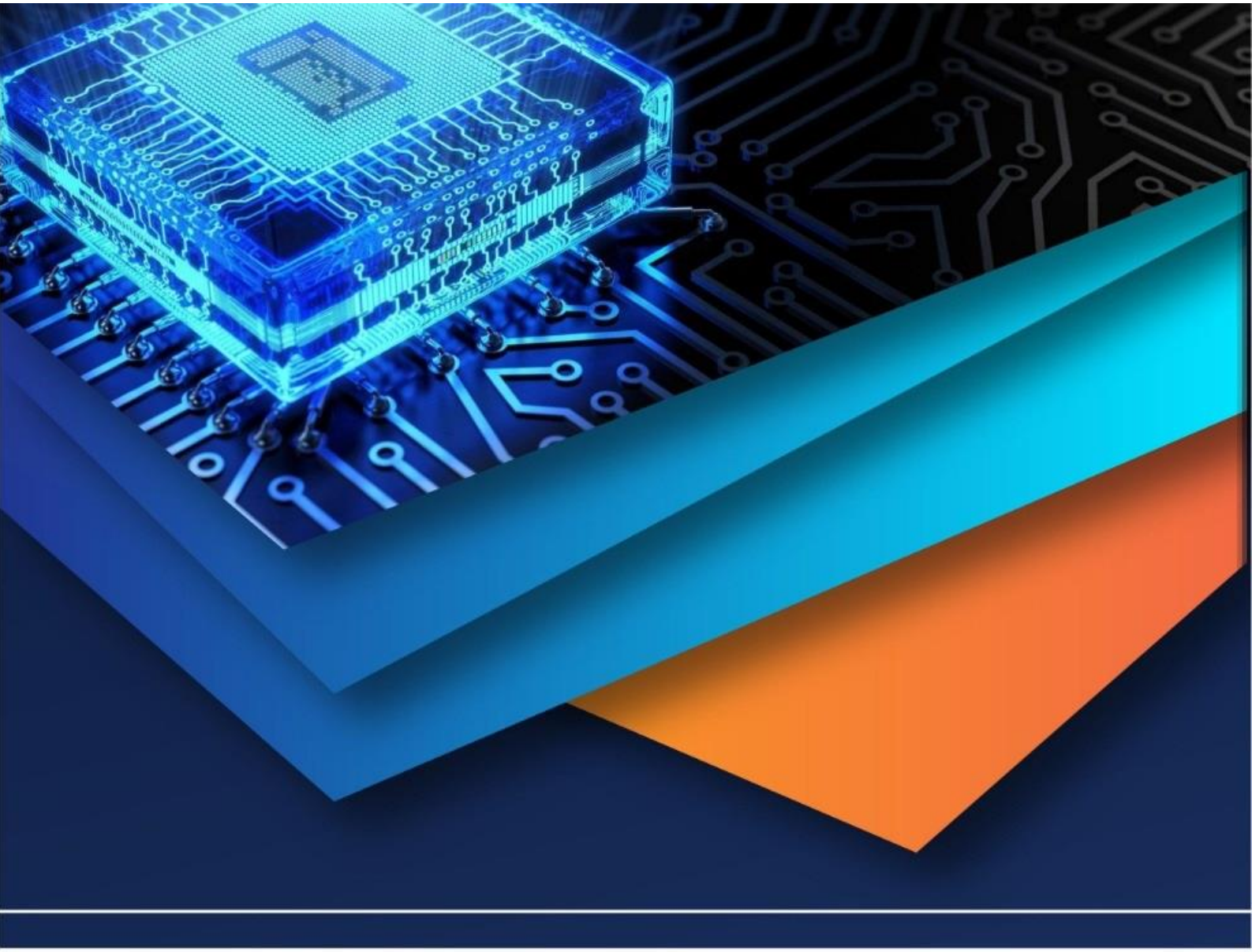
V. ACKNOWLEDGMENT

The authors are highly thankful to the Centre for Energy and Environment and Department of Mechanical Engineering, Delhi Technology University, Delhi (India) for their assistance and facilities in compiling this review manuscript

REFERENCES

[1] water distribution/<https://www.dw.com/en/are-we-running-out-of-fresh-water/a-40241057/>
[2] Waterborne diseases /< https://www.who.int/news-room/fact-sheets/detail/drinking-water />
[3] Setoodeh N, Rahimi R, Ameri A (2011) Modelling and determination of heat transfer coefficient in-basin solar still. Desalination, 268,103-110.
[4] Khare V R, Singh P A, Kumar H, Khatri R (2017) modeling and performance enhancement of single slope solar still using computational fluid dynamics. Energy Procedia,109, 447—455.
[5] Panchal H.N, Shah P K (2011) Modelling and verification of single slope still using ANSYS-CFX. International Journal of Energy and Environment,2(6), 985-998.
[6] Maheswari C.U, Reddy B.V, Sree A.N, Reddy V, Reddy S.P, Prasad R.R, Varma B.H.K (2016) Computational fluid dynamics analysis of single basin double slope solar still. Invention Journal of Research Technology in Engineering & Management, 2, 01-05.
[7] Gokilavani NS, Prabhakaran D, Kannadasan T (2014) Experimental studies and CFD modeling on solar distillation system. Int J Innov Res Sci Eng Technol, 3(9), 15818-15822.
[8] Mahendren R, Senthilkumar R, Irfan Ali MD (2011) modeling of solar still using granular activated carbon in Matlab. Bonfring Int J Power Sys and Int circuits, 1, 5-10.
[9] Nayak AK, Dev R (2016) thermal modeling and performance study by modified double slope solar still. Int J Res in Eng and Technol, 5, 19-23.
[10] Tripathi R, Tiwari GN (2005) effect of water depth on internal heat and mass transfer for active solar distillation. Desalination, 173, 187-200.
[11] Chaibi MT (2000) analysis by simulation of a solar still integrated with a greenhouse roof. Desalination, 128, 123-138. [15] Singh N (2013) performance analysis of single slope solar stills at different inclination angles: an indoor simulation. Int J Curr Eng Technol, 3(2), 2277- 4106.
[12] Ali Saeed M M, Mohsen MS, Hachim D M, Hameed H G (2019) numerical simulation for single slope solar still performance with optimal amount of nano-PCM. Journal of advanced Res in Fluid Mech and Thermal Sci, 63, 302-316.
[13] Ghoneyem A, Ileri A (1997) software to analyze solar stills and an experimental study on the effects of the cover. Desalination, 114, 37-44.
[14] Hamadou OA, Abdellatif K (2014) Modeling an active solar still for sea water desalination process optimization. Desalination, 354, 1-8.
[15] Kumar A, Kumar R (2010) Numerical Simulation of Single and Double slope solar still for Temperature Distribution. Int Conf on Recent Adv in Mech Engg, 6, 221-231.
[16] Zerrouki M, Settou N, Marif Y, Belhadj MM (2014) Simulation Study of Capillary film solar still coupled with a conventional solar still in South Algeria. Energy Convers Manag, 85, 112-119.
[17] Maalem MS, Benzaoui A, Bouhenna A(2014) Modeling of simultaneous transfers of heat and mass in a trapezoidal solar distiller. Desalination, 344, 371-382.
[18] Thakur AK, Pathak S.K (2017) Single basin solar still with varying depth of water: optimization by the computational method. Iranian Journal of Energy & Environment, 8(3), 216-223.
[19] Bhaisare A, Hiwakar A, Sakhare A, Ukey S, Purty S, Wasnik U, P V.K (2019) Brackish Water Distillation System for Gorewada Water Treatment Plant, Nagpur By Using Solar Energy- A Case Study. World Journal of Engineering Research and Technology,198-215.
[20] Badusha R, Arjunan T.V (2013) Performance analysis of single slope solar still. international journal of mechanical engineering and robotics research,3(3), 2278-0149.
[21] Aybar H S, Assefi H (2009) simulation of a solar still to investigate water depth and glass angle. Desalination and water treatment, 7, 35-40.

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)

arXiv:2101.00353v1 [math.CV] 2 Jan 2021

ON A GENERALIZED BRIOT-BOUQUET DIFFERENTIAL SUBORDINATION

S. SIVAPRASAD KUMAR AND PRIYANKA GOEL*

Abstract. We introduce and study the following special type of differential subordination implication:

$$p(z)Q(z)+\frac{zp'(z)}{\beta p(z)+\alpha}\prec h(z)\quad\Rightarrow p(z)\prec h(z), \tag{0.1}$$

which involves generalization of the Briot-Bouquet differential subordination, where $Q(z)$ is analytic and $0\neq\beta,\alpha\in\mathbb{C}$. Further, some special cases of our result are also discussed. Finally, analogues of open door lemma and integral existence theorem with applications to univalent functions are obtained.

1. Introduction

Let $\mathcal{H}$ be the class of analytic functions defined on the open unit disc $\mathbb{D}:=\{z\in\mathbb{C}:|z|<1\}$. For any positive integer n and complex number a, $\mathcal{H}[a,n]$ denotes the subclass of $\mathcal{H}$ consisting of functions of the form $f(z)=a+a_nz^n+a_{n+1}z^{n+1}+\cdots$. Let $\mathcal{A}_n$ be the class of functions of the form $f(z)=z+a_{n+1}z^{n+1}+\cdots$ and denote $\mathcal{A}:=\mathcal{A}_1$. The subclass of $\mathcal{A}$ consisting of univalent functions is denoted by $\mathcal{S}$. A function $f\in\mathcal{A}$ is said to be typically real in $\mathbb{D}$ if for every non-real z in $\mathbb{D}$, we have $sign(\operatorname{Im}f(z))=sign(\operatorname{Im}z)$. The class of all such functions is denoted by $\mathcal{TR}$. Let $\mathcal{P}$ be the class of analytic functions of the form $p(z)=1+c_1z+c_2z^2+\cdots$ such that $\operatorname{Re}p(z)>0$ for all z in $\mathbb{D}$, $\mathcal{P}$ is known as the Carathéodory class. Given two analytic functions f and F, we say that f is subordinate to F, denoted by $f\prec F$ if there exists an analytic function $\omega(z)$, with $\omega(0)=0$ and $|\omega(z)|<1$, such that $f(z)=F(\omega(z))$. In particular, if F is univalent then $f\prec F$ if and only if $f(0)=F(0)$ and $f(\mathbb{D})\subset F(\mathbb{D})$. Assume $\mathcal{R}$ to be the class of functions with bounded turning consisting of all the functions $f\in\mathcal{A}$ such that $\operatorname{Re}f'(z)>0$ $(z\in\mathbb{D})$, clearly $\mathcal{R}\subset\mathcal{S}$. A function f in $\mathcal{S}$ is said to be starlike if and only if $\operatorname{Re}(zf'(z)/f(z))>0$ in $\mathbb{D}$ and the class of starlike functions is denoted by $\mathcal{S}^*$. Similarly, the class of convex functions, denoted by $\mathcal{K}$ consists of all those functions f in $\mathcal{S}$ for which $\operatorname{Re}(1+zf''(z)/f'(z))>0$ in $\mathbb{D}$. Let $\mathcal{S}^*(\alpha)$ $(0\leq\alpha<1)$ be the subclass of $\mathcal{S}^*$ consisting of the functions f satisfying $\operatorname{Re}(zf'(z)/f(z))>\alpha$. We say that a function f is strongly starlike of order γ $(0<\gamma\leq1)$, whenever $|\arg(zf'(z)/f(z))|<\gamma\pi/2$ and the class of strongly starlike functions is denoted by $\mathcal{SS}^*(\gamma)$, note that $\mathcal{SS}^*(1)=\mathcal{S}^*(0)=\mathcal{S}^*$. Coman [6] defined that a function $f\in\mathcal{A}$ is said to be almost strongly starlike of order $\alpha,\alpha\in(0,1]$, with respect to the function

2010 *Mathematics Subject Classification.* 30C45, 30C55, 30C80.

Key words and phrases. Starlike functions, Briot Bouquet, Subordination.

*Corresponding author

The work presented here was supported by the Council of Scientific and Industrial Research(CSIR). Ref.No.:08/133(0018)/2017-EMR-I.

$g \in \mathcal{S}^*(1-\alpha)$ if

$$\frac{g(z)f'(z)}{g'(z)f(z)} \prec \left(\frac{1+z}{1-z}\right)^{\alpha} \text{ or equivalently, } \left|\arg \frac{g(z)f'(z)}{g'(z)f(z)}\right| < \alpha\frac{\pi}{2}$$

and concluded that such functions are starlike and hence univalent. Recently Antonino and Miller [3] defined the class of F-starlike functions, denoted by $\mathcal{FS}^*$ as

$$\mathcal{FS}^* = \{f \in \mathcal{A} : \operatorname{Re}\left(\frac{F(z)f'(z)}{F'(z)f(z)}\right) > 0\},$$

where F is fixed univalent function on the closed unit disk $\mathbb{D}$, with at most a single pole on $\partial\mathbb{D}$ and $F(0) = 0$. In the year 1992, Ma and Minda [13] came up with a generalized subclass of $\mathcal{S}^*$ as well as $\mathcal{K}$, defined using subordination. They considered an analytic function φ, having positive real part, which is starlike with respect to $\varphi(0) = 1$, symmetric about the real axis and satisfies $\varphi'(0) > 0$ on $\mathbb{D}$. Various subclasses of $\mathcal{S}^*$ are evolved for different choices of φ in due course of time (see [5, 9, 14, 20, 24]). Note that φ is a typically real function as $\varphi'(0) > 0$. Miller and Mocanu [16] introduced the theory of differential subordination as an analogue to differential inequalities, using the concept of subordination. The gist of this whole theory is the following implication:

$$\psi(p(z), zp'(z), z^2p''(z)) \prec h(z) \Rightarrow p(z) \prec q(z), \qquad z \in \mathbb{D}, \tag{1.1}$$

where $\psi(p(z), zp'(z), z^2p''(z))$ is analytic in $\mathbb{D}$. Given a complex function $\psi : \mathbb{C}^3 \times \mathbb{D} \to \mathbb{C}$ and a function h, univalent in $\mathbb{D}$, if p is analytic in $\mathbb{D}$ and satisfies the differential subordination

$$\psi(p(z), zp'(z), z^2p''(z)) \prec h(z), \tag{1.2}$$

then p is called a solution (1.2). The univalent function q is called a dominant of the solutions of the differential subordination, if $p \prec q$ for all p satisfying (1.2). A dominant $\tilde{q}$ that satisfies $\tilde{q} \prec q$ for all dominants q is said to be best dominant of (1.2). The implication (1.1) gives rise to three types of differential subordination problems stated in [16, Chapter.2] and a good deal of literature associated with them is available (see [1, 4, 8, 9, 11, 12, 17, 19]). Using this theory, a special type of first order differential subordination, known as Briot-Bouquet differential subordination defined by

$$p(z) + \frac{zp'(z)}{\beta p(z) + \alpha} \prec h(z), \tag{1.3}$$

was studied by Miller and Mocanu [16]. Many implication results were proved later associating (1.3). Ruschewyh and Singh [22] considered Briot-Bouquet differential subordination in a more particular form given by

$$p(z) + \frac{zp'(z)}{\beta p(z) + \alpha} \prec \frac{1+z}{1-z}$$

with $\alpha \geq 0$ and $\beta > 0$. Later it was generalized to the form given by (1.3), in which $h(z)$ is taken to be a univalent function and $\alpha, \beta \neq 0$ are extended to complex numbers. This particular differential subordination has vast number of applications in the univalent function theory, see [2, 6, 7, 15, 18, 23] and the references therein. Motivated by the above work, in the present investigation, we consider the following differential subordination

$$p(z)Q(z) + \frac{zp'(z)}{\beta p(z) + \alpha} \prec h(z),$$

which is clearly a generalization of the Briot-Bouquet differential subordination and find conditions on α, β and $Q(z)$ so that the implication (0.1) holds. Further, we establish certain subordination results analogous to open door lemma and integral existence theorem. Apart from deriving other similar results, we also find sufficient conditions for starlikeness and univalence as applications of our results. For a better understanding of our main results, one may refer to [16, Ch.2] for the prerequisites.

2. Generalized Briot-Bouquet Differential Subordination

We present here all implication results pertaining to the proposed generalized Briot-Bouquet differential subordination. We begin with the following result:

Theorem 2.1. *Let h be convex in $\mathbb{D}$ and $\alpha, \beta \in \mathbb{C}$ with $\beta \neq 0$. If $Q \in \mathcal{H}[1, n]$ be such that the following conditions hold:*

$$(i)\ \operatorname{Re}\left(\frac{1}{\beta h(z)+\alpha}\right) > 0 \quad (z \in \mathbb{D}).$$

$$(ii)\ \operatorname{Re}\left(\frac{1}{\beta h(\zeta)+\alpha} + (Q(z)-1)\frac{h(\zeta)}{\zeta h'(\zeta)}\right) > 0 \quad (z \in \mathbb{D},\ \ \zeta \in h^{-1}(p(D))),$$

where $D = \{z \in \mathbb{D} : p(z) = h(\zeta)$ for some $\zeta \in \partial\mathbb{D}\}$. If p is analytic in $\mathbb{D}$ with $p(0) = h(0)$ and

$$p(z)Q(z) + \frac{zp'(z)}{\beta p(z)+\alpha} \prec h(z), \tag{2.1}$$

then $p(z) \prec h(z)$.

Proof. Let us suppose p is not subordinate to h. Then by [16, Lemma 2.2d, pp.24] there exists $z_0 \in \mathbb{D}$, $\zeta_0 \in \partial\mathbb{D}$ and $m \geq 1$ such that $p(z_0) = h(\zeta_0)$ and $z_0 p'(z_0) = m\zeta_0 h'(\zeta_0)$ and therefore we have

$$\psi_0 := \psi(p(z_0), z_0p'(z_0)) = \psi(h(\zeta_0), m\zeta_0h'(\zeta_0)) = h(\zeta_0)Q(z_0) + \frac{m\zeta_0h'(\zeta_0)}{\beta h(\zeta_0)+\alpha},$$

which yields

$$\operatorname{Re}\frac{\psi_0 - h(\zeta_0)}{\zeta_0 h'(\zeta_0)} = \operatorname{Re}\left((Q(z_0)-1)\frac{h(\zeta_0)}{\zeta_0h'(\zeta_0)} + \frac{m}{\beta h(\zeta_0)+\alpha}\right).$$

Using the fact that $m \geq 1$ together with (i) and (ii), we have

$$\operatorname{Re}\frac{\psi_0 - h(\zeta_0)}{\zeta_0 h'(\zeta_0)} \ \geq\ \operatorname{Re}\left((Q(z_0)-1)\frac{h(\zeta_0)}{\zeta_0h'(\zeta_0)} + \frac{1}{\beta h(\zeta_0)+\alpha}\right) > 0,$$

which implies

$$\left|\arg\frac{\psi_0 - h(\zeta_0)}{\zeta_0 h'(\zeta_0)}\right| < \frac{\pi}{2}.$$

Since $h(\mathbb{D})$ is convex, $h(\zeta_0) \in h(\partial\mathbb{D})$ and $\zeta_0 h'(\zeta_0)$ is the outward normal to $h(\partial\mathbb{D})$ at $h(\zeta_0)$, we conclude that $\psi_0 \notin h(\mathbb{D})$, which contradicts (2.1) and hence $p(z) \prec h(z)$. □

Remark 2.2. If we take $Q(z) = 1$ in Theorem 2.1, it reduces to [16, Theorem 3.2a].

Corollary 2.3. *Let h be convex in $\mathbb{D}$ and $\alpha, \beta \in \mathbb{C}$ with $\beta \neq 0$. If $Q \in \mathcal{H}[1, n]$ and p is analytic in $\mathbb{D}$ with $p(0) = h(0) = (k-1)/4$, where $k \geq 1$, be such that*

$$\operatorname{Re}\left(\frac{1}{\beta h(\zeta) + \alpha}\right) > k|Q(z) - 1| - \operatorname{Re}(Q(z) - 1) \quad (z \in \mathbb{D},\ \zeta \in \partial\mathbb{D}), \tag{2.2}$$

then

$$p(z)Q(z) + \frac{zp'(z)}{\beta p(z) + \alpha} \prec h(z) \quad \Rightarrow \quad p(z) \prec h(z).$$

Proof. Since $k \geq 1$, from (2.2) it is clear that for $\zeta \in \partial\mathbb{D}$,

$$\operatorname{Re}\left(\frac{1}{\beta h(\zeta) + \alpha}\right) > 0. \tag{2.3}$$

Since h is convex, the above inequality holds on $\mathbb{D}$ as well. Also, we can say that $\tilde{h}(z) := h(z) - h(0) \in \mathcal{C}$. Using Marx Strohhäcker theorem [16], we have $\operatorname{Re}(\zeta\tilde{h}'(\zeta)/\tilde{h}(\zeta)) > 1/2$, which is equivalent to

$$\left|\frac{\tilde{h}(\zeta)}{\zeta\tilde{h}'(\zeta)} - 1\right| \leq 1,$$

which implies

$$\left|\frac{h(\zeta)}{\zeta h'(\zeta)} - 1\right| \leq 1 + \frac{|h(0)|}{|h'(\zeta)|}. \tag{2.4}$$

Since $\tilde{h} \in \mathcal{C}$, we have $|\tilde{h}'(z)| \geq 1/(1+r)^2$ on $|z| = r$ [10, Theorem.9, pp 118]. We know that $\zeta \in \partial\mathbb{D}$, so we have $|h'(\zeta)| = |\tilde{h}'(\zeta)| \geq 1/4$. Thus (2.4) reduces to

$$\left|\frac{h(\zeta)}{\zeta h'(\zeta)} - 1\right| \leq k.$$

Note that if $X,\ Y \in \mathbb{C}$ and $|X - 1| \leq K$, then

$$\operatorname{Re}(X.Y) = \operatorname{Re} Y + \operatorname{Re} Y(X-1)) \geq \operatorname{Re} Y - |Y|K.$$

Using this inequality, we can say that

$$\begin{aligned}\operatorname{Re}\left((Q(z)-1)\frac{h(\zeta)}{\zeta h'(\zeta)} + \frac{1}{\beta h(\zeta) + \alpha}\right) \ &\geq \ \operatorname{Re}(Q(z)-1) - k|Q(z)-1| \\ &\quad + \operatorname{Re}\left(\frac{1}{\beta h(\zeta) + \alpha}\right),\end{aligned}$$

which by using (2.2) implies

$$\operatorname{Re}\left((Q(z)-1)\frac{h(\zeta)}{\zeta h'(\zeta)} + \frac{1}{\beta h(\zeta) + \alpha}\right) > 0. \tag{2.5}$$

From (2.3) and (2.5), we may conclude that the conditions (i) and (ii) of Theorem 2.1 are satisfied and as its application, the result follows. □

Corollary 2.4. *Let $Q \in \mathcal{H}[1, 1]$ be a function such that $|Q(z)| \leq M$ $(z \in \mathbb{D})$ for some $M > 0$ and $\alpha,\ \beta \in \mathbb{C}$ with $\beta \neq 0$. Suppose p is analytic and h is convex in $\mathbb{D}$ with $p(0) = h(0) = 1$ such that*

$$\operatorname{Re}\left(\frac{1}{\beta h(\zeta) + \alpha}\right) > 6(M+1) \quad (\zeta \in \partial\mathbb{D}), \tag{2.6}$$

then

$$p(z)Q(z)+\frac{zp'(z)}{\beta p(z)+\alpha}\prec h(z) \quad \Rightarrow \quad p(z)\prec h(z).$$

Proof. We know that $-\operatorname{Re}(Q(z)-1)\leq |Q(z)-1|$ and since $p(0)=1$, in view of Corollary 2.3, we have $k=5$. Thus

$$k|Q(z)-1|-\operatorname{Re}(Q(z)-1)\leq 6|Q(z)-1|\leq 6(|Q(z)|+1)\leq 6(M+1). \tag{2.7}$$

Since (2.2) holds due to (2.6) and (2.7) and therefore the result follows from Corollary 2.3. □

Corollary 2.5. *Let h be convex in $\mathbb{D}$ with $h(0)=1$ and $\alpha,\beta\in\mathbb{C}$ with $\beta\neq 0$. Let $g\in\mathcal{A}$ be defined as*

$$g(z)=z\exp\int_0^z\frac{Q(t)-1}{t}dt, \tag{2.8}$$

such that h and Q satisfy the conditions (i) and (ii) of Theorem 2.1. If $f\in\mathcal{A}$ and F is given by

$$F(z)=I[f,g]=\left(\frac{\alpha+\beta}{g^{\alpha}(z)}\int_0^z g'(t)g^{\alpha-1}(t)f^{\beta}(t)dt\right)^{1/\beta}, \tag{2.9}$$

then

$$\frac{zf'(z)}{f(z)}\prec h(z) \quad \Rightarrow \quad \frac{zF'(z)/F(z)}{zg'(z)/g(z)}\prec h(z).$$

Proof. From (2.8), we have $Q(z)=zg'(z)/g(z)$ and let us suppose $p(z)=zF'(z)/(Q(z)F(z))$. Then by differentiating (2.9) and appropriately replacing the expressions, we have

$$p(z)Q(z)+\frac{zp'(z)}{\beta p(z)+\alpha}=\frac{zf'(z)}{f(z)}.$$

Since $zf'(z)/f(z)\prec h(z)$, the result now follows from Theorem 2.1. □

If we take $h(z)=((1+z)/(1-z))^{\gamma}$ with $\gamma\in(0,1]$ in Corollary 2.5, we obtain the following result.

Corollary 2.6. *Let $f\in\mathcal{A}$ and g,Q and F are as defined in* (2.8) *and* (2.9) *respectively such that* $\operatorname{Re}(Q(z)-1)>1-\gamma$. *Then*

$$f\in\mathcal{SS}^*(\gamma)\ \Rightarrow\ F \text{ is almost strongly starlike of order } \gamma \text{ w.r.t the function } g.$$

If we take $h(z)=(1+z)/(1-z)$ in Corollary 2.5, we obtain the following result.

Corollary 2.7. *Let $f\in\mathcal{A}$ and g,Q and F are as defined in* (2.8) *and* (2.9) *respectively such that $f\in\mathcal{S}^*$ and g is univalent on $\overline{\mathbb{D}}$, then F is a g-starlike function.*

Now we list some of the special cases of Theorem 2.1 here below:

Corollary 2.8. *Let $Q\in\mathcal{H}[1,1]$ and α,β be non-negative real numbers with $\beta\neq 0$ such that $|Q(z)-1|<1/(\beta e+\alpha)$ on $\mathbb{D}$. Suppose p is analytic with $p(0)=1$, then*

$$p(z)Q(z)+\frac{zp'(z)}{\beta p(z)+\alpha}\prec e^z \quad \Rightarrow \quad p(z)\prec e^z.$$

Proof. If we take $h(z) = e^z$, then for $\alpha \geq 0$ and $\beta > 0$, we have

$$\operatorname{Re}\left(\frac{1}{\beta e^z + \alpha}\right) \geq \frac{1}{\beta e + \alpha} > 0. \tag{2.10}$$

Further, we observe for $z \in \mathbb{D}$ and $\zeta \in \partial\mathbb{D}$,

$$\operatorname{Re}\left(\frac{Q(z)-1}{\zeta} + \frac{1}{\beta e^z + \alpha}\right) \geq -|Q(z)-1| + \frac{1}{\beta e + \alpha} > 0. \tag{2.11}$$

From (2.10) and (2.11), we may conclude that both the conditions of Theorem 2.1 are satisfied and thus the result now follows from Theorem 2.1. □

Corollary 2.9. *Let $Q \in \mathcal{H}[1,1]$ and $\alpha \geq 0, \beta > 0$ be such that*

$$|Q(z)-1| < \operatorname{Re}(Q(z)-1) + \frac{1}{2(\sqrt{2}\beta + \alpha)}. \tag{2.12}$$

Suppose p is analytic with $p(0) = 1$, then

$$p(z)Q(z) + \frac{zp'(z)}{\beta p(z) + \alpha} \prec \sqrt{1+z} \quad \Rightarrow \quad p(z) \prec \sqrt{1+z}.$$

Proof. Let $h(z) = \sqrt{1+z}$, then condition (i) of Theorem 2.1 is satisfied clearly as for $\alpha \geq 0$ and $\beta > 0$,

$$\operatorname{Re}\left(\frac{1}{\beta\sqrt{1+z} + \alpha}\right) \geq \frac{1}{\beta\sqrt{2} + \alpha} > 0, \quad z \in \mathbb{D}.$$

Now for $z \in \mathbb{D}$ and $\zeta \in \partial\mathbb{D}$, we have from (2.12)

$$\begin{aligned} \operatorname{Re}\left(2(Q(z)-1)\left(1 + \frac{1}{\zeta}\right) + \frac{1}{\beta\sqrt{1+z} + \alpha}\right) &\geq 2\operatorname{Re}(Q(z)-1) - 2|Q(z)-1| \\ &\quad + \frac{1}{\sqrt{2}\beta + \alpha} \\ &> 0, \end{aligned}$$

which implies that condition (ii) of Theorem 2.1 is satisfied. Thus the result follows at once from Theorem 2.1. □

Corollary 2.10. *Let $Q \in \mathcal{H}[1,1]$ be such that $Q'(0) > 0$ and α, β be non-negative real numbers with $\beta \neq 0$. Suppose p is analytic with $p(0) = 1$ and $p'(0) > 0$, then for $0 < \gamma \leq 1$,*

$$p(z)Q(z) + \frac{zp'(z)}{\beta p(z) + \alpha} \prec \left(\frac{1+z}{1-z}\right)^{\gamma} \quad \Rightarrow \quad p(z) \prec \left(\frac{1+z}{1-z}\right)^{\gamma}.$$

Proof. Let $h(z) = ((1+z)/(1-z))^{\gamma}$. Then condition (i) of Theorem 2.1 clearly holds. For condition (ii) to hold, we need to show that

$$\operatorname{Re}\left((Q(z)-1)\frac{1-\zeta^2}{2\gamma\zeta} + \frac{1}{\beta\left(\frac{1+\zeta}{1-\zeta}\right)^{\gamma} + \alpha}\right) > 0 \quad (z \in \mathbb{D},\ \zeta \in \partial\mathbb{D}).$$

Let $Q(z) = 1 + a_1 z + a_2 z^2 + \cdots$ and define $R(z) = Q(z) - 1$, then $R(0) = 0$. Since $Q'(0) > 0$, it is typically real and it is easy to conclude that $R(z)$ is typically real. We know that $\zeta \in h^{-1}(p(D))$, where $D := \{z \in \mathbb{D} : p(z) = h(\zeta) \text{ for some } \zeta \in \mathbb{D}\}$. Since

$h(z) = ((1+z)/(1-z))^\gamma$ is typically real and $p'(0) > 0$, we have $sign(\operatorname{Im} z) = sign(\operatorname{Im}\zeta)$. Now we consider

$$\begin{aligned}\operatorname{Re}\left((Q(z)-1)\frac{1-\zeta^2}{2\gamma\zeta}\right) &= \frac{1}{2\gamma}\left(\operatorname{Re}(Q(z)-1)\operatorname{Re}\left(\frac{1-\zeta^2}{\zeta}\right) - \operatorname{Im}(Q(z)-1)\operatorname{Im}\left(\frac{1-\zeta^2}{\zeta}\right)\right)\\ &= -\frac{1}{2\gamma}\operatorname{Im}(Q(z)-1)\operatorname{Im}\left(\frac{1-\zeta^2}{\zeta}\right).\end{aligned}$$

Taking $\zeta = e^{i\theta}$ $(0 \le \theta \le 2\pi)$, we have

$$-\operatorname{Im}\left(\frac{1-\zeta^2}{\zeta}\right) = 2\sin\theta \begin{cases} > 0, & \theta \in (0,\pi),\\ < 0, & \theta \in (\pi, 2\pi).\end{cases}$$

Since $Q(z)-1$ is typically real, $sign(\operatorname{Im}(Q(z)-1)) = sign(\operatorname{Im} z)$. Thus we have

$$\operatorname{Re}\left((Q(z)-1)\frac{1-\zeta^2}{2\gamma\zeta}\right) = 2\sin\theta\operatorname{Im}(Q(z)-1) \ge 0.$$

Also for $\alpha \ge 0$ and $\beta > 0$, we have $\operatorname{Re}\left(\beta\left(\frac{1+z}{1-z}\right)^\gamma + \alpha\right) > 0$ and therefore

$$\operatorname{Re}\left((Q(z)-1)\frac{1-\zeta^2}{2\gamma\zeta} + \frac{1}{\beta\left(\frac{1+\zeta}{1-\zeta}\right)^\gamma + \alpha}\right) > 0.$$

Therefore the result follows at once from Theorem 2.1. □

Now we use a different technique to prove the next two results, which demonstrates the similar implication for Janowski functions.

Theorem 2.11. *Let $p(z)$ be analytic in $\mathbb{D}$ with $p(0) = 1$ and $Q \in \mathcal{H}[1,n]$ be a function such that $|Q(z)| < M$ for some $M > 0$. Let $-1 \le B < A < 1$ and $-1 < E < D \le 1$ satisfy*

$$\begin{aligned}(A-B)(1-A)(1+E) &> (1+|A|)(\beta+\alpha+|\beta A+\alpha B|)((1+D)(1-B)\\ &\quad +M(1+E)(1-A))\end{aligned} \tag{2.13}$$

for some α and β, where $\alpha+\beta > 0$. If

$$p(z)Q(z) + \frac{zp'(z)}{\beta p(z)+\alpha} \prec \frac{1+Dz}{1+Ez},$$

then $p(z) \prec (1+Az)/(1+Bz)$.

Proof. Let us define $P(z)$ and $\omega(z)$ as

$$P(z) := p(z)Q(z) + \frac{zp'(z)}{\beta p(z)+\alpha}, \qquad \omega(z) := \frac{p(z)-1}{A-Bp(z)}.$$

Then $\omega(z)$ is meromorphic in $\mathbb{D}$ and $\omega(0) = 0$. By the definition of $P(z)$ and $\omega(z)$, we have

$$P(z) = \frac{1+A\omega(z)}{1+B\omega(z)}Q(z) + \frac{(A-B)z\omega'(z)}{(1+B\omega(z))[\beta(1+A\omega(z)) + \alpha(1+B\omega(z))]}.$$

Now we need to show that $|\omega(z)|< 1$ in $\mathbb{D}$. On the contrary, let us assume that there exists a point $z_0 \in \mathbb{D}$ such that

$$\max_{|z|\leq|z_0|}|\omega(z)|= |\omega(z_0)|= 1.$$

Then by [21, Lemma 1.3, pp.28], there exists $k \geq 1$ such that $z_0\omega'(z_0) = k\omega(z_0)$. Now by taking $\omega(z_0) = e^{i\theta}$ $(0 \leq \theta \leq 2\pi)$, we have

$$\begin{aligned} |P(z_0)| &= \left|\frac{1+A\omega(z_0)}{1+B\omega(z_0)}Q(z_0) + \frac{(A-B)k\omega(z_0)}{(1+B\omega(z_0))(\beta(1+A\omega(z_0))+\alpha(1+B\omega(z_0)))}\right| \\ &\geq \left|\frac{1+Ae^{i\theta}}{1+Be^{i\theta}}\right|\left(\left|\frac{(A-B)k}{(1+Ae^{i\theta})(\beta(1+Ae^{i\theta})+\alpha(1+Be^{i\theta}))}\right| - |Q(z_0)|\right). \end{aligned}$$

We know that for $p > 0$, $|p + qe^{i\theta}|^2= p^2 + q^2 + 2pq\cos\theta$, attains its maximum at $\theta = 0$ if $q > 0$ and at $\theta = \pi$ if $q \leq 0$. So, $\max_{0\leq\theta\leq2\pi}|p + qe^{i\theta}|^2= (p + |q|)^2$. Thus

$$|P(z_0)|\geq \frac{1-A}{1-B}\left(\frac{(A-B)k}{(1+|A|)(\beta+\alpha+|\beta A+\alpha B|)} - M\right).$$

Clearly the expression on the right hand side is an increasing function of k and attains its minimum at $k = 1$. So

$$|P(z_0)|\geq \frac{1-A}{1-B}\left(\frac{(A-B)}{(1+|A|)(\beta+\alpha+|\beta A+\alpha B|)} - M\right).$$

Now from (2.13), we have

$$|P(z_0)|> \frac{1+D}{1+E},$$

which contradicts the fact that $P(z) \prec (1+Dz)/(1+Ez)$ and hence completes the proof. □

Using Theorem 2.11, we obtain the following corollary.

Corollary 2.12. *Let $p(z)$ be analytic in $\mathbb{D}$ with $p(0) = 1$ and Q be a function such that*

$$z\exp\int_0^z \frac{Q(t)-1}{t}dt \in \mathcal{S}^*(\phi).$$

Let $-1 \leq B < A < 1$, $-1 < E < D \leq 1$ and α, β be such that $\alpha + \beta > 0$. Then

$$p(z)Q(z) + \frac{zp'(z)}{\beta p(z)+\alpha} \prec \frac{1+Dz}{1+Ez} \Rightarrow p(z) \prec \frac{1+Az}{1+Bz},$$

whenever any of the following cases hold:

- *(i)* $\phi(z) = e^z$ *and* $(A-B)(1-A)(1+E) > (1+|A|)(\beta+\alpha+|\beta A+\alpha B|)((1+D)(1-B)+e(1+E)(1-A))$
- *(ii)* $\phi(z) = \sqrt{1+z}$ *and* $(A-B)(1-A)(1+E) > (1+|A|)(\beta+\alpha+|\beta A+\alpha B|)((1+D)(1-B)+\sqrt{2}(1+E)(1-A))$
- *(iii)* $\phi(z) = 2/(1+e^{-z})$ *and* $(A-B)(1-A)(1+E)(1+e) > (1+|A|)(\beta+\alpha+|\beta A+\alpha B|)((1+D)(1-B)(1+e)+2e(1+E)(1-A))$
- *(iv)* $\phi(z) = 1+ze^z$ *and* $(A-B)(1-A)(1+E) > (1+|A|)(\beta+\alpha+|\beta A+\alpha B|)((1+D)(1-B)+(1+e)(1+E)(1-A))$
- *(v)* $\phi(z) = z+\sqrt{1+z^2}$ *and* $(A-B)(1-A)(1+E) > (1+|A|)(\beta+\alpha+|\beta A+\alpha B|)((1+D)(1-B)+(1+\sqrt{2})(1+E)(1-A))$.

Since the proof of Theorem 2.11 loses its validity when $A = 1$ or $E = -1$, the theorem does not reduce to the case when $P(z)$ and $p(z)$, both are subordinate to $(1 + az)/(1 - z)$ for $0 \leq a \leq 1$. This case, we handle in the following result with a different approach.

Theorem 2.13. *Let $\alpha \geq 0,\ \beta > 0,\ 0 \leq a \leq 1$. Assume $p(z)$ to be analytic in $\mathbb{D}$ with $p(0) = 1$ and $p'(0) > 0$. Also, let $Q(z) \in \mathcal{H}[1, 1]$ be such that $Q'(0) > 0$ and $\operatorname{Re} Q(z) < 1$. If p satisfies*

$$p(z)Q(z) + \frac{zp'(z)}{\beta p(z) + \alpha} \prec \frac{1 + az}{1 - z}, \tag{2.14}$$

then $p(z) \prec (1 + az)/(1 - z)$.

Proof. We need to show that $p(z) \prec q(z) = (1 + az)/(1 - z)$. For if $p \not\prec q$, then using [16, Lemma 2.2d, pp.24], there exists a point $z_0 = r_0 e^{i\theta_0} \in \mathbb{D}$, $\zeta_0 \in \partial\mathbb{D}\backslash\{1\}$ and $m \geq 1$ such that $p(z_0) = q(\zeta_0)$, $z_0 p'(z_0) = m\zeta_0 q'(\zeta_0)$ and $p(\mathbb{D}_{r_0}) \subset q(\mathbb{D})$. Since ζ_0 is a boundary point, we may assume that $\zeta_0 = e^{i\theta}$ for $\theta \in [0, 2\pi]$. Then

$$\begin{aligned} p(z_0)Q(z_0) + \frac{z_0 p'(z_0)}{\beta p(z_0) + \alpha} &= q(\zeta_0)Q(z_0) + \frac{m\zeta_0 q'(\zeta_0)}{\beta q(\zeta_0) + \alpha} \\ &= \left(\frac{1 - a}{2} + i\frac{1 + a}{2}\cot\theta/2\right) Q(z_0) \\ &\quad - \frac{m(a + 1)e^{i\theta}}{(1 - e^{i\theta})((\beta + \alpha) + (\beta a - \alpha)e^{i\theta})}. \end{aligned}$$

Taking $Q(z_0) = u + iv$, we have

$$\begin{aligned} \operatorname{Re}\left(p(z_0)Q(z_0) + \frac{z_0 p'(z_0)}{\beta p(z_0) + \alpha}\right) &= \frac{(1 - a)u}{2} - \frac{(1 + a)\cot(\theta/2)v}{2} \\ &\quad - \frac{m(1 + a)(\beta(1 - a) + 2\alpha)(1 - \cos\theta)}{(\beta(1 - a) + 2\alpha)^2(1 - \cos\theta)^2 + (\beta(a + 1)\sin\theta)^2}. \end{aligned}$$

Since $Q'(0) > 0$, it is clear that $Q(z)$ is typically real and thus

$$sign(v) = sign(\operatorname{Im}(Q(z_0))) = sign(\operatorname{Im}(z_0)). \tag{2.15}$$

We also have $p'(0) > 0$ and $q(z) = (1 + az)/(1 - z)$ is typically real, which implies

$$sign(\operatorname{Im}(z_0)) = sign((\operatorname{Im}(p(z_0)))) = sign((\operatorname{Im}(q(\zeta_0)))) = sign(\operatorname{Im}(\zeta_0)) = sign(\sin\theta). \tag{2.16}$$

From (2.15) and (2.16), we obtain $sign(v) = sign(\sin\theta)$, which is sufficient to conclude that $v\cot\theta/2 > 0$ for $\theta \in [0, 2\pi]$. Also we know that $\alpha \geq 0,\ \beta > 0,\ m \geq 1$ and $0 \leq a \leq 1$. Therefore

$$\operatorname{Re}\left(p(z_0)Q(z_0) + \frac{z_0 p'(z_0)}{\beta p(z_0) + \alpha}\right) < \frac{(1 - a)u}{2} < \frac{1 - a}{2},$$

which contradicts (2.14) and hence the result. □

3. Other Subordination Results and Applications

We begin with the following analogue of open door lemma:

Theorem 3.1. *Let $p(z)$ be analytic in $\mathbb{D}$ with $p(0) = 1$ and $Q(z) \in \mathcal{P}$. Suppose that $\alpha \geq 0, \beta > 0$ and p satisfies*

$$p(z)Q(z) + \frac{zp'(z)}{\beta p(z) + \alpha} = 1, \tag{3.1}$$

then $\operatorname{Re} p(z) > 0$.

Proof. Define the analytic function g by

$$g(z) = z \exp \int_0^z \frac{Q(t) - 1}{t} dt.$$

Then one can verify that the function $p(z)$, given by

$$p(z) = \frac{g^\alpha(z) z^\beta}{\beta} \left(\int_0^z g^{\alpha-1}(t) g'(t) t^\beta \right)^{-1} - \frac{\alpha}{\beta} \tag{3.2}$$

is analytic in $\mathbb{D}$ and $p \in \mathcal{H}[1, n]$. The logarithmic differentiation of (3.2) reveals that p is a solution of the differential equation (3.1). We now use [16, Theorem 2.3i, pp.35] to prove that $\operatorname{Re} p(z) > 0$. Let $\Omega = \{1\}$ and $\psi(r, s; z) = rQ(z) + s/(\beta r + \alpha)$, then (3.1) can be written as

$$\{\psi(p(z), zp'(z); z) | z \in \mathbb{D}\} \subset \Omega.$$

In view of [16, Theorem 2.3i, pp.35], it is sufficient to show that $\psi \in \Psi_n[\Omega, 1]$, which means admissibility conditions defined for the class $\Psi_n[\Omega, 1]$ are satisfied by ψ or equivalently:

$$\psi(\rho i, \sigma; z) = \frac{\sigma}{\beta \rho i + \alpha} + Q(z)\rho i \neq 1, \tag{3.3}$$

where $\rho \in \mathbb{R},\ \sigma \leq -\frac{n}{2}(1+\rho^2),\ z \in \mathbb{D}$ and $n \geq 1$. Suppose on the contrary if we assume (3.3) is false, then there exists some values $\rho_0,\ \sigma_0$ and z_0 such that

$$\frac{\sigma_0}{\beta \rho_0 i + \alpha} + i\rho_0 Q(z_0) = 1. \tag{3.4}$$

If we let $Q(z) = S(z) + iT(z)$, then (3.4) yields

$$\frac{\sigma_0 \alpha}{\alpha^2 + \beta^2 \rho_0^2} - \rho_0 T(z_0) = 1 \quad \text{and} \quad -\frac{\beta \sigma_0 \rho_0}{\alpha^2 + \beta^2 \rho_0^2} + \rho_0 S(z_0) = 0. \tag{3.5}$$

Since $\sigma_0 < 0$ and $\alpha \geq 0$, we have $\rho_0 \neq 0$. Therefore from (3.5), we deduce that

$$\operatorname{Re} Q(z_0) = S(z_0) = \frac{\beta \sigma_0}{\alpha^2 + \beta^2 \rho_0^2} \leq -\frac{\beta n (1 + \rho_0^2)}{2(\alpha^2 + \beta^2 \rho_0^2)} < 0,$$

which contradicts the hypothesis and hence it follows that $\operatorname{Re} p(z) > 0$. □

Corollary 3.2. *Let $f \in \mathcal{A}$ be such that $f(z)/z \neq 0$ in $\mathbb{D}$ and $f \in \mathcal{R}$. Then* $\operatorname{Re} f(z)/z > 0$.

Proof. Let $Q(z) = f'(z)$ then $Q \in \mathcal{P}$ as $f \in \mathcal{R}$. Choose $p(z) = z/f(z), \beta = 1$ and $\alpha = 0$, then clearly $p(0) = 1$ and $p(z)$ satisfies (3.1). Now by an application of Theorem 3.1, it follows that $\operatorname{Re} p(z) > 0$ and hence the result. □

Here below we consider an analogue of integral existence theorem:

Theorem 3.3. *Let $\varphi, \phi \in \mathcal{H}[1,n]$, with $\varphi(z)\phi(z) \neq 0$ in $\mathbb{D}$. Let λ, η, γ and δ be complex numbers with $\eta \neq 0$, $\lambda + \delta = \eta + \gamma = 1$ and α, β be non negative real numbers with $\beta \neq 0$. Let $g \in \mathcal{A}_n$ and suppose that*

$$Q(z) = \lambda \frac{zg'(z)}{g(z)} + \frac{z\varphi'(z)}{\varphi(z)} + \delta \prec \frac{1+z}{1-z}. \tag{3.6}$$

If F is defined by

$$F(z) = \left(\frac{\alpha+\beta}{\phi^\beta(z) g^{\lambda\alpha}(z) \varphi^\alpha(z) z^{\delta\alpha+\beta\gamma}} \int_0^z g^{\lambda\alpha}(t)\varphi^\alpha(t) t^{\beta+\delta\alpha-1} Q(t)dt \right)^{\frac{1}{\eta\beta}}, \tag{3.7}$$

then $F \in \mathcal{A}_n, F(z)/z \neq 0$ and for $z \in \mathbb{D}$,

$$\operatorname{Re}\left(\frac{\eta \frac{zF'(z)}{F(z)} + \frac{z\phi'(z)}{\phi(z)} + \gamma}{\lambda \frac{zg'(z)}{g(z)} + \frac{z\varphi'(z)}{\varphi(z)} + \delta} \right) > 0. \tag{3.8}$$

Proof. Let $p(z)$ be given by

$$p(z) = \frac{1}{\beta} g^{\lambda\alpha}(z)\varphi^\alpha(z) z^{\beta+\delta\alpha} \left(\int_0^z g^{\lambda\alpha}(t)\varphi^\alpha(t) t^{\beta+\delta\alpha-1} Q(t)dt \right)^{-1} - \frac{\alpha}{\beta}. \tag{3.9}$$

Then

$$p(z) = 1 + p_n z^n + p_{n+1} z^{n+1} + \cdots$$

is analytic in $\mathbb{D}$ and $p \in \mathcal{H}[1,n]$. The logarithmic differentiation of (3.9) shows that $p(z)$ satisfies (3.1) with $Q(z)$ as given in (3.6). From hypothesis, we have $Q(z) \prec (1+z)/(1-z)$. Thus the hypothesis of the Theorem. 3.1 is fulfilled by p as well as Q and hence it follows that $\operatorname{Re} p(z) > 0$. Using (3.7) and (3.9), we get

$$F(z) = \left(\frac{(\alpha+\beta) z^{\beta(1-\gamma)}}{\phi^\beta(z)(\beta p(z) + \alpha)} \right)^{\frac{1}{\eta\beta}} = z \left(\frac{\alpha+\beta}{\phi^\beta(z)(\beta p(z)+\alpha)} \right)^{\frac{1}{\eta\beta}}. \tag{3.10}$$

Clearly the expression in the bracket is analytic and non-zero, so we deduce that $F \in \mathcal{A}_n$ and $F(z)/z \neq 0$. By differentiating (3.10) logarithmically, we obtain

$$\eta \frac{zF'(z)}{F(z)} = 1 - \gamma - \frac{z\phi'(z)}{\phi(z)} - \frac{zp'(z)}{\beta p(z) + \gamma},$$

which by using (3.1) simplifies to

$$\eta \frac{zF'(z)}{F(z)} + \frac{z\phi'(z)}{\phi(z)} + \gamma = p(z)Q(z).$$

Therefore

$$\operatorname{Re} \frac{1}{Q(z)} \left(\eta \frac{zF'(z)}{F(z)} + \frac{z\phi'(z)}{\phi(z)} + \gamma \right) > 0,$$

which is equivalent to (3.8) and hence completes the proof. □

We now obtain the following special case of the above theorem when $\delta = \gamma = 0$, $\lambda = \eta = 1$ and $\varphi(z) = 1 = \phi(z)$.

Corollary 3.4. *Let $g \in \mathcal{S}^*$, $\alpha \geq 0, \beta > 0$ and F is defined as*

$$F(z) = \frac{\alpha+\beta}{g^\alpha(z)} \int_0^z g'(t) g^{\alpha-1}(t) t^\beta dt,$$

then F is almost strongly starlike with respect to g and hence univalent.

Theorem 3.5. *Let $f, g \in \mathcal{A}_n$ and $\phi \in \mathcal{H}[1,n]$ with $\phi(0) = 1$ and $\phi(z) \neq 0$ in $\mathbb{D}$. Let β, α and σ be complex numbers such that $\beta \neq 0$ and $\mathrm{Re}(\beta h(z) + \alpha) > 0$, where h is a convex function with $h(0) = 1$ in $\mathbb{D}$. Suppose*

$$\beta \frac{zf'(z)}{f(z)} + \sigma \frac{zg'(z)}{g(z)} \prec \beta h(z) + \sigma. \tag{3.11}$$

If F is defined by

$$F(z) = \left(\frac{\beta+\alpha}{z^\alpha \phi(z)} \int_0^z f^\beta(t) g^\sigma(t) t^{\alpha-\sigma-1} dt \right)^{\frac{1}{\beta}}, \tag{3.12}$$

then $F(z) \in \mathcal{A}_n$ and

$$\frac{zF'(z)}{F(z)} + \frac{1}{\beta} \frac{z\phi'(z)}{\phi(z)} \prec h(z). \tag{3.13}$$

Proof. Let $p(z)$ be given by

$$p(z) = \frac{1}{\beta} f^\beta(z) g^\sigma(z) z^{\alpha-\sigma} \left(\int_0^z f^\beta(t) g^\sigma(t) t^{\alpha-\sigma-1} dt \right)^{-1} - \frac{\alpha}{\beta}. \tag{3.14}$$

Then $p(z)$ is analytic in $\mathbb{D}$ and $p(0) = 1$. Upon logarithmic differentiation of (3.14), we deduce that $p(z)$ satisfies

$$p(z) + \frac{zp'(z)}{\beta p(z) + \alpha} = \frac{zf'(z)}{f(z)} + \frac{\sigma}{\beta} \left(\frac{zg'(z)}{g(z)} - 1 \right). \tag{3.15}$$

From (3.11) and (3.15), it can be concluded that

$$p(z) + \frac{zp'(z)}{\beta p(z) + \alpha} \prec h(z).$$

By applying [16, Theorem 3.2a, pp.81], we obtain $p(z) \prec h(z)$. Substituting (3.12) in (3.14), we obtain

$$p(z) = \frac{1}{\beta} \left(\frac{(\alpha+\beta) f^\beta(z) g^\sigma(z)}{F^\beta(z) \phi(z) z^\sigma} - \alpha \right).$$

Differentiating logarithmically the following equation

$$\beta p(z) + \alpha = \frac{(\alpha+\beta) f^\beta(z) g^\sigma(z)}{F^\beta(z) \phi(z) z^\sigma}$$

and using (3.15), we obtain

$$p(z) = \frac{zF'(z)}{F(z)} + \frac{1}{\beta} \frac{z\phi'(z)}{\phi(z)}$$

and thus (3.13) follows. □

We now derive another result analogous to open door lemma as follows:

Lemma 3.6. *Let n be a positive integer and α, β be non-negative real numbers with $\beta \neq 0$. Let $Q \in \mathcal{H}[1,n]$ satisfy*

$$Q(z) \prec 1 + z + \frac{nz}{\beta + \alpha(1+z)} \equiv h(z). \tag{3.16}$$

If $p \in \mathcal{H}[1,n]$ satisfies the differential equation

$$p(z)Q(z) + \frac{zp'(z)}{\beta p(z) + \alpha} = 1, \tag{3.17}$$

then $p(z) \prec 1/(1+z)$.

Proof. Let us set $q(z) = 1/(1+z)$, then

$$h(z) = \frac{1}{q(z)} - \frac{nzq'(z)}{q(z)(\beta q(z) + \alpha)}.$$

If $g(z) = z/(\beta + \alpha(1+z))$, then we have

$$\operatorname{Re}\frac{zg'(z)}{g(z)} = 1 - \operatorname{Re}\frac{\alpha z}{\beta + \alpha(1+z)} > 1 - \frac{\alpha}{\beta + 2\alpha} = \frac{\beta + \alpha}{\beta + 2\alpha} > 0.$$

Since g is starlike and

$$\begin{aligned}
\operatorname{Re}\frac{zh'(z)}{g(z)} &= \operatorname{Re}\left((\beta + \alpha(1+z))\left(1 + \frac{n(\beta+\alpha)}{(\beta+\alpha(1+z))^2}\right)\right) \\
&= \beta + \alpha(1 + \operatorname{Re} z) + n(\beta+\alpha)\operatorname{Re}\left(\frac{1}{\beta + \alpha(1+z)}\right) \\
&\geq \beta + \frac{n(\beta+\alpha)}{\beta + 2\alpha} > 0,
\end{aligned}$$

we deduce that h is close to convex and hence univalent in $\mathbb{D}$. Now we consider the boundary curve of h defined as

$$w = h(e^{i\theta}) = u(\theta) + iv(\theta), \quad \theta \in (-\pi, \pi).$$

Suppose

$$\begin{aligned}
r(\theta) &= |h(e^{i\theta})| = |e^{-i\theta}h(e^{i\theta})| = \left|e^{-i\theta} + 1 + \frac{n}{\beta + \alpha(1+e^{i\theta})}\right| \\
&= \left|1 + \cos\theta + \frac{n(\beta + \alpha(1+\cos\theta))}{d(\theta)} - i\left(\sin\theta + \frac{n\alpha\sin\theta}{d(\theta)}\right)\right|,
\end{aligned} \tag{3.18}$$

where $d(\theta) = (\beta + \alpha(1+\cos\theta))^2 + (\alpha\sin\theta)^2$. After simplifying (3.18), we have

$$r(\theta) = \sqrt{2(1+\cos\theta) + \frac{n^2 + 2n(2\alpha+\beta)(1+\cos\theta)}{d(\theta)}}.$$

By using (3.16) and (3.17), we deduce that

$$Q(z) = \frac{1}{p(z)} - \frac{zp'(z)}{p(z)(\beta p(z) + \alpha)} \prec h(z). \tag{3.19}$$

On the contrary, if $p \not\prec q$ then by [16, Lemma 2.2d], there exist points $z_0 \in \mathbb{D}$ and $\zeta_0 \in \partial\mathbb{D}$ and $m \geq n$, such that $p(z_0) = q(\zeta_0)$ and $z_0 p'(z_0) = m\zeta_0 q'(\zeta_0)$. From (3.19), we have

$$Q(z_0) = \frac{1}{q(\zeta_0)} - \frac{m\zeta_0 q'(\zeta_0)}{q(\zeta_0)(\beta q(\zeta_0) + \alpha)} = 1 + \zeta_0 + \frac{m\zeta_0}{\beta + \alpha(1+\zeta_0)}.$$

For $\zeta_0 = e^{i\theta}$, we have

$$|Q(z_0)| = \sqrt{2(1+\cos\theta) + \frac{m^2 + 2m(2\alpha+\beta)(1+\cos\theta)}{d(\theta)}} \geq r(\theta) \quad \theta \in (-\pi, \pi),$$

where $r(\theta)$ is given in (3.18). This implies that $Q(z_0) \notin h(\mathbb{D})$, which is a contradiction and thus $p(z) \prec 1/(1+z)$. □

Theorem 3.7. *Let n be a positive integer and α, β be non negative real numbers with $\beta \neq 0$. Let $f \in \mathcal{A}_n$ and $F = I_{\alpha,\beta}[f]$ be defined as*

$$I_{\alpha,\beta}[f] = \left(\frac{\alpha+\beta}{f^\alpha(z)} \int_0^z f^{\alpha-1}(t) f'(t) t^\beta dt\right)^{1/\beta}. \tag{3.20}$$

If

$$\frac{zf'(z)}{f(z)} \prec 1 + z + \frac{nz}{\beta + \alpha(1+z)},$$

then $|zf'(z)/f(z)| < 2|zF'(z)/F(z)|$.

Proof. Let $f \in \mathcal{A}$ satisfy (3.20) and define

$$p(z) = z^\beta f^\alpha(z) \left(\beta \int_0^z f^{\alpha-1}(t) f'(t) t^\beta dt\right)^{-1} - \frac{\alpha}{\beta}.$$

By the series expansion, it is easy to verify that p is well defined and $p \in \mathcal{H}[1, n]$. If we let $Q(z) = zf'(z)/f(z)$, then it is easy to show that p satisfies (3.17). Hence by Lemma 3.6 we deduce that $p(z) \prec 1/(1+z)$. Since $p(z) \neq 0$, we can define the analytic function $F \in \mathcal{A}_n$ by

$$F(z) = z\left(\frac{\alpha+\beta}{\beta p(z) + \alpha}\right)^{1/\beta}.$$

A simple calculation shows that this function coincides with the function given in (3.20). So we obtain

$$\frac{F(z)Q(z)}{zF'(z)} = \frac{1}{p(z)} \prec 1 + z,$$

which further implies $\left|\dfrac{zf'(z)/f(z)}{zF'(z)/F(z)} - 1\right| < 1$ and hence the result follows at once. □

Theorem 3.8. *Let n be a positive integer and α, β be non-negative real numbers with $\beta \neq 0$. Let $f \in \mathcal{A}$ satisfy*

$$\frac{f(z)}{zf'(z)} - \left(\frac{zf''(z)}{f'(z)} + 1 - \frac{zf'(z)}{f(z)}\right)\left(\beta\frac{zf'(z)}{f(z)} + \alpha\right)^{-1} \prec 1 + z + \frac{nz}{\beta + \alpha(1+z)}.$$

then $zf'(z)/f(z) \prec 1/(1+z)$.

Proof. Let $p(z) = zf'(z)/f(z)$. Then

$$Q(z) = \frac{f(z)}{zf'(z)} - \left(\frac{zf''(z)}{f'(z)} + 1 - \frac{zf'(z)}{f(z)}\right)\left(\beta\frac{zf'(z)}{f(z)} + \alpha\right)^{-1}$$

satisfies the following differential equation

$$p(z)Q(z) + \frac{zp'(z)}{\beta p(z) + \alpha} = 1.$$

Now applying Lemma 3.6, we obtain $p(z) \prec 1/(1+z)$ and that completes the proof. □

If we take $\beta = 1, \alpha = 0$ and $n = 1$ in the above theorem, we obtain the following corollary, which is a particular case of a result of Tuneski [25, Theorem 2.5].

Corollary 3.9. *Let $f \in \mathcal{A}$ satisfies*

$$\left|\frac{f(z)f''(z)}{(f'(z))^2}\right| < 2,$$

then $zf'(z)/f(z) \prec 1/(1+z)$.

Theorem 3.10. *Let n be a positive integer and α, β be non-negative real numbers with $\beta \neq 0$. Let $f \in \mathcal{A}$ satisfy*

$$\frac{f(z)}{z} + \left(\frac{zf'(z)}{f(z)} - 1\right)\left(\frac{\beta z}{f(z)} + \alpha\right)^{-1} \prec 1 + z + \frac{nz}{\beta + \alpha(1+z)},$$

then $f(z)/z \prec 1 + z$.

Proof. Taking $p(z) = z/f(z)$ and

$$Q(z) = \frac{f(z)}{z} + \left(\frac{zf'(z)}{f(z)} - 1\right)\left(\frac{\beta z}{f(z)} + \alpha\right)^{-1},$$

the proof goes similarly as that of Theorem 3.8 and the result follows at once. □

By taking $\beta = 1, \alpha = 0$ and $n = 1$ in Theorem 3.10, we obtain the following result.

Corollary 3.11. *Let $f \in \mathcal{A}$ be such that $f'(z) \prec 1+2z$, then $f(z)/z \prec 1+z$, or equivalently*

$$|f'(z) - 1| < 2 \Rightarrow |f(z)/z - 1| < 1.$$

We now prove the following lemma in order to derive some sufficient conditions for starlikeness:

Lemma 3.12. *Let n be a positive integer and α, β be non negative integers with $\beta \neq 0$. Suppose that either $\alpha < \beta < 3\alpha$ or $\beta < \alpha < 3\beta$ and $Q \in \mathcal{H}[1, n]$ satisfy*

$$Q(z) \prec \frac{1+z}{1-z} + \frac{2nz}{(1-z)((\alpha+\beta) + (\alpha-\beta)z)} = h(z). \tag{3.21}$$

If $p \in \mathcal{H}[1, n]$ satisfies the differential equation

$$p(z)Q(z) + \frac{zp'(z)}{\beta p(z) + \alpha} = 1, \tag{3.22}$$

then $p(z) \prec (1-z)/(1+z)$.

Proof. Let us set $q(z) = (1-z)/(1+z)$, then

$$h(z) = \frac{1}{q(z)} - \frac{nzq'(z)}{q(z)(\beta q(z)+\alpha)}.$$

We know that the Koebe function $k(z) = z/(1-z)^2$ is starlike and

$$\begin{aligned}
\operatorname{Re}\frac{zh'(z)}{k(z)} &= \operatorname{Re}\left((1-z)^2\left(\frac{2}{(1-z)^2} + \frac{2n((\alpha+\beta)+(\alpha-\beta)z^2)}{(1-z)^2((\alpha+\beta)+(\alpha-\beta)z)^2}\right)\right) \\
&= \operatorname{Re}\left(2 + \frac{2n((\alpha+\beta)+(\alpha-\beta)z^2)}{((\alpha+\beta)+(\alpha-\beta)z)^2}\right) \\
&= 2 + 2n\operatorname{Re}\left(\frac{(\alpha+\beta)+(\alpha-\beta)z^2}{((\alpha+\beta)+(\alpha-\beta)z)^2}\right).
\end{aligned}$$

Now taking $\alpha+\beta = a$, $\alpha-\beta = b$ and $z = e^{i\theta}$ for $\theta \in (-\pi,\pi)$, we have

$$\begin{aligned}
\left|\arg\left(\frac{a+be^{2i\theta}}{(a+be^{i\theta})^2}\right)\right| &= \left|\arg\left(a+be^{2i\theta}\right) - 2\arg\left(a+be^{i\theta}\right)\right| \\
&= \left|\arctan\left(\frac{b\sin 2\theta}{a+b\cos 2\theta}\right) - 2\arctan\left(\frac{b\sin\theta}{a+b\cos\theta}\right)\right| \\
&\le \left|\arctan\left(\frac{b\sin 2\theta}{a+b\cos 2\theta}\right)\right| + 2\left|\arctan\left(\frac{b\sin\theta}{a+b\cos\theta}\right)\right|.
\end{aligned}$$

Since $\arctan x$ is an increasing function in $(-\pi,\pi)$, we now find the maximum of $g(x) := b\sin x/(a+b\cos x)$ for $x \in (-\pi,\pi)$. Clearly, $a > 0$ and after some elementary calculations, we deduce that $g(x)$ attains its maximum at $x = \pi - \arccos(b/a)$ when $b > 0$ and at $x = -\arccos(-b/a)$ when $b < 0$, $b/\sqrt{a^2-b^2}$ and $-b/\sqrt{a^2-b^2}$ are the corresponding maximum values. Thus we have

$$\left|\arg\left(\frac{a+be^{2i\theta}}{(a+be^{i\theta})^2}\right)\right| \le 3\arctan\left(\frac{|b|}{\sqrt{a^2-b^2}}\right).$$

For the case when $\beta < \alpha < 3\beta$, we observe that

$$3\arctan\left(\frac{|b|}{\sqrt{a^2-b^2}}\right) = 3\arctan\left(\frac{\alpha-\beta}{2\sqrt{\alpha\beta}}\right) < 3\arctan\left(\frac{1}{3}\sqrt{\frac{\alpha}{\beta}}\right) < 3\arctan\left(\frac{1}{\sqrt{3}}\right) = \frac{\pi}{2}.$$

The other case can also be verified in the similar way and it can be concluded that

$$\left|\arg\left(\frac{(\alpha+\beta)+(\alpha-\beta)z^2}{((\alpha+\beta)+(\alpha-\beta)z)^2}\right)\right| < \frac{\pi}{2},$$

which further implies that $\operatorname{Re}(zh'(z)/k(z)) > 0$. So h is close to convex and hence univalent in $\mathbb{D}$. Now we consider the boundary curve of h defined as

$$h(e^{i\theta}) = u(\theta) + iv(\theta), \quad \theta \in (-\pi,\pi).$$

Since $e^{i\theta}$ is a boundary point, without loss of generality we may assume that $(1+e^{i\theta})/(1-e^{i\theta}) = i\gamma$ and thus

$$r(\theta) = |h(e^{i\theta})| = \left|i\gamma + \frac{2ne^{i\theta}}{(1-e^{2i\theta})(\beta/i\gamma+\alpha)}\right| = \left|i\gamma - \frac{n}{\sin\theta(\beta/\gamma+i\alpha)}\right|.$$

So

$$r(\theta)=\sqrt{\gamma^2+\frac{n^2+2n\alpha\gamma\sin\theta}{\sin^2\theta((\beta/\gamma)^2+\alpha^2)}}=\sqrt{\gamma^2+\frac{n^2+4n\alpha\gamma^2/(1+\gamma^2)}{\sin^2\theta((\beta/\gamma)^2+\alpha^2)}}. \tag{3.23}$$

From (3.21) and (3.22), we deduce that

$$Q(z)=\frac{1}{p(z)}-\frac{zp'(z)}{p(z)(\beta p(z)+\alpha)}\prec h(z). \tag{3.24}$$

On the contrary if p is not subordinate to q, then by [16, Lemma.2.2d], there exist points $z_0\in\mathbb{D}$ and $\zeta_0\in\partial\mathbb{D}$ and $m\geq n$, such that $p(z_0)=q(\zeta_0)$ and $z_0p'(z_0)=m\zeta_0q'(\zeta_0)$. From (3.24), we have

$$Q(z_0)=\frac{1}{q(\zeta_0)}-\frac{m\zeta_0q'(\zeta_0)}{q(\zeta_0)(\beta q(\zeta_0)+\alpha)}=\frac{1+\zeta_0}{1-\zeta_0}+\frac{2m\zeta_0}{(1-\zeta_0)((\beta+\alpha)+(\alpha-\beta)\zeta_0)}.$$

For $\zeta_0=e^{i\theta}$, we have

$$|Q(z_0)|=\sqrt{\gamma^2+\frac{m^2+4m\alpha\gamma^2/(1+\gamma^2)}{\sin^2\theta((\beta/\gamma)^2+\alpha^2)}}\geq r(\theta)\quad\theta\in(-\pi,\pi),$$

where $r(\theta)$ is given in (3.23). This implies that $Q(z_0)\notin h(\mathbb{D})$, which is a contradiction and thus we have $p(z)\prec(1-z)/(1+z)$. □

Theorem 3.13. *Let n be a positive integer and α,β be non negative real numbers with $\beta\neq0$, either $\alpha<\beta<3\alpha$ or $\beta<\alpha<3\beta$. Let $f\in\mathcal{A}_n$ and $F=A_{\alpha,\beta}[f]$ is as defined in* (3.20)*. If*

$$\frac{zf'(z)}{f(z)}\prec\frac{1+z}{1-z}+\frac{2nz}{(1-z)((\alpha+\beta)+(\alpha-\beta)z)},$$

then $\operatorname{Re}\left(\dfrac{zF'(z)/F(z)}{zf'(z)/f(z)}\right)>0.$

The proof of Theorem 3.13 follows by an application of Lemma 3.12, similar to that of Theorem 3.7 and therefore it is omitted here.

Theorem 3.14. *Let n be a positive integer and α,β be non-negative real numbers with $\beta\neq0$. Suppose that either $\alpha<\beta<3\alpha$ or $\beta<\alpha<3\beta$ and $f\in\mathcal{A}$ satisfies*

$$\Theta(f)\prec\frac{1+z}{1-z}+\frac{2nz}{(1-z)((\alpha+\beta)+(\alpha-\beta)z)},$$

where

$$\Theta(f)=\frac{f(z)}{zf'(z)}-\left(\frac{zf''(z)}{f'(z)}+1-\frac{zf'(z)}{f(z)}\right)\left(\beta\frac{zf'(z)}{f(z)}+\alpha\right)^{-1},$$

then $zf'(z)/f(z)\prec(1-z)/(1+z)$.

The proof is omitted here as it is much akin to Theorem 3.8 and can be easily done by using Lemma 3.12.

Remark 3.15. If we take $n=1$, $\beta=1$ and $\alpha=0$ in Theorem 3.14, it reduces to a result of Tuneski [25, Theorem 2.1].

Theorem 3.16. *Let n be a positive integer and α, β be non-negative real numbers with $\beta \neq 0$. Suppose that either $\alpha < \beta < 3\alpha$ or $\beta < \alpha < 3\beta$ and $f \in \mathcal{A}$ satisfies*

$$\frac{f(z)}{z} + \left(\frac{zf'(z)}{f(z)} - 1\right)\left(\frac{\beta z}{f(z)} + \alpha\right)^{-1} \prec \frac{1+z}{1-z} + \frac{2nz}{(1-z)((\alpha+\beta)+(\alpha-\beta)z)},$$

then $f(z)/z \prec (1+z)/(1-z)$.

Proof. Taking $p(z) = z/f(z)$ and

$$Q(z) = \frac{f(z)}{z} + \left(\frac{zf'(z)}{f(z)} - 1\right)\left(\frac{\beta z}{f(z)} + \alpha\right)^{-1},$$

the result follows by an application of Lemma 3.12. □

By taking $\beta = 1, \alpha = 0$ and $n = 1$, we obtain the following result

Corollary 3.17. *Let $f \in \mathcal{A}$ be such that*

$$f'(z) \prec \frac{1+z}{1-z} + \frac{2z}{(1-z)^2},$$

then

$$\frac{f(z)}{z} \prec \frac{1+z}{1-z}.$$

4. Concluding Remarks

The work carried out in this article is not an extension of any earlier work but deals with new ideas and therefore there will be no parallel comparisons. As a matter of future scope, these ideas of transforming differential subordinations into integral operators and vice versa can be extended for other similar subordination cases too. Note that the lemmas obtained in this paper are not confined to finding some starlikeness or univalence criteria alone, but also paves way in establishing some exceptional implication results.

Funding

The present work is supported by the Council of Scientific and Industrial Research(CSIR). Ref.No.:08/133(0018)/2017-EMR-I

References

[1] Ali, R.M., Cho, N.E., Ravichandran, V., Kumar, S.S.: Differential subordination for functions associated with the lemniscate of Bernoulli. Taiwanese J. Math. **16**, no. 3, 1017–1026(2012)
[2] Ali, R.M., Ravichandran, V., Seenivasagan, N.: On Bernardi's integral operator and the Briot-Bouquet differential subordination. J. Math. Anal. Appl. **324**, no. 1, 663–668(2006)
[3] Antonino, J.A., Miller, S.S.: F-Classes of Univalent Functions and F-Differential Subordinations. Comput. Methods Funct. Theory (2020). https://doi.org/10.1007/s40315-020-00321-0
[4] Cho, N.E., Kumar, S., Kumar, V., Ravichandran, V.: Differential subordination and radius estimates for starlike functions associated with the Booth lemniscate. Turkish J. Math. **42**, no. 3, 1380–1399(2018)
[5] Cho, N.E., Kumar, V., Kumar, S.S., Ravichandran, V.: Radius problems for starlike functions associated with the sine function. Bull. Iranian Math. Soc. **45**, no. 1, 213-232(2019)
[6] Coman, D.: On a generalized type of the Briot-Bouquet differential subordination, Mathematica (Cluj) **35(58)**, no. 1, 15–19(1993)

[7] Dziok, J.: On some applications of the Briot-Bouquet differential subordination, J. Math. Anal. Appl. **328**, no. 1, 295–301(2007)
[8] Ebadian, A., Bulboaca, T., Cho, N.E., Adegani, E.A.: Coefficient bounds and differential subordinations for analytic functions associated with starlike functions. Rev. R. Acad. Cienc. Exactas Fís. Nat. Ser. A Mat. RACSAM 114(3), 114–128 (2020)
[9] Goel, P., Kumar, S.S.: Certain class of starlike functions associated with modified sigmoid function, Bull. Malays. Math. Sci. Soc. **43**, no. 1, 957–991(2020)
[10] Goodman, A.W.: *Univalent functions. Vol. I*, Mariner Publishing Co., Inc., Tampa, FL, 1983
[11] Kumar, S.S., Goel, P.: Starlike functions and higher order differential subordinations. RACSAM 114, 192 (2020). https://doi.org/10.1007/s13398-020-00925-0
[12] Kumar, S.S., Kumar, V., Ravichandran, V., Cho, N.E.: Sufficient conditions for starlike functions associated with the lemniscate of Bernoulli, J. Inequal. Appl. **176**, 13 (2013)
[13] Ma, W.C., Minda, D.: A unified treatment of some special classes of univalent functions, in *Proceedings of the Conference on Complex Analysis (Tianjin)*, 157–169, Conf. Proc. Lecture Notes Anal., I, Int. Press, Cambridge, MA(1992)
[14] Mendiratta, R., Nagpal, S., Ravichandran, V.: On a subclass of strongly starlike functions associated with exponential function, Bull. Malays. Math. Sci. Soc. **38**, no. 1, 365–386(2015)
[15] Miller, S.S., Mocanu, P.T.: Briot-Bouquet differential superordinations and sandwich theorems, J. Math. Anal. Appl. **329**, no. 1, 327–335(2007)
[16] Miller, S.S., Mocanu, P.T.: *Differential subordinations*, Monographs and Textbooks in Pure and Applied Mathematics, 225, Marcel Dekker, Inc., New York, 2000
[17] Mostafa, A.O., Aouf,M.K., Zayed, M.K., Bulboac˘a, T.:Multivalent functions associated with Srivastava– Saigo–Owa fractional differintegral operator. Rev. R. Acad.Cienc. Exactas Fís. Nat. Ser.AMat.RACSAM 112(4), 1409–1429 (2018)
[18] Nunokawa, M., Sokół, J.: On some differential subordinations, Studia Sci. Math. Hungar. **54**, no. 4, 436–445(2017)
[19] D. Răducanu, Differential subordination associated with generalized Mittag-Leffler function, Rev. R. Acad. Cienc. Exactas Fís. Nat. Ser. A Mat. RACSAM **113** (2019), no. 2, 435–452.
[20] Raina, R. K., Sharma, P., Sokól, J.: Certain classes of analytic functions related to the crescent-shaped regions. J. Contemp. Math. Anal. 53, no. 6, 355–362 (2018)
[21] Ruscheweyh, S.: *Convolutions in geometric function theory*, Séminaire de Mathématiques Supérieures, 83, Presses de l'Université de Montréal, Montreal, QC, 1982.
[22] Ruscheweyh, S., Singh, V.: On a Briot-Bouquet equation related to univalent functions, Rev. Roumaine Math. Pures Appl. **24**, no. 2, 285–290(1979)
[23] Sim, Y.J., Kwon, O.S.: The Briot-Bouquet differential subordination associated with vertical strip domains: Honam Math. J. **39**, no. 4, 503–514(2017)
[24] Sokół, J.: Radius problems in the class $\mathscr{SL}^*$, Appl. Math. Comput. **214**, no. 2, 569–573(2009)
[25] Tuneski, N.: On certain sufficient conditions for starlikeness, Int. J. Math. Math. Sci. **23**, no. 8, 521–527(2000)

Department of Applied Mathematics, Delhi Technological University, Delhi–110 042, India

Email address: spkumar@dce.ac.in

Department of Applied Mathematics, Delhi Technological University, Delhi–110 042, India

Email address: priyanka.goel0707@gmail.com

Journal of Thermal Engineering, Vol. 6, No. 4, pp. 559-576, July, 2020
Yildiz Technical University Press, Istanbul, Turkey

OPTIMUM PARAMETRIC ANALYSIS BASED ON THERMODYNAMIC MODELING OF A COMPRESSION ABSORPTION CASCADE REFRIGERATION SYSTEM

Naushad Ahmad Ansari[1*], Akhilesh Arora[1], Samsher[1], Manjunath K[1]

ABSTRACT

In the present study a cascade refrigeration system (CRS) comprising of a vapor compression refrigeration (VCR) system in the low temperature (LT) stage and a single stage vapor absorption refrigeration (VAR) system in high temperature (HT) stage is analyzed. The scope of this work focuses on the effect of different parameters on the performance of the CRS with the help of a mathematical model. The energy and exergy analysis of the CRS is carried out with CO_2 and NH_3 as refrigerants in the VCR stage and the pair of H_2O-LiBr fluids in the VAR stage. It is observed that COP and exergetic efficiency of the CRS reduce with increase in cascade condenser temperature with CO_2 as a refrigerant in VCR stage whereas with NH_3 as a refrigerant in VCR stage, the COP increases with increase in cascade condenser temperature, it attains a maximum value and then decreases.

Keywords: *Absorption, Compression, COP, Generator, Cascade Condenser, Exergetic Efficiency*

INTRODUCTION

A cascade refrigeration system is suitable for low temperature applications in the temperature range from -30°C to -100°C in various industries such as pharmaceutical, food, chemical, blast freezing and liquefaction of gases. For achieving temperature below -40°C by using a simple vapor compression system or a single stage absorption system is uneconomical as suggested by Chakravarthy et al. (1). The disadvantage of the VCR systems in this type of applications is their high electricity consumption. This disadvantage can be overcome by a novel cascade refrigeration system comprising of VCR system in low temperature stage and a VAR system in high temperature stage. The VAR system can be operated by waste heat assumed to be supplied in the generator of absorption section by industrial waste heat as done by Cimsit and Ozturk (2) and therefore electricity consumption can be reduced. Reddy et al (3) investigated that a natural gas-fired combined cycle power plant is better compared to a coal-fired supercritical thermal power plant from energetic and exergetic efficiency view point. Reddy et al. (4) did exergy analysis of vapor compression refrigeration system with few selected refrigerants, on comparing the performance they found that refrigerant R134a has shown better results in all respect whereas refrigerant R407C has shown poor results. The energy analysis of a compression absorption cascade refrigeration system with CO_2 and NH_3 as refrigerants in the compression stage and the pair of NH_3–H_2O in the absorption stage was done and by Seara et al. (5) and the COP of compression system to be 2.602 and 2.463 with CO_2 and NH_3 respectively at an evaporator temperature of -45^{0}C was computed.

Cimsit and Ozturk (2) in the study using H_2O-LiBr and NH_3–H_2O in absorption stage and R-134A, R-410A and NH_3 in vapour compression stage investigated that overall COP of the system improves by 33% with H_2O-LiBr, electrical energy consumption in the cascade refrigeration cycle is 48-51% lower than the conventional vapour compression cycle and also heat energy requirement reduces by 35% at generator side. Riffat and Shankland (6) did the energy analysis of various configuration of vapor absorption and integrated compression absorption systems for water-lithium bromide systems and obtained higher COP for integrated system compared to individual one. Kairouani and Nehdi (7) in a proposed cascade cycle which is powered by geothermal energy with NH_3–H_2O in absorption system and R717, R22 and A134A in vapour compression system reported that COP of cascade cycle to be 37-54% higher than the vapour compression cycle and calculated the COP as 5.5. Meng et al. (8) in their study proposed a hybrid system utilising low grade energy from solar energy to run the absorption stage saving conventional energy used NH_3-H_2O in absorption stage and R134A in vapour compression stage. Colorado and Velazquez (9) performed exergy based thermodynamic analysis of compression-absorption refrigeration cycle using R134a, NH3 and CO2 in VCR cycle and H2O-LiBr in VAR cycle to find out best working substance and suitable operating parameters. They concluded that highest irreversibility takes place in cascade condenser. Garimella et al. (10) using a computational model of a waste heat driven cascade system having H2O-LiBr in VAR stage and CO_2 in sub-critical VCR stage predicted that their novel cascade system consumes 31% less electricity compared to a conventional VCR cascade system. Seyfouri and Ameri (11) proposed a configuration

This paper was recommended for publication in revised form by Regional Editor Vineet Veer Tyagi
[1]Mechanical Engineering Department, Delhi Technological University, Bawana Road, Delhi-42, INDIA
**E-mail address: naushad.nsr@gmail.com*
Orcid id: 0000-0002-4736-4261, 0000-0002-5984-406X, 0000-0003-4846-1343, 0000-0001-7732-954X
Manuscript Received 30 July 2017, Accepted 08 November 2017

of integrated refrigeration system of compression chiller and absorption chiller where a micro turbine is used to run the compression chiller at the low temperature stage and waste heat of micro turbine is used to run the absorption chiller thus saving the energy. Wang et al. (12) developed solar-assisted cascade refrigeration system. The system consists of electricity-driven vapour compression refrigeration system and solar-driven vapour absorption refrigeration system. They predicted that the system gives higher COP as compared with the conventional vapour compression refrigeration system. Chen et al. (13) investigated a system comprising of three sub systems namely power generation subsystem, an ammonia-water two stage absorption refrigeration subsystem and a CO_2 compression refrigeration sub-system has concluded that COP of such two-stage absorption system is 50% higher than the conventional two stage absorption refrigeration system when the refrigeration temperature is lower than -30^0C. Arora et al. (14) predicted that optimum solution distribution ratio was same for maximum COP and maximum exergetic efficiency for a double effect parallel flow absorption system with LiBr-H_2O as working pair of fluid. Kutlu, et al. (15) theoretically analyzed a two-phase ejector refrigeration system with bi-evaporator refrigeration and predicted the COP improvement up-to 21% in Mediterranean climate region. Dixit et al. (16) did the analysis on a triple effect refrigeration cycle which was waste driven and consisted of an ejector organic Rankine cycle, vapor absorption cycle and vapor compression cycle and computed the energy efficiency to be 21.79% while exergetic efficiency to be around 12.91%. Arora et al. (17) carried out the thermodynamic analysis of a water/lithium-Bromide based half effect absorption refrigeration system and predicted a same optimum intermediate pressure for maximum COP and maximum exergetic efficiency for the system while further, predicting maximum COP range to be 0.415-0.438 and maximum exergetic efficiency range to be 6.96-13.74%. Jain et al. (18) did the optimisation with combination of R410a as a refrigerant in VCR stage and aqua Li-Br solution in VAR stage in similar cascade refrigeration system where non-linear programming has been used to optimize the system. Jain et al. (19) did thermodynamic study of similar cascade refrigeration system using water and Li-Br solution in VAS stage and R22, R 410A, R 407C and R134A in VCR stage, while comparing their performances accordingly found that R134a as a suitable refrigerant. Jain et al. (20) did the multi-objective optimization using NSGA-II technique on the performance of a 170kW vapour-absorption cascade refrigeration system and concluded that multi-objective design criteria is better than two single-objective thermodynamic and total product cost optimized designs. Dixit et al. (21) carried a study on similar cascade refrigeration cycle with CO_2, NH_3 and R-134a in compression stage and H_2O-LiBr in absorption stage predicted that combination of R-134a with H_2O-LiBr is most suitable of the three combinations based on comparing operational parameters of such a system. Kaushik et al. (22) studied a similar system and predicted the optimum generator temperature for maximum COP to be 83.5^0C and for maximum exergetic efficiency to be 72.25^0C with ammonia as a refrigerant in the low temperature cycle.

Megdouli et al (23) analyzed a hybrid vapor compression refrigeration (HVCR) system, which combines a vapor compression refrigeration (VCR) system and an ejector refrigeration (ER) system. Thermodynamic investigations using CO2 as a refrigerant, were performed with energetic and exergetic methods, and the comparative analyses with the VCR system were conducted. The results indicated that for the same cooling capacity, the coefficient of performance (COP) of the HVCR system showed 25% higher COP and the total mechanical power consumption was reduced by 20% than that of conventional VCR system, respectively. Patel et al (24) in their study on cascaded vapor compression-absorption system powered by solar-biomass organic Rankine cycle with working fluid Toluene and R245fa proposed and analyzed that Thermo-economic performance of the system is influenced by the type of working fluid of organic Rankine cycle, electricity and heating cost, biomass type, discount rate, solar collector field and location of installation. Chen et al (25) proposed and analyzed a novel heat driven absorption compression system to reach a temperature as low as -60^0C. The proposed system had three sub-systems namely a power cycle, an ammonia water absorption cycle and a CO2 vapor compression refrigeration cycle. They concluded that the results predicted were adaptable and suitable for practical applications. Yingjie Xu et al (26) proposed a low-grade heat driven cascade system named as ejection-compression refrigeration system and predicted that the proposed system with low electricity consumption had higher COP compared to a conventional vapor compression system and conventional ejector compression refrigeration system. Yingjie Xu et al (27) studied and compared two low grade heat driven compression-absorption cycles. Their study was based on energy, exergy, environmental and economic. They compared a novel absorption-compression cycle with evaporator sub-cooler and a conventional absorption-compression cycle with an evaporator-condenser. They predicted that novel absorption-compression system was more economical when waste heat was used whereas the conventional absorption-compression system was more economical when solar heat was used.

This paper seeks to remedy the highlighted gap of knowledge in the field of the study by conducting the thermodynamic analysis based on energy and exergy analysis to achieve a compromise between the optimum generator temperature, cascade condenser temperature, cascade approach and other variables at which maximum COP and maximum energetic efficiency occurs. Thus, it could be possible to formulate the theoretical framework of the procedures for energy saving and optimization which is environment friendly.

DESCRIPTION OF THE COMPRESSION ABSORPTION CASCADE REFRIGERATION SYSTEM

The compression-absorption cascade system which is being considered is shown schematically in Figure1(a). It consists of a single stage compression system in the low temperature stage and a single effect absorption system in the high temperature stage. The compression system comprises of the evaporator, compressor, condenser and an expansion device. The absorption system uses the pair water-lithium bromide and its major components are the absorber, generator, condenser, evaporator, solution heat exchanger, pump, solution throttle valve and a refrigerant throttle valve. Both systems share a common heat exchanger (cascade condenser), which operates simultaneously as the condenser of the compression system and as the evaporator of the absorption system. The compression system is used for the generation of the cooling at low temperature and the absorption system rejects heat to the surroundings, as shown in Figure1(a). This refrigeration system would decrease the electricity consumption compared to the two stages compression systems, since it is only required to operate the compression system at the low stage; meanwhile the absorption system is driven by heat.

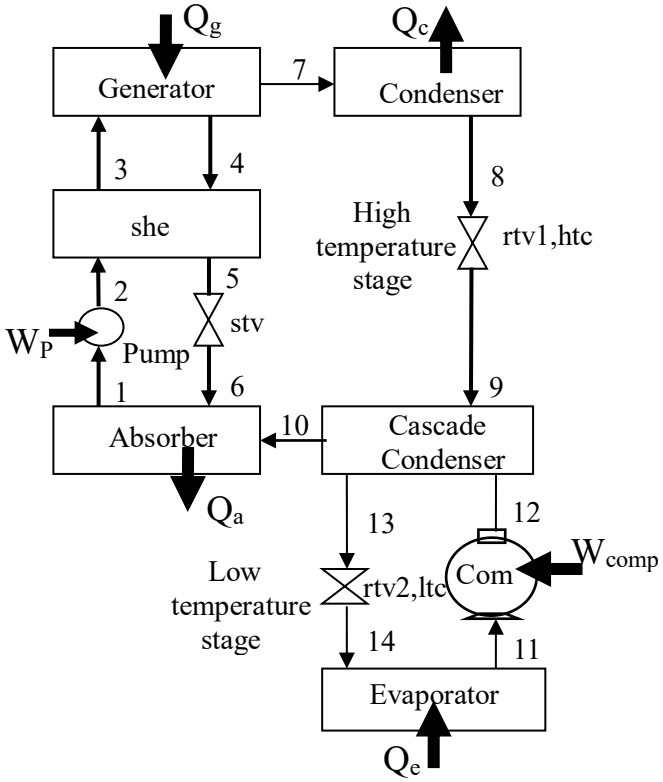


Figure 1(a). Schematic diagram of compression-absorption cascade refrigeration system

Where, Q_a= heat rejected from absorber, W_p= pump work, Q_g= heat generated in generator, Q_c= heat rejected from condenser, W_{comp}= compressor work, Q_e= heat absorbed in evaporator, stv= solution thermostatic expansion valve, rtv1= refrigerant thermostatic expansion valve in htc, htc= high temperature circuit, rtv2= refrigerant thermostatic expansion valve in ltc, ltc= low temperature circuit, she= solution heat exchanger.

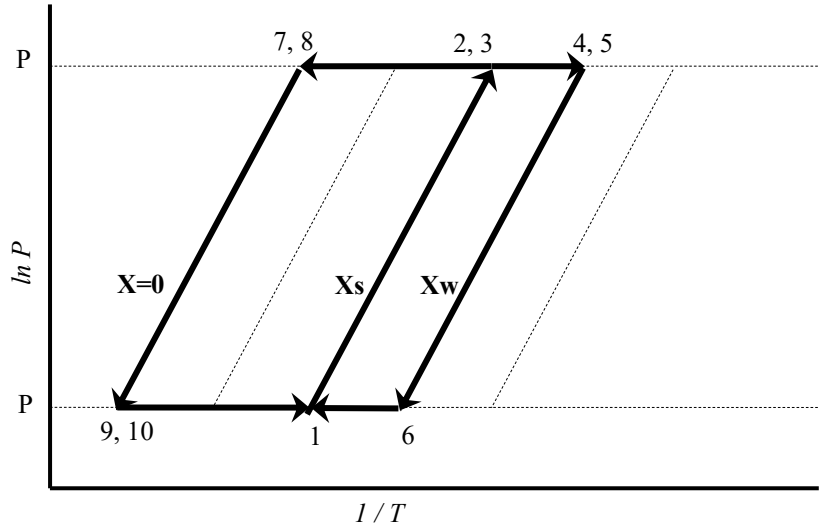


Figure 1(b). ln P-1/T diagram for water lithium bromide Single Effect Generation VAR system

THERMODYNAMIC ANALYSIS OF THE COMPRESSION ABSORPTION CASCADE REFRIGERATION SYSTEM

The thermodynamic analysis of compression absorption system involves the principles of mass conservation, energy conservation and exergy balance.

Mass Balance

The mass flow rate of refrigerant through each component of low temperature circuit is $\dot{m}_{r_vcr}$ (kg/s). It is calculated using equation (1).

$$\dot{Q}_e = \dot{m}_{r_vcr}(h_{11} - h_{14}) \tag{1}$$

Where, $\dot{Q}_e$ is heat absorbed in evaporator (kW) and h is enthalpy (kW/kg)

Mass balance equations in high temperature circuit are specified below.

Mass balance at absorber or generator

$$\dot{m}_s = \dot{m}_r + \dot{m}_w \tag{2}$$

Where, $\dot{m}_s$ is the mass of solution (kg/s) flowing through absorber and generator.
$\dot{m}_r$is the mass flow rate (kg/s) through condenser and cascade condenser.
$\dot{m}_w$ is the mass (kg/s) of absorbent flowing through solution heat exchanger.

Energy Balance

The energy balance equations for the system shown in Figure 1(a) are given below:
Energy balance in high temperature stage

$$\dot{Q}_a = \dot{m}_r h_{10} + \dot{m}_w h_6 - \dot{m}_s h_1 \tag{3}$$

$$\dot{Q}_g = \dot{m}_r h_7 + \dot{m}_w h_4 - \dot{m}_s h_3 \tag{4}$$

$$\dot{Q}_c = \dot{m}_r(h_7 - h_8) \tag{5}$$

$$\dot{Q}_{she} = \dot{m}_s(h_3 - h_2) = \dot{m}_w(h_4 - h_5) \tag{6}$$

$$\dot{W}_p = \dot{m}_s(h_2 - h_1) \tag{7}$$

Energy balance in low temperature stage

$$\dot{W}_{comp} = \dot{m}_{r_vcr}(h_{12} - h_{1_1}) \quad (8)$$

$$\dot{Q}_e = \dot{m}_{r_vcr}(h_{11} - h_{14}) \quad (9)$$

$$Power\ \text{input} = \dot{W}_{\text{comp}} + \dot{W}_p \quad (10)$$

Energy input

$$\dot{E}_{in} = \dot{Q}_g + \dot{Q}_e + \dot{W}_p + \dot{W}_{comp} \quad (11)$$

Energy output

$$\dot{E}_{out} = \dot{Q}_a + \dot{Q}_c \quad (12)$$

Coefficient of Performance (COP)

$$COP = \frac{\dot{Q}_e}{\dot{Q}_g + \dot{W}_p + \dot{W}_{comp}} \quad (13)$$

Exergy Balance

Exergy destruction in components of the cascade system is furnished below:

Exergy analysis in high temperature stage

$$\dot{ED}_a = \dot{m}_r(h_{10} - T_o s_{10}) + \dot{m}_w(h_6 - T_o s_6) - \dot{m}_s(h_1 - T_o s_1) \quad (14)$$

$$\dot{ED}_g = \dot{m}_s(h_3 - T_o s_3) - \dot{m}_w(h_4 - T_o s_4) - \dot{m}_r(h_7 - T_o s_7) + \dot{Q}_g\left(1 - \frac{T_0}{Tg}\right) \quad (15)$$

$$\dot{ED}_c = \dot{m}_r((h_7 - h_8) - T_o(s_7 - s_8)) \quad (16)$$

$$\dot{ED}_{cc} = \dot{m}_r((h_9 - h_{10}) - T_o(s_9 - s_{10})) + \dot{m}_{r_vcr}((h_{12} - h_{13}) - T_o(s_{12} - s_{13})) \quad (17)$$

$$\dot{ED}_{she} = \dot{m}_s((h_2 - h_3) - T_o(s_2 - s_3)) + \dot{m}_w((h_4 - h_5) - T_o(s_4 - s_5)) \quad (18)$$

$$\dot{ED}_{rtv_htc} = \dot{m}_r T_o(s_9 - s_8) \quad (19)$$

$$\dot{ED}_{stv} = \dot{m}_w T_o(s_6 - s_5) \quad (20)$$

Exergy analysis in low temperature stage

$$\boldsymbol{\dot{ED}_{comp} = \dot{m}_{r_vcr} T_0(s_{12} - s_{11})} \quad (21)$$

$$\dot{ED}_e = \dot{m}_{r_vcr}(h_{14} - T_0 s_{14}) + \dot{Q}_e\left(1 - \frac{T_0}{T_r}\right) - \dot{m}_{r_vcr}(h_{11} - T_0 s_{11}) \quad (22)$$

$$\dot{ED}_{rtv_ltc} = \dot{m}_{r_vcr}(h_{13} - T_0 s_{13}) - -\dot{m}_{r_vcr}(h_{14} - T_0 s_{14}) = \dot{m}_{r_vcr} T_0(s_{14} - s_{13}) \quad (23)$$

$$\dot{ED}_{total} = \dot{ED}_a + \dot{ED}_g + \dot{ED}_c + \dot{ED}_{cc} + \dot{ED}_{she} + \dot{ED}_{rtv_htc} + \dot{ED}_{stv} + \dot{ED}_{rtv_ltc} + \dot{ED}_{comp} + \dot{ED}_e \quad (24)$$

Exergetic Efficiency

$$\eta_{e_x} = \frac{Exergy\ in\ product}{Exergy\ of\ fuel} = \frac{\dot{EP}}{\dot{EF}} = \frac{\dot{EF} - \dot{ED}_{total}}{\dot{EF}} = 1 - \frac{\dot{ED}_{total}}{\dot{EF}} \quad (25)$$

$$\dot{EF} = \dot{W}_{comp} + \dot{W}_p + \dot{Q}_g \left(1 - \frac{T_0}{Tg}\right) \quad (26)$$

$$\dot{EP} = \dot{Q}_e \left|\left(1 - \frac{T_0}{T_r}\right)\right| \quad (27)$$

$$\eta_{e_x} = \frac{\dot{Q}_e \left|\left(1 - \frac{T_0}{T_r}\right)\right|}{\dot{W}_{comp} + \dot{W}_p + \dot{Q}_g \left(1 - \frac{T_0}{Tg}\right)} \quad (28)$$

Model validation

The present work has been validated with the Cimsit and Ozturk (2) investigation with LiBr-H_2O and NH_3 as working fluids in VAR cycle and VCR cycle respectively. The difference between the present data and predicted by Cimsit and Ozturk (2) in Table 1 falls within the accepted range.

Table 1. Result validation

Parameters	Present Model	Cimsit and Ozturk (2012)	Difference %	Operating Parameters
$\dot{Q}_a$ (kW)	72.76	73.15	0.53	T_c= 40^{0}C
$\dot{Q}_g$ (kW)	76.81	76.45	0.47	T_g= 90^{0}C
$\dot{Q}_c$ (kW)	61.17	61.06	0.1	T_{cc}= 10^{0}C
$\dot{W}_{COMP}$ (kW)	7.83	8.08	-3.09	T_{e_vcr}= -10^{0}C
COP	0.594	0.592	0.34	ϵ_{she}= 0.60 η_{comp} =72%
				$\dot{Q}_e$= 50 (kW)

Thermodynamic modeling of a similar system done by Kaushik et al. (22) predicted the optimum generator temperature for maximum COP to be 83.5ºC and maximum exergetic efficiency to be 72.25ºC with ammonia as fluid in low temperature stage which can be validated for the same as calculated in the present study to be 84.21ºC and 72.59ºC respectively. The error may be due to approximation but in agreement could be accepted. Similarly, at higher absorber and condenser temperature (40ºC) and other same parameters Dixit et al. (21) investigated cascade condenser temperature to be 12.47ºC and 8.64ºC for maximum COP for ammonia and carbon dioxide respectively. In the present study the same has been computed as 13ºC and 9ºC for ammonia and carbon dioxide respectively. The error may be due to approximation and could be agreed upon.

RESULTS AND DISCUSSION

The present analysis has been done considering two different natural refrigerants, carbon dioxide and ammonia in the compression system. The absorption system uses the water-lithium bromide pair. A simple steady state simulation model based on sequential modular approach has been developed and implemented in a computer program using EES software taken from Klein and Alvarado (28). The model equations are formulated from species, mass, energy and exergy balances. State equations used for the carbon-dioxide and ammonia at equilibrium and the thermodynamic properties are taken from built-in functions in EES. The properties of water-lithium bromide solution are taken from Pa´tek and Klomfar (29). The parameters assumed for computation of results are mentioned below:

1. Cooling capacity ($\dot{Q}_e$) : 100 kW
2. Isentropic efficiency of compressor, (η_{comp}) : 60-80 %
3. Evaporator temperature, (T_{e_vcr}) : -35°C to -45°C
4. Cascade Condenser temperature, (T_{cc}) : 2-14°C
5. Generator temperature (T_g) : 55-115°C
6. Absorber temperature (T_a) : 25-40°C
7. Effectiveness of solution heat exchanger, ($\in_{she}$) : 0.6-0.8
8. Condenser temperature (T_c) : 25-40°C
9. Approach in cascade condenser (A) : 0-10°C

The cascade condenser temperature is assumed to vary between 2-14°C because the lowest temperature in water lithium bromide system depends upon the freezing point of water and which can't be below 0°C.

Effect of Cascade Condenser Temperature on COP and Exergetic Efficiency

Since the selection of cascade condenser temperature depends upon the freezing point of refrigerant (water), hence for a given set of parameters viz. evaporator temperature (depends upon application), condenser and absorber temperatures (depends upon ambient conditions) there must exist a unique generator temperature corresponding to which COP is maximum. Similar to this, the exergetic efficiency will be maximum corresponding to some other value of generator temperature. This point is highlighted in Figure 2 for ammonia and carbon dioxide cascade systems. It can be observed that maximum COP and maximum exergetic efficiency occur corresponding to different generator temperatures. The maximum COP occurs corresponding to generator temperature of 84.21°C and maximum exergetic efficiency occurs corresponding to 72.59°C for both the refrigerants. It may be predicted that under similar condition of the cycle the optimum generator temperature remains same irrespective of the refrigerants in low temperature stage.

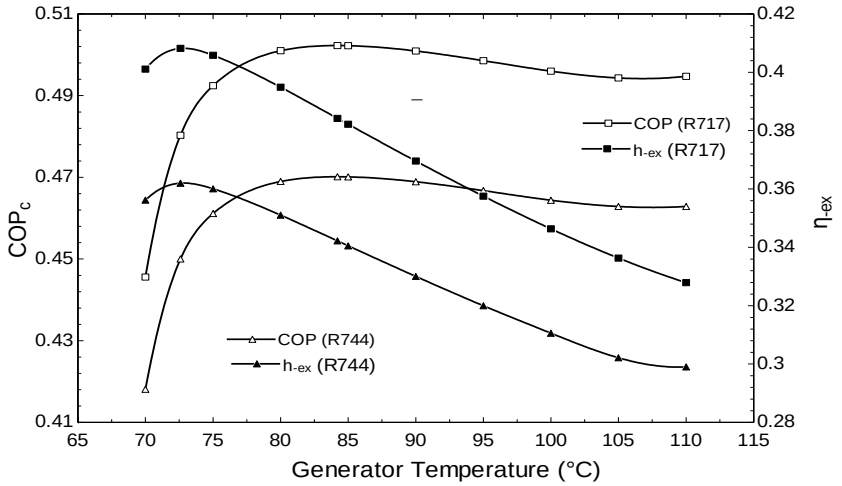
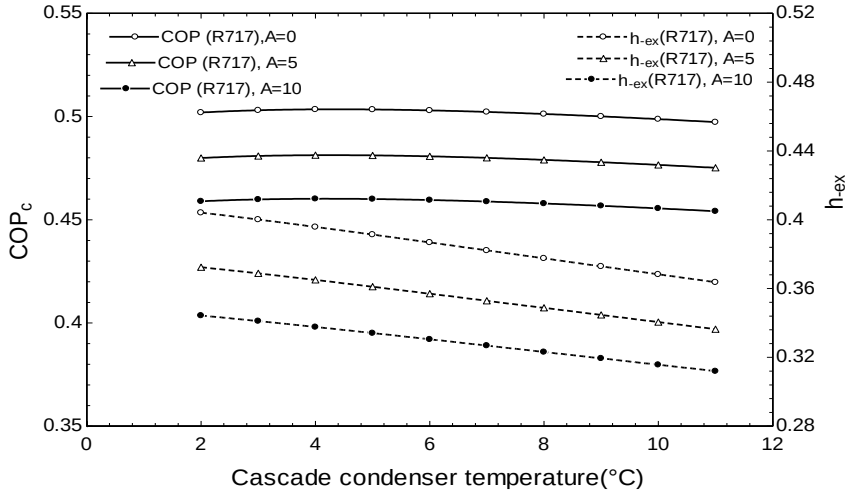
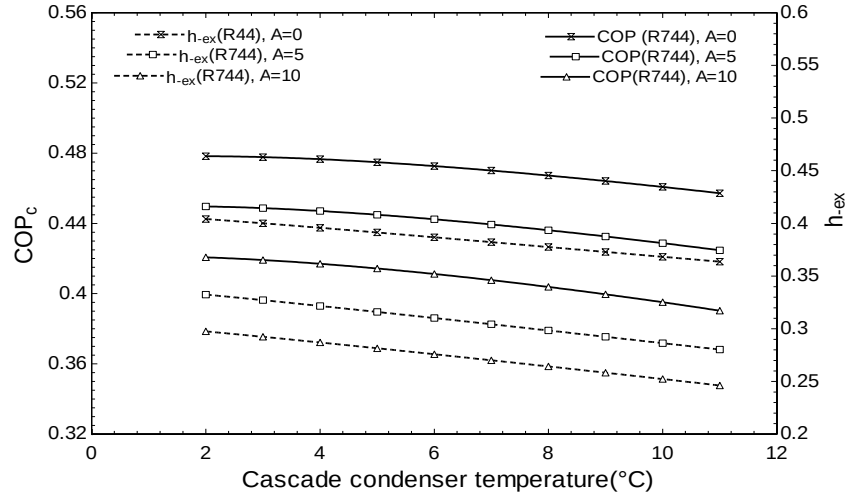


Figure 2. Variation of COP and η_ex versus generator temperature for ammonia "R717" and carbon-dioxide "R744" (Ta = Tc = 35°C, Te_vcr = -40°C,ε_she = 0.75, η_comp =0.8, Qe_vcr = 100 kW, A = 0°C,To= 298.15 K)



(a)



(b)

Figure 3. Effect of cascade condenser temperature on COP and exergetic efficiency for **(a)** ammonia "R717" and **(b)** carbon dioxide "R744" (Ta = Tc = 35°C, Te_vcr = -40°C, Tg = 84.5°C, ε_she = 0.75, η_comp =0.8 , Qe_vcr = 100 kW, A = 0 °C , To = 298.15 K)

Figure 3(a) and (b) show the effect of cascade condenser temperature on COP and exergetic efficiency for ammonia and carbon-dioxide in low temperature circuit. It is observed that both COP and exergetic efficiency reduce with increase in cascade condenser temperature for carbon-dioxide cycle. In ammonia cascade system, the COP first increases, achieves a maximum value at cascade temperature of 4°C and then reduces. Also, at different approach (A=0, 5 and 10) the maximum COP occurs at same cascade temperature of 4°C. It is observed that with increase in approach COP decreases. However, the difference between maximum and minimum values is very small. The carbon-dioxide cycle cascade system has the lower values of COP and exergetic efficiency in comparison to ammonia cycle cascade system. The increase in cascade condenser temperature results in increase in pressure ratio across compressor, increase in mass flow rate in compression circuit and increase in absorber pressure. The first two factors account for increase in compressor power and the last factor i.e. increase in absorber pressure accounts for decrease in strong solution concentration *'Xs'*. The weak solution concentration '*Xw'* remains constant. Thus, the solution circulation ratio (given by *Xw/(Xw-Xs)* decreases. The reduction in solution circulation ratio reduces the generator heat duty. Thus, COP of the system may increase or decrease depending upon the increase in compressor power requirement and reduction in generator heat duty.

Further the optimum generator temperature corresponding to maximum COP is higher in comparison to optimum generator temperature corresponding to maximum exergetic efficiency. Also, at optimum generator temperature the cascade condenser temperature corresponding to maximum COP is higher in comparison to

cascade condenser temperature corresponding to maximum exergetic efficiency. So, it could be a tradeoff to choose between maximum COP and maximum exergetic efficiency the appropriate values of temperatures in both the conditions.

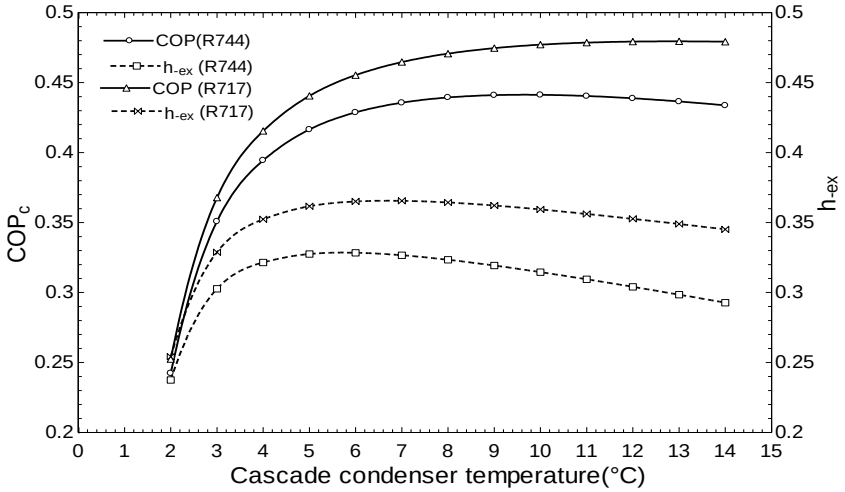


Figure 3(c). Effect of cascade condenser temperature on COP and exergetic efficiency for carbon dioxide "R744" and ammonia "R714" (Ta = Tc = 40°C, Te_vcr = -40°C, Tg = 84.5°C, ε_she = 0.75, η_comp =0.8 , Qe_vcr = 100 kW, A = 0 °C , To = 298.15 K)

Figure 3(c) shows the effect of cascade condenser temperature at (optimum generator temperature) on COP and exergetic efficiency but at higher absorber and condenser temperature which is 40°C in this study for both the ammonia and carbon dioxide as refrigerants in low temperature stage. It is observed that with the increase in cascade temperature both the COP and exergetic efficiency increases significantly, reaches to a maximum value and then start to decrease. For ammonia maximum COP and maximum exergetic efficiency occur at 13°C and 7°C respectively whereas for carbon dioxide maximum COP and maximum exergetic efficiency occur at 9°C and 6°C respectively.

Figure 4(a) represents the values of maximum COP and maximum exergetic efficiency with variation in cascade condenser temperature for ammonia. The optimum generator temperature for ammonia has also been computed corresponding to both maximum COP and maximum exergetic efficiency. It is observed that the optimum generator temperature decreases with increase in cascade condenser temperature. The optimum generator temperature for maximum COP is higher than optimum generator temperature for maximum exergetic efficiency. Figure 4(b) illustrates the comparison of maximum exergetic efficiencies of ammonia and carbon-dioxide and corresponding optimum generator temperatures. The values of exergetic efficiency are also calculated for generator temperature equal to 84.5ºC. It is observed that if generator temperature is not optimum then exergetic efficiency continues to decrease with increase in cascade condenser temperature. The maximum exergetic efficiency and corresponding generator temperature also decrease with increase in cascade condenser temperature. The maximum exergetic efficiency and optimum generator temperature both are lower for carbon-dioxide cycle cascade system in comparison to ammonia cycle cascade system. The cascade condenser temperature corresponding to the optimum generator temperature is certainly optimum cascade condenser temperature. However the maximum exergetic efficiency is not the global maximum exergetic efficiency rather it is localized maximum value of exergetic efficiency corresponding to given conditions.

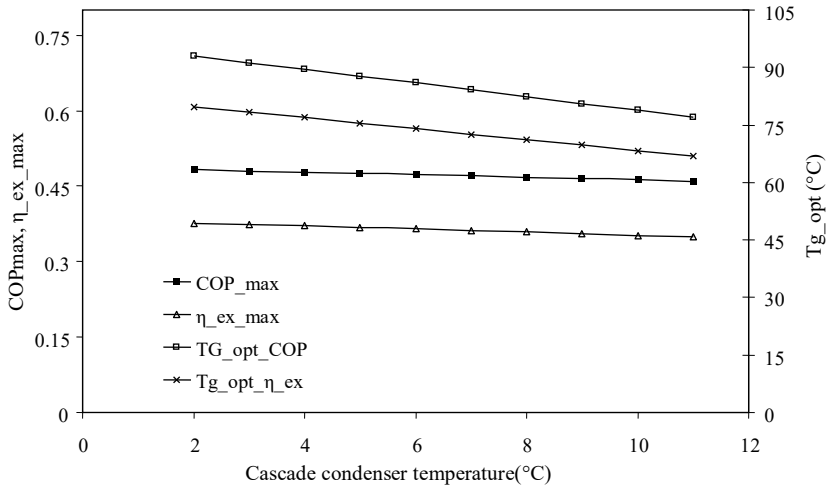


Figure 4(a). COP_max and η_ex_max versus cascade condenser temperature and corresponding Tg_opt for ammonia "R714" (Ta = Tc = 35 °C, Te_vcr = -40 °C, ε_she = 0.75, η_comp =0.8, Qe_vcr = 100 kW, A = 0 °C, To = 298.15 K)

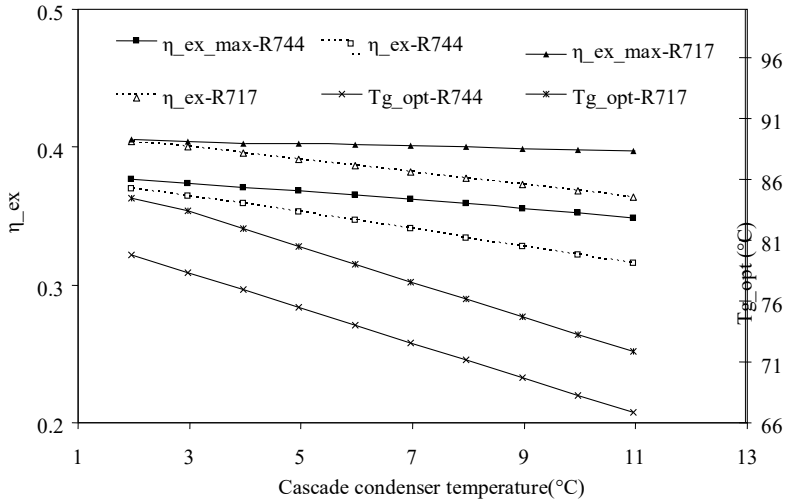
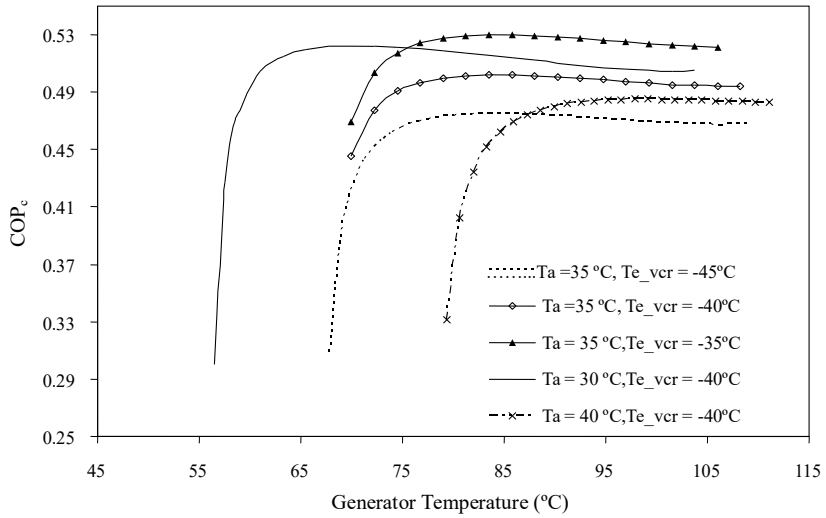


Figure 4(b). Exergetic efficiency versus cascade condenser temperature and Tg_opt corresponding maximum exergetic efficiency for ammonia "R744" and carbon dioxide "R744" (Ta = Tc = 35 °C, Te_vcr = -40 °C, ε_she=0.75, η_comp =0.8, Qe_vcr = 100 kW, A = 0 °C, To = 298.15 K)

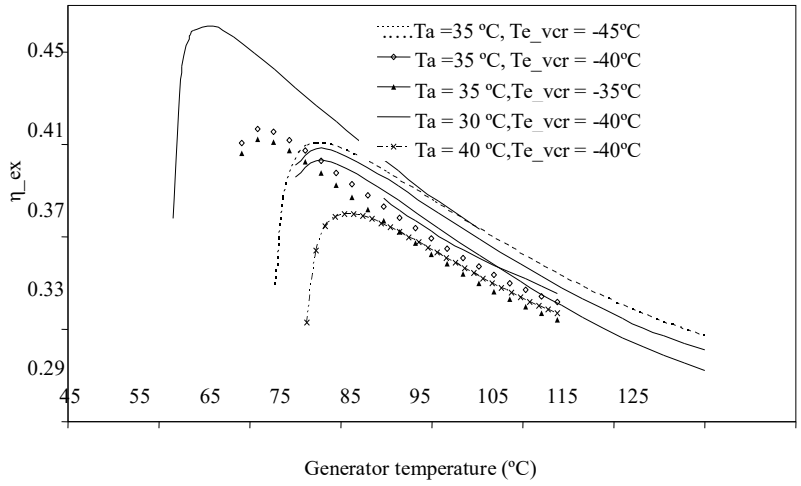
Effect of Generator Temperature

Figure 5(a) and (b) illustrate the effect of generator temperature on COP and exergetic efficiency at different absorber temperatures and different evaporator temperatures. It is observed that with increase in generator temperature (for a specific absorber and specific evaporator temperature), COP increases initially, attains a maximum value and with further increase in generator temperature COP decreases marginally. The trend followed by exergetic efficiency curves is similar to trend followed by COP curves with the exception that the fall of exergetic efficiency is significant after it attains maximum value. At particular evaporator and cascade condenser temperatures, the compressor power required for a constant cooling load remains constant. This implies that COP of the low temperature stage is constant. Thus the variation in COP of the cascade system is dependent on variations in COP of the high temperature i.e. COP of the single effect absorption system. The increase in generator temperature causes the solution circulation ratio to reduce and the heat duty of the generator reduces. The reduction in generator heat duty increases the COP of the high temperature stage and consequently the COP of the cascade system increases initially. However with further increase in generator temperature, the temperature difference between generator and sub-cooled solution entering the generator increases. Thus the irreversibility in generator

increases. The increase in generator temperature also accounts for increase in temperatures of weak solution exiting the generator and the refrigerant exiting the generator.



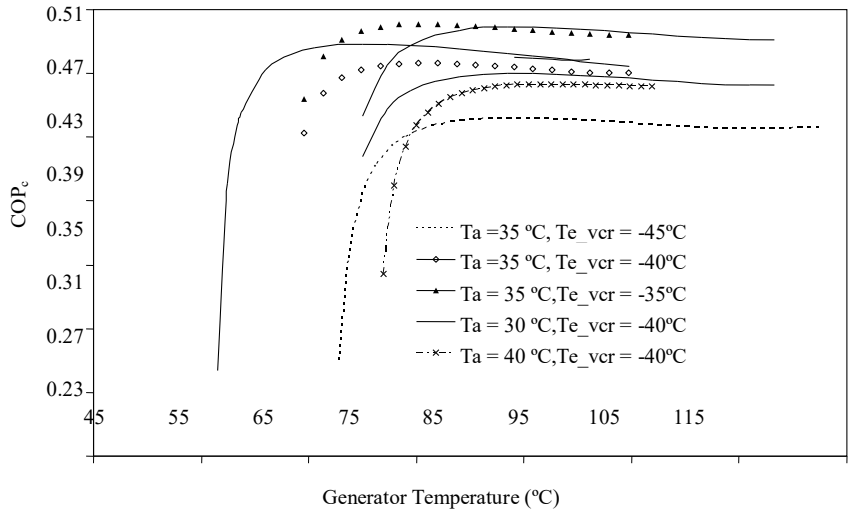
(a)



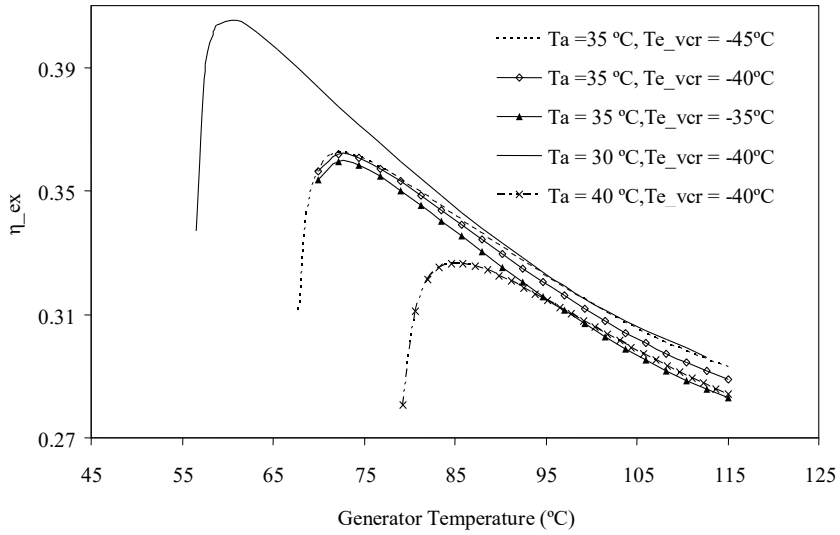
(b)

Figure 5. Effect of generator temperature on **(a)** COP and **(b)** exergetic efficiency for range of evaporator and absorber temperatures (Refrigerant ammonia "R714", Ta = Tc, Tcc = 7ºC, effectiveness = 0.75, Qe_vcr = 100 kW, compressor efficiency = 0.8, Approach = 0 ºC)

Thus, irreversibility in absorber and condenser increases, hence the positive effect of increase of generator temperature on COP of high temperature stage is counterbalanced by negative effects of increase in irreversibility in generator, absorber and condenser. Hence COP of the cascade system becomes constant and even reduces marginally with increase in generator temperature. The reduction in generator heat duty brings down the exergy input initially due to fall in solution circulation ratio. The exergetic efficiency, therefore, increases and achieves a maximum value. The rate of decrease of solution circulation ratio reduces with further increase in generator temperature and therefore exergy input (given by $\dot{Q}_g\left(1-\frac{T_0}{T_g}\right)+\dot{W}_p+\dot{W}_{c_vcr}$) increases whereas output exergy (given by $\dot{Q}_e\left|\left(1-\frac{T_0}{T_r}\right)\right|$) remains constant and hence the exergetic efficiency drops sharply. The trend of various parameters which constitute input exergy is shown in Figure 7.



(a)



(b)

Figure 6. Effect of generator temperature on **(a)** COP and **(b)** exergetic efficiency for range of evaporator and absorber temperatures (Refrigerant carbon-dioxide "R744", Ta = Tc, Tcc = 7ºC, effectiveness = 0.75, Qe_vcr = 100 kW, compressor efficiency = 0.8, Approach = 0 ºC)

Figure 5 (a) and (b) also show that with increase in evaporator temperature the COP increases whereas exergetic efficiency decreases. The COP increases because increase in evaporator temperature reduces the pressure ratio across the compressor and hence the compressor power required reduces which in turn reduces the total input energy required. The exergetic efficiency reduces because of reduction in output exergy. The input exergy required also reduces however the rate of decrease in input exergy is lower in comparison to rate of decrease of output exergy. The Figure 5(a) and (b) also show the effect of absorber temperature on COP and exergetic efficiency. The reduction in absorber temperature accounts for reduction in solution circulation ratio which reduces generator heat duty and increases COP. The reduction in generator heat duty also accounts for reduction in input exergy, it results in increase in exergetic efficiency. Simultaneously, the optimum generator temperature also reduces with reduction in absorber temperature. This indicates that an absorption compression cascade refrigeration system requires lower temperature heat source when absorber temperature is reduced.

Figure 6(a) and (b) illustrate the variation of COP and exergetic efficiency for carbon-dioxide cascade system. The comparison of Figures (5) and (6) illustrates that the COP and exergetic efficiency values for carbon-dioxide cascade cycle are lower in comparison to ammonia cascade cycle.

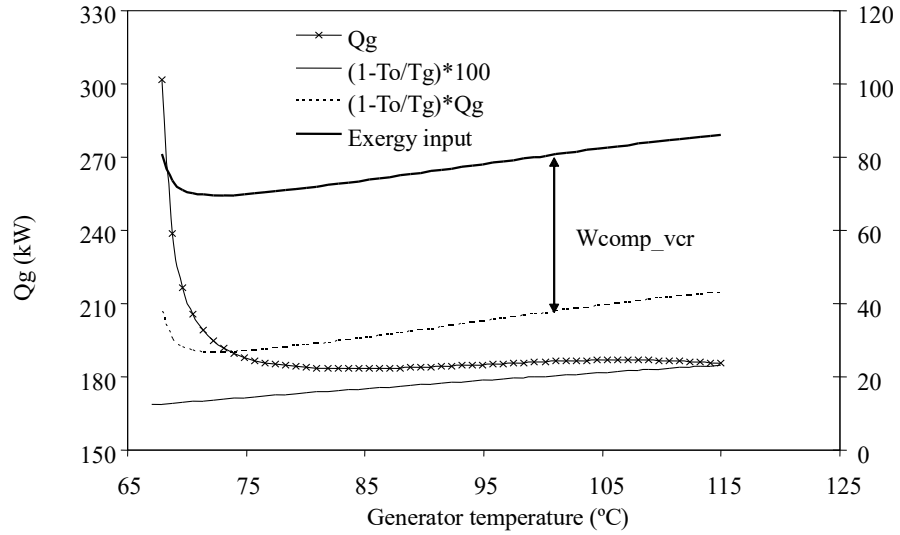


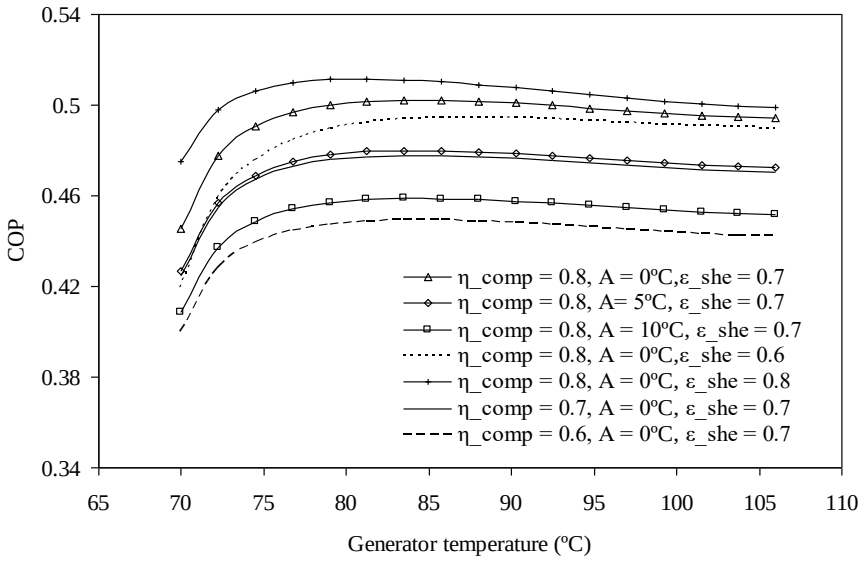
Figure 7. Variation of parameteric factors which constitute input exergy (Ta = Tc=35ºC, Tcc = 7ºC, effectiveness = 0.75, Qe_vcr = 100 kW, compressor efficiency= 0.8, Approach = 0 ºC)

Effect of Effectiveness of Solution Heat Exchanger, Approach in Cascade Condenser and Efficiency of Compressor in Low Temperature Stage

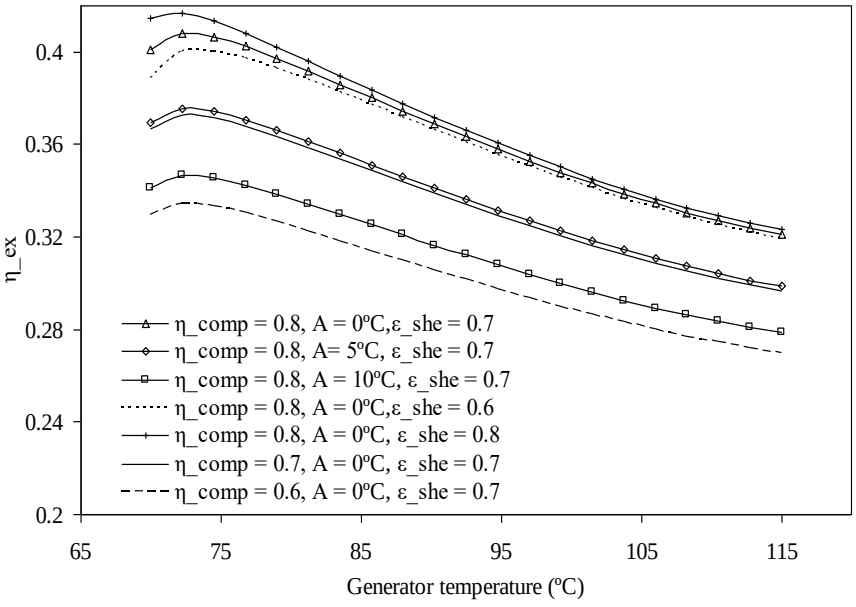
Figures 8 and 9 show the effect of effectiveness of solution heat exchanger, approach in cascade condenser and efficiency of compressor in low temperature stage on the COP and exergetic efficiency. The reduction in effectiveness of solution heat exchanger is responsible for decrease in COP and exergetic efficiency. The reduction in effectiveness of solution heat exchangers causes the temperature difference to increase between the strong solution exiting the solution heat exchanger and the generator temperature. Thus the heat duty and irreversibility in generator increase. Moreover at lower values of effectiveness of solution heat exchanger, the temperature difference between entering and exiting streams of strong and weak solutions is larger in comparison to when the effectiveness is higher. Thus at lower values of effectiveness, the irreversibility in solution heat exchanger is more. The above factors lead to increase the total exergy destruction in the cycle and the COP and exergetic efficiency reduce.

The increase in approach from 0ºC to 10ºC causes the absorber pressure to increase which increases the solution circulation ratio. The increase in solution circulation ratio increases the generator heat duty which brings down the COP. It also increases the irreversibility in generator and cascade condenser. Thus total exergy destruction increases and hence exergetic efficiency reduces.

The increase in isentropic efficiencies of compressor reduces the compressor power required and hence input energy and input exergy reduce thereby increasing the COP and the exergetic efficiency.

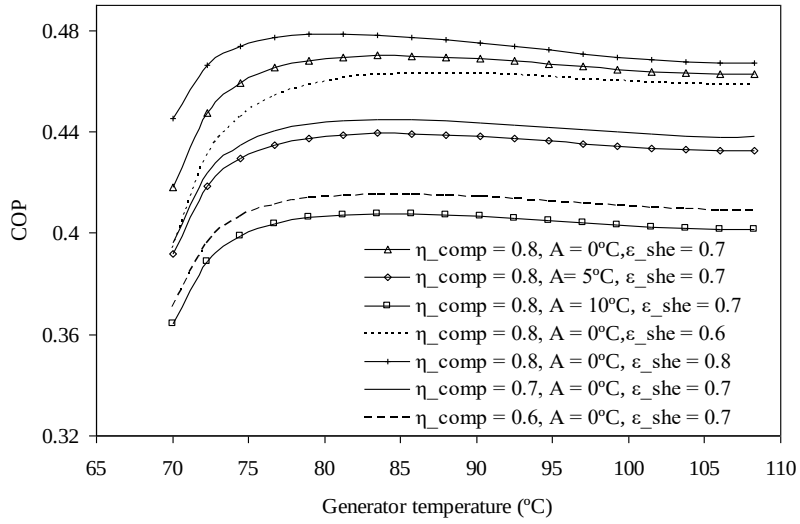


(a)

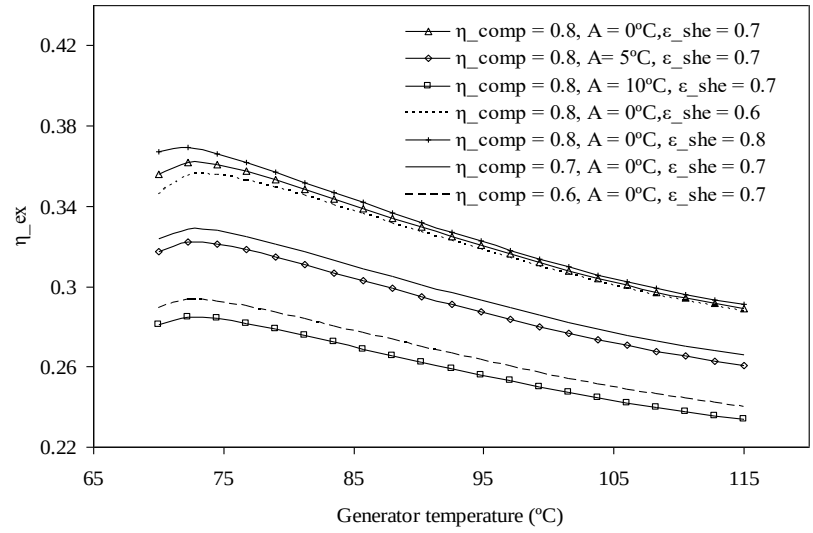


(b)

Figure 8. Effect of generator temperature on **(a)** COP and **(b)** exergetic efficiency for varying approach, effectiveness of 'she' and compressor efficiency (Refrigerant ammonia "R714", Ta = Tc = 35ºC, Tcc = 7ºC, Te_vcr = -40ºC, Qe_vcr = 100 kW)



(a)



(b)

Figure 9. Effect of generator temperature on **(a)** COP and **(b)** exergetic efficiency for varying approach, effectiveness of 'she' and compressor efficiency (Refrigerant carbon-dioxide "R744", Ta = Tc = 35ºC, Tcc = 7ºC, Te_vcr = -40ºC, Qe_vcr = 100 kW)

CONCLUSIONS

In the thermodynamic analysis of a compression absorption cascade refrigeration system, the following conclusions are drawn from it:

The COP and exergetic efficiency reduce with increase in cascade condenser temperature for carbon dioxide as refrigerant in compression system irrespective of approach temperature. The drop in COP is large at higher approach. In case of ammonia as refrigerant in compression system, the COP increases initially with increase in cascade condenser temperature, attains a maximum value and then decreases. However, the difference between maximum and minimum values of COP is very nominal at different approach. The exergetic efficiency decreases with increase in cascade condenser temperature.

- The maximum values of COP and exergetic efficiency occur corresponding to different generator temperatures. The maximum COP occurs at a generator temperature higher than the generator temperature corresponding to maximum exergetic efficiency irrespective of cascade condenser temperature and approach. The optimum generator temperature corresponding to maximum COP and maximum exergetic

efficiency reduce with increase in cascade condenser temperature. The optimum generator temperature corresponding to maximum COP and maximum exergetic efficiency for carbon dioxide cascade cycle and ammonia cascade cycle is same at similar variables but value of maximum COP and maximum exergetic efficiency of ammonia cascade cycle is higher in comparison to carbon dioxide cascade cycle.

- The COP decreases and exergetic efficiency increases with decrease in evaporator temperature for ammonia cascade cycle. It is significant that with increase in absorber temperature and condenser temperature i.e. at 40 ºC, both COP and exergetic efficiency increases, reaches to a maximum value and then decrease. Moreover, the optimum generator temperatures corresponding to maximum COP and maximum exergetic efficiency are also observed to shift towards higher values. Similar trends occur in case of carbon dioxide cascade cycle also. However the maximum values of COP and exergetic efficiency are lower for carbon dioxide cascade cycle in comparison to ammonia cascade cycle.
- Increase in approach in cascade condenser results in decrease in the COP and exergetic efficiency for both the refrigerants. The increase in effectiveness of solution heat exchanger increases the COP and exergetic efficiency marginally. The COP and exergetic efficiency increases significantly with the increase in isentropic efficiency of compressor.

Finally, it is concluded that VAR cycle of such cascade system can run using heat energy also from industrial waste heat. Thus, saving lots of electrical energy and by incorporating the appropriate parameters heat energy could also be saved. Results also predict that environment friendly refrigerant ammonia is a better refrigerant compared to carbon dioxide for low stage compression system in a compression absorption cascade refrigeration system.

NOMENCLATURE

A	Approach in Cascade condenser
CO_2	Carbon dioxide
COP	Coefficient of performance (dimensionless)
CRS	Compression refrigeration system
$\dot{E}D$	Exergy destruction (kW)
$\dot{E}D_{total}$	Total exergy destruction (kW)
$\dot{E}F$	Exergy of fuel (kW)
$\dot{E}P$	Exergy in product (kW)
H_2O	Water
HT	High temperature
Li-Br	Lithium bromide
LT	Low temperature
NH_3	Ammonia
P	Pressure
$\dot{Q}$	heat rate (kW)
s	Entropy (kJ/kgK)
T_g	Generator temperature (K)
T_0	Ambient temperature (K)
T_r	Refrigerant temperature (K)
VCR	Vapor compression refrigeration
VAR	Vapor absorption refrigeration
$\dot{W}_P$	Pump work (kW)
$\dot{W}_{COMP}$	Compressor work (kW)
X	Solution Concentration
Xs	Strong solution concentration
Xw	Weak solution concentration

GREEK LETTERS

ε_she	Effectiveness of solution heat exchanger
η_comp	Compressure efficiency

η_ex	Exergetic efficiecy

Subscripts

a	Absorber
c	Condenser
comp	Compressor
cc	Cascade Condenser
e	Evaporator
g	Generator
rtv_htc	Refrigerant thermostatic expansion valve of high temperature circuit
rtv_ltc	Refrigerant thermostatic expansion valve of low temperature circuit
she	Solution heat exchanger
stv	Solution thermostatic expansion valve

REFERENCES

[1] Chakravarthy, V.S., Shah, R.K., Venkatarathnam, G. (2011). A Review of Refrigeration Methods in the Temperature Range 4–300 K, J. Therm. Sci. Eng. Appl., 3, 02081.
[2] Cimsit, C., Ozturk IT. (2012). Analysis of compression-absorption cascade refrigeration cycle, Appl Therm Eng, 40, 311-317.
[3] Reddy, V.S., Kaushik, S.C., Tyagi, S.K. (2014). Exergetic analysis and evaluation of coal-fired supercritical thermal power plant and natural gas-fired combined cycle power Plant, Clean Techn Environ Policy, 16, 489–499.
[4] Reddy, V.S., Panwar, N.L., Kaushik, S.C. (2012). Exergetic analysis of a vapour compression refrigeration system with R134a, R143a, R152a, R404A, R407C, R410A, R502 and R507A, Clean Technologies and Environmental Policy, 14(1), 47-53.
[5] Seara, F., Sieres, J., Vazquez, M. (2006). Compression–absorption cascade refrigeration system, Applied Thermal Engineering 26, 502–512.
[6] Riffat, S.B., Shankland, N. (1993). Integration of absorption and vapor-compression systems, Applied Energy, 46(4), 303-316.
[7] Kairouani, L., Nehdi, E. (2006). Cooling performance and energy saving of a compression-absorption refrigeration system assisted by geothermal energy, Appl Therm Eng., 26, 288-94.
[8] Meng, X., Zheng, D., Wang, J., Li, X. (2013). Energy saving mechanism analysis of the absorption-compression hybrid refrigeration cycle, Renewable Energy, 57, 43-50.
[9] Colorado, D., Velazquez, V.M. (2013). Exergy analysis of a compression-absorption cascade system for refrigeration, Int. J. Energy Res, 37, 1851-65.
[10] Garimella, S., Brown, A.M., Nagavarapu, A.K. (2011). Waste heat driven absorption/vapor-compression cascade refrigeration system for megawatt scale, high-flux, low-temperature cooling, Int. J. Refrig., 34, 1776–1785.
[11] Seyfouri, Z., Ameri, M. (2012). Analysis of integrated compression-absorption refrigeration systems powered by a microturbine, International journal of Refrigeration, 35, 1639-1646.
[12] Wang, L., Aihua, M., Yingying, T., Xiaolong, Cui., Hongli, Cui. (2012). Study on Solar-Assisted Cascade Refrigeration System, International Conference on Future Energy, Environment, and Materials. Energy Procedia, 16, 1503 – 1509.
[13] Chen, Y., Han, W., Sun, L., Jin, H. (2015). A new absorption–compression refrigeration system using a mid-temperature heat source for freezing application, The 7th International Conference on Applied Energy – ICAE (2015), Energy Procedia, 75, 560 – 565.
[14] Arora, A., Dixit, M., Kaushik, S.C. (2016). Energy and exergy analysis of a double effect parallel flow LiBr/H_2O absorption refrigeration system, Journal of Thermal Engineering, 2(1), 541-549.
[15] Kutlu, Ç., Ünal, Ş., Erdinc, M. T. (2016). Thermodynamic analysis of bi-evaporator ejector refrigeration cycle using R744 as natural refrigerant, Journal of Thermal Engineering, 2(2), 735-740.
[16] Dixit, M., Arora, A., Kaushik, S. C. (2016). Energy and exergy analysis of a waste heat driven cycle for triple effect refrigeration, Journal of Thermal Engineering, 2(5), 954-961.
[17] Arora, A., Dixit, M., Kaushik, S.C. (2016). Computation of optimum parameters of a half effect water-lithium bromide vapour absorption refrigeration system", Journal of Thermal Engineering, 2(2), 683-692.

[18] Jain, V., Sachdeva, G., Kachhwaha, S.S. (2013). NLP model based thermo-economic optimization of vapour compression–absorption cascaded refrigeration system, Energy Conversion and Management, 93, 49–62.
[19] Jain, V., Kachhwaha, S.S., Sachdeva, G. (2013). Thermodynamic performance analysis of a vapour compression–absorption cascaded refrigeration system, Energy Conversion and Management, 75, 685–700.
[20] Jain, V., Sachdeva, G., Kachhwaha, S.S., Patel, B. (2016). Thermo-economic and environmental analyses based multi-objective optimization of vapor compression–absorption cascaded refrigeration system using NSGA-II technique, Energy Conversion and Management, 113, 230–242.
[21] Dixit, M., Kaushik, S.C., Arora, A. (2015). Energy and exergy analysis of absorption-compression cascade refrigeration system, Journal of Thermal Engineering, 1(1), 1-12.
[22] Kaushik, S.C., Arora, A., Bilga P. S. (2016). Alternatives in Refrigeration and Air Conditioning, I.K. International House Publishing,1/e, India.
[23] Megdouli, K., Tashtoush, B.M., Nahdi, E., Elankhdar, M., Mhimid, A., Kairouani, L. (2017). Performance analysis of a combined vapor compression cycle and ejector cycle for refrigeration cogeneration, International journal of refrigeration, 74, 517–527.
[24] Patel, B., Desai, N.B., Kachhwaha, S.S. (2017). Thermo-economic analysis of solar-biomass organic Rankine cycle powered cascaded vapor compression-absorption system, Solar energy, 157, 920-933.
[25] Chen, Y., Han, W., Jin, H. (2017). Proposal and analysis of a novel heat driven absorption-compression refrigeration system at low temperatures, Applied Energy, 185, 2106-2116.
[26] Yingjie, Xu., Jiang, N., Qin Wang, Q., Han, X., Gao, Z., Chen, G. (2017). Proposal and thermodynamic analysis of an ejection-compression refrigeration cycle driven by low-grade heat, Energy Conversion and Management, 145, 343-352.
[27] Yingjie, Xu., Jiang, N., Pan, F., Wang, Q., Gao, Z., Chen, G. (2017). Comparative study on two low-grade heat driven absorption-compression refrigeration cycles based on energy, exergy, economic and environmental (4E) analyses, Energy Conversion and Management, 133, 535-347.
[28] Klein, S.A., Alvarado, F. (2005). Engineering Equation Solver Version 7.441, F-Chart software Middleton: WI.
[29] Pa.´tek, J., Klomfar, J. (2006). A computationally effective formulation of the thermodynamic properties of water–lithium bromide solutions from 273K to 500 K over full composition range, International Journal of Refrigeration, 29, 566–578.

2020 IEEE International Conference for Innovation in Technology (INOCON)
Bengaluru, India. Nov 6-8, 2020

PLC and SCADA based Industrial Automated System

Basant Tomar
Department of Electrical Engineering
Delhi Technological University
Delhi-110042, India
basanttomar18@gmail.com

Narendra Kumar
Department of Electrical Engineering
Delhi Technological University
Delhi-110042, India
ndeshwalus@gmail.com

***Abstract*— In every process industry, there is a repetition of the tasks. So, in order to get the better productivity in limited time with lesser error probability, we are incorporating Automation into the process industries. In recent trends of automation like Industry 4.0, manual operations are being replaced with fully or semi auto-controlled, reconfigurable operations by incorporating more advanced technologies. Primitive methods of filling of different types of liquids in a fixed proportion in different containers or bottles involve manual operations. This hampers production rate and sometimes quality of product consistency due to the involvement of human intervention. In this paper, a simulated prototype of SCADA (Supervisory Control and Data Acquisition) and PLC (Programmable Logic Controller) based process of bottle filling in beverage industry is designed to fill the different types of containers with different composition of liquids. The entire simulated operation involves distinct stages of sorting the containers, filling liquids, capping, labelling of different containers and counting the units per batch. The Metallic base and Non-Metallic bottles are sorted by assistance of Inductive Proximity Sensor. During filling stage, the Metallic base bottle is filled with each of the available three different liquids composition equally and the Non-Metallic bottle is filled only with two initial type of liquids composition. In later stages, bottles are labelled separately. Both type of bottles are detected with the help of Photo-electric Retro-reflective type sensors at Filling, Capping and Labelling stages.**

Keywords*— *Industry 4.0, Automated Process, PLC (Programmable Logic Controller), Variable Frequency Drive, SCADA (Supervisory Control and Data Acquisition), IPR (Inductive Proximity sensor).

I. INTRODUCTION

Automation entail controlling of any operation with very little or no human intervention. This is the main reason that industries are opting for automated processes which provide flexible and cost-effective operations with competitiveness and efficiency. Traditionally, the industrial processes were conventional and mainly operated manually which made them time consuming and not so cost effective. But with the admittance of automation into industries, all such concerns were rectified. Automation also curtails the probable errors, reoccurring in operations which ultimately results in conservation of energy and resources.

Industry 4.0, the latest scenario in automation, has been developed to offer various leverages like decentralization of process, virtualization, interoperability, real-time data acquisition etc. there is not a unique way of describing or defining Industry 4.0 [1]. It may be perceived as a comprehensive way of combining complex machineries, sensors, actuators and software interlinked in a network for predicting, organizing and utilizing the control outcomes for business processes.

The purpose here is to visualize an automated process industry bottle filling operation in SCADA system and regulation of total operation is carried out by the assistance of PLC. In this operation, an OMRON (NX1P2-9024DT1) PLC is used and ladder logic is developed in OMRON's Sysmac Studio Programming Software. The visualization of the operation is implemented in Wonderware Intouch SCADA software.

PLCs are majorly used in automated process control systems for being simple, easily programmable, less time consuming as compared to implement replace hundreds of hard-wired controllers like relays, timers etc. [2]-[4] and handle multiple inputs and outputs; immune to electrical noises and also provide resistance against vibrations and impacts.

The entire paper is organized into four sections which includes Introduction; Prototype of Automated Beverage Industry Operation; Operation followed by Results and Discussions.

II. PROTOTYPE OF AUTOMATED BAVEREGE INDUSTRY OPERATION

Almost in every beverage industry, filling of different liquids process is the most commonly practiced [5]-[7]. All industries concerned with beverages are designed in order to keep drinks of different material compositions into different container types. The whole container filling and processing is partitioned into following different stages

- Sortation stage: The containers or bottles that are moving over the conveyor belt will be sorted as Metallic base container & Non-Metallic container. The sorting is carried out by the assistance of (IPR) Inductive proximity IME08 sensor. This sensor is capable to sense the metallic things only and ignore all other material type like non-metallic container
- Liquid filling stage: Filling of containers with different materials compositions. Metallic base containers will be having three liquids composition with 33% each but the Non-Metallic containers will be having only two liquid compositions with 33% each. Photo-electric Retro reflective sensor will be used in this stage along with valves.
- Capping Stage: As soon as filling of containers is completed, capping work unit will come into picture to apply caps and tightens them over filled one

978-1-7281-9744-9/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:06:18 UTC from IEEE Xplore. Restrictions apply.

container. Here also Photo-electric retro reflective sensor will be used.

- Labelling Stage: Caping of the containers is followed by the labelling work unit. For differentiating containers, we will be using labels of different colors for particular container type. Here also Photo-electric retro reflective sensor will be used for detecting container.
- Counting Stage: Counting process of metallic and non- metallic containers at the end in each and every case.

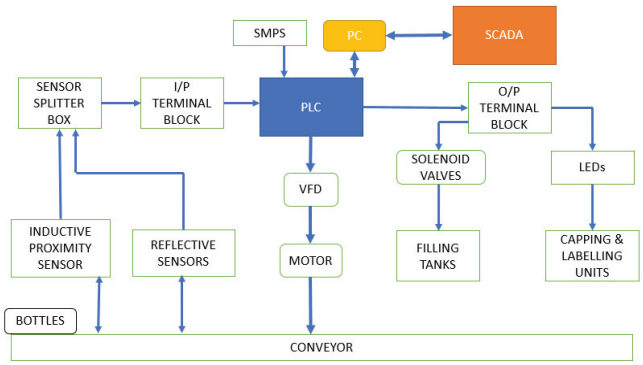


Fig. 1. Framework

The communication between SCADA & PLC is via PC through wired or wireless media. The PLC needs regulated power supply of 24V DC to perform its functions. SMPS (switched mode power supply) provide this required voltage to PLC.

Programmable Logic Controller is the center of the entire process. Input data or signals from input devices like sensors is feed into PLC, input sensor's information is processed, logical operations will be performed over inputs and a desired response is provided for the output connected devices at the output terminal block such as solenoid valves to filling tanks, LEDs to cap & label units, VFD to motor to conveyor belt. Conveyor system motion is governed by a DC motor, which get driven by Drive in order to provide precise and accuracy in motion. Both type of Containers will be sensed by Inductive Proximity sensor IPR and Photo-electric Retro-reflective sensors present over the conveyor. Both type of sensors are linked with a box called sensor splitter, which indicates the status of all sensors. Then infeed signal reaches the PLC input ports or terminals. Output devices are linked with PLC via output ports or terminals. Different type of liquid is poured into both type of containers through valves connected to the filling tanks. Labelling & Capping Units are also controlled with Reflective sensors signal conditioning to the PLC

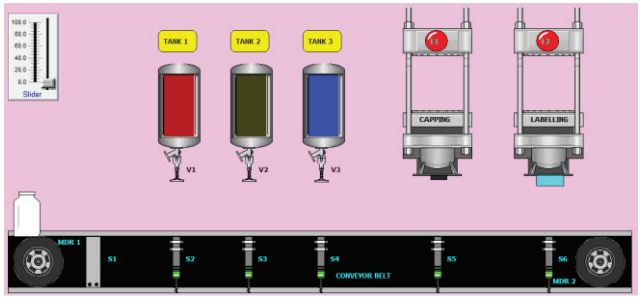


Fig. 2. Process Model in SCADA Wonderware Intouch software

TABLE I. COMPONENTS INVOLVED IN THE OPERATION

Input Components	Output Components	Additional Components
One Inductive Proximity Sensor (S1); Five Photo-electric Retro-reflective type sensors (S2-S6)	Three Valves (V1-V3); Conveyor belt motor; Two LEDS (L1, L2) for capping process unit and labelling process unit.	One PLC unit; Conveyor system; Two motor dependent or driven rollers; Three Tanks for different liquid types; Units for capping Unit for labelling

III. OPERATION

The whole system operation is demonstrated with the help of flow charts discussed below:

In this flow chart functioning of the conveyor belt motor is depicted. Accuracy in the dynamics of the conveyor belt is one of the keenest factors of this operation. The conveyor belt mobility is governed or we can say totally influenced by input signals sent by the sensors (IR sensors and reflective sensors) and output signal send by the Drive. Inductive Proximity Sensor (IPR S1) does not directly influence the stopping and motion of the conveyor belt but whenever either of five Photoelectric Retro Reflective type sensors (S2, S3, S4, S5, S6) senses any of the bottles, metallic base or non-metallic then the conveyor belt motor stops its motion and resumed after a fixed time interval when either of the processes (bottle filling, capping of bottles and labelling of the bottles) is finally done.

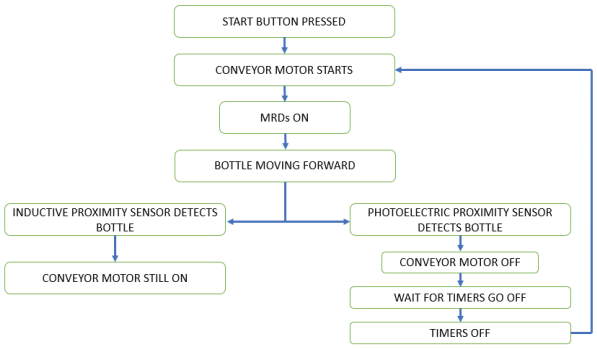


Fig. 3. Conveyor Belt Functioning

In this second flow chart, the entire operation of sorting of containers and filling of each type of containers is depicted. When any type of container is moving on the conveyor belt, then Inductive proximity sensor (IPR S1) will decide the type of the container, whether it is metallic or non-metallic. If the presence of metallic container is there, then the Inductive PR Sensor S1 turns green, else there will be no change in PR sensor in case of non-metallic container. After that, both type of containers moves further and in later stage, they are detected by the Photoelectric retro- reflective type sensor (S2), which results into the opening of solenoid valve (V1) and as a result of this the tank1 starts filling the liquid into the container and this task lasts till the 33% or approximately one-third of the container gets filled. Then timer of the tank1 goes OFF and the motion of the conveyor belt starts again and continues till the next upcoming Photoelectric retro-reflective type sensor (S3) detects the container, which results into the opening of solenoid valve (V2) and as a result of this the tank2 starts filling the liquid into the container and this task lasts till the 66% or

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:06:18 UTC from IEEE Xplore. Restrictions apply.

approximately two-third of the container gets filled. Then timer of the tank2 goes OFF and motion of the conveyor belt starts again and continues till the next upcoming Photoelectric retro-reflective type sensor (S4) detects the container, which results into the opening of solenoid valve (V3) and as a result of this the tank3 starts filling the liquid into the container and this task lasts till the 99% or approximately full container gets filled. But if the container that was not detected by the IPR S1 then as a action of that the solenoid valve (V3) will not open and as a result the tank 3 will not fill the non-metallic container in this scenario. As a consequence of this, the non-metallic container will be filled by 66% or two-third of the liquid in the entire process.

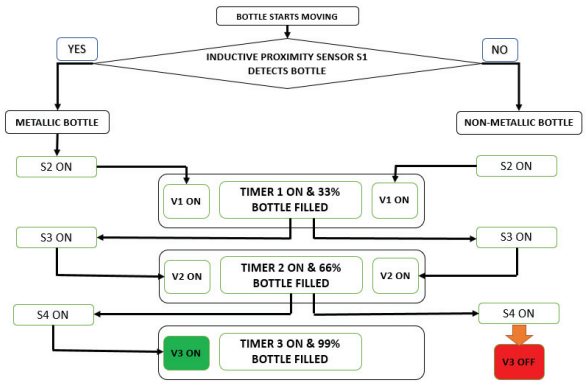


Fig. 4. Sorting and Filling

Next flow chart shows the process of capping of the containers along with the process of labelling of the containers. When Photoelectric retro-reflective sensor (S5) detects any type of the containers then the LED L1 of the unit of capping turns green and capping initiated to put the cap over the containers. During this entire process the conveyor belt motion is stopped. And after the competition of the process the motor resumes again. Now when Photoelectric retro-reflective sensor (S6) detects any type of the containers then the LED L2 of the unit of labelling turns green and labelling initiated to put the labels over the containers. During this entire process the conveyor belt motion is stopped. And after the competition of the process the motor resumes again

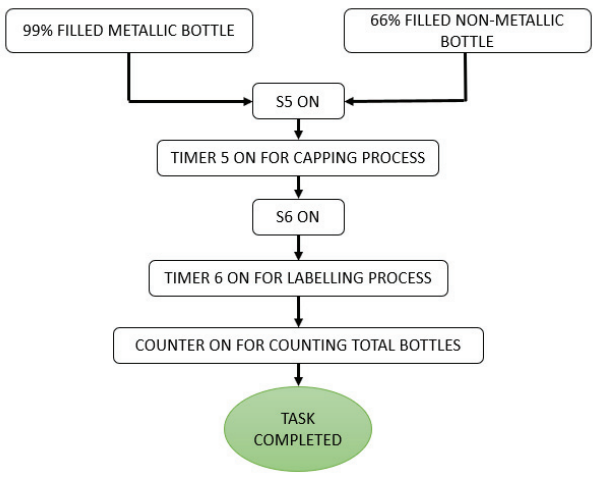


Fig. 5. Capping and Labelling

IV. RESULTS AND DISCUSSION

1. Inductive Proximity (IPR) sensor (S1) is able to identify metallic base bottle and bypass others.

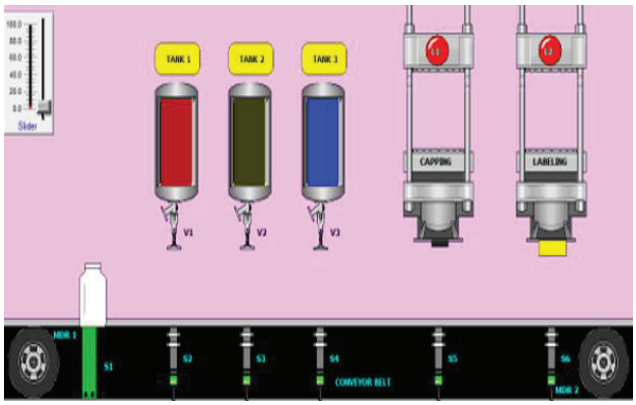


Fig. 6. Sorting

2. (S2) Photo-electric retro-reflective sensor detect both containers, conveyor motor stops, (V1) valve opens and containers got filled one-third.

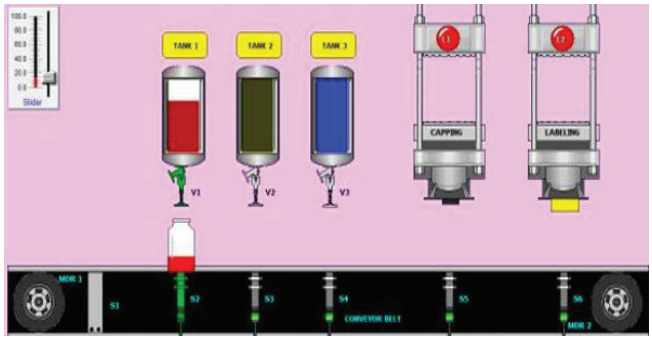


Fig. 7. Tank1 Level

3. (S3) Photo-electric retro-reflective sensor detect both containers, conveyor motor stops, (V2) valve opens and containers got filled two-third.

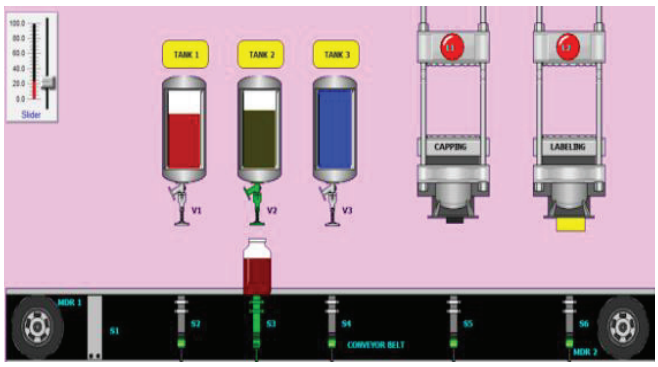


Fig. 8. Tank2 Level

4. (S4) Photo-electric retro-reflective sensor detect both containers, conveyor motor stops but (V2) valve opens only in the case of metallic base container (when S1 is high) and it gets completely filled but in case of non-metallic container, S1 is low, (V2) valve does not open and container remains filled two-third.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:06:18 UTC from IEEE Xplore. Restrictions apply.

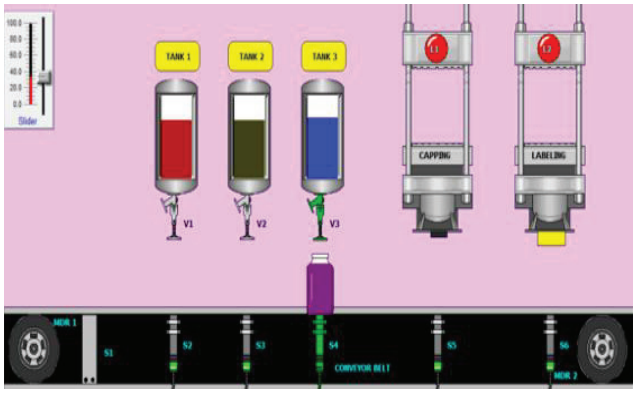


Fig. 9. Tank3 Level (Metallic Base Bottle)

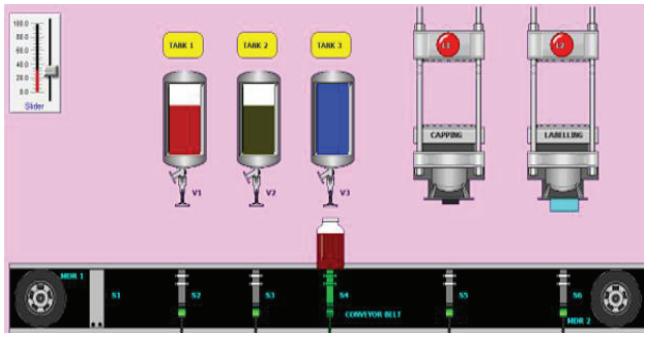


Fig. 10. Tank3 Level (Non-Metallic Bottle)

5. (S5) Photo-electric retro-reflective sensor detect both containers, conveyor motor stops, (L1) LED turns green and capping unit applies cap to both containers.

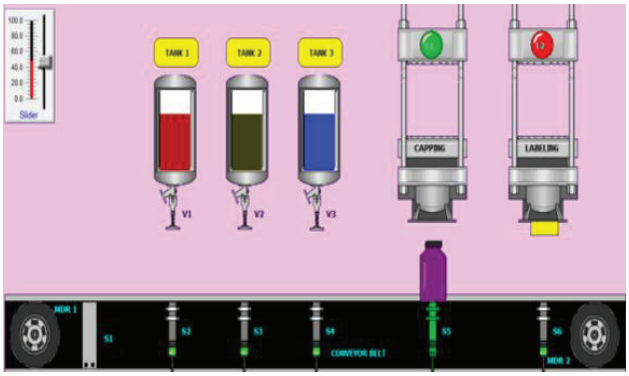


Fig. 11. Capping (Metallic Base Bottle)

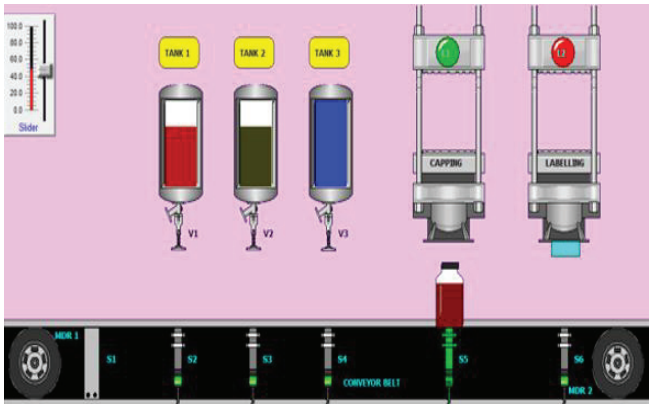


Fig. 12. Capping (Non-Metallic Bottle)

6. (S6) Photo-electric retro-reflective sensor detect both containers, conveyor motor stops, (L2) LED turns green and labelling unit applies cap to both containers but with different colors.

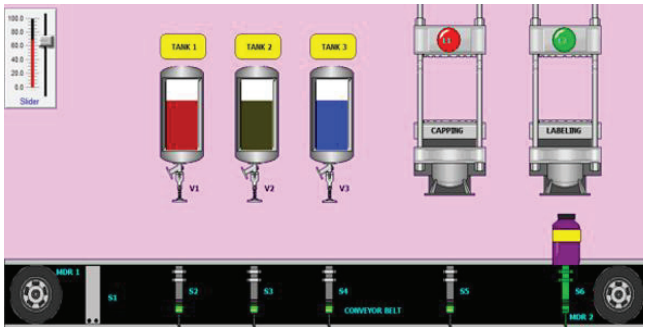


Fig. 13. Labelling (Metallic Base Bottle)

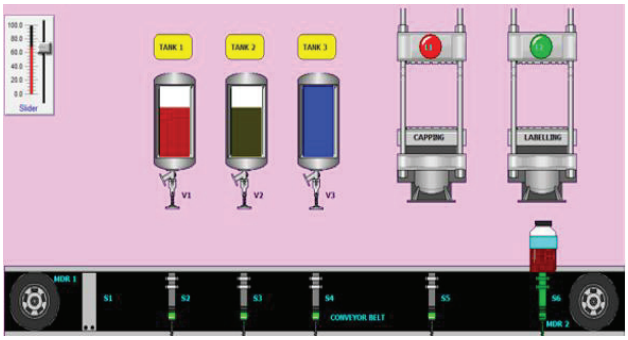
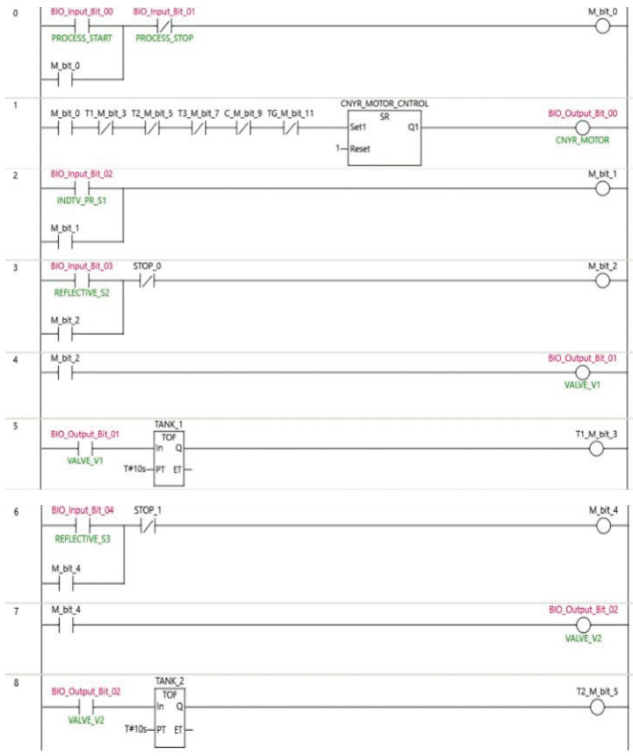


Fig. 14. Labelling (Non-Metallic bottle)

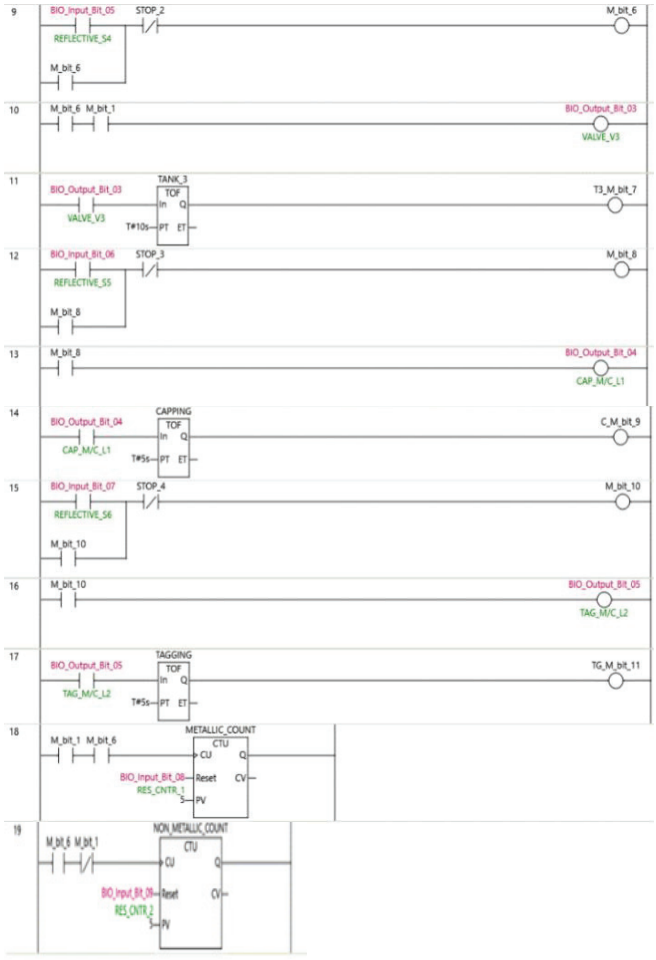
APPENDICES

Appendix A:



Appendix B:
Continued…

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:06:18 UTC from IEEE Xplore. Restrictions apply.



REFERENCES

[1] Kahiomba Sonia Kiangala and Zenghui Wang (2019), "An Industry 4.0 approach to develop auto parameter configuration of a bottling process in a small to medium scale industry using PLC and SCADA", Procedia Manufacturing Volume 35, 2019, Pages 725-730

[2] Htet Htet Aung, Thae Thae Ei Aung (2019), "Simulation and Implementation of PLC based for Nonstop Filling Process using PLCSIM and HMI", International Journal of Creative and Innovative Research in All Studies, Volume 2, Issue 3

[3] Zar Kyi Win, Tin Tin New (2019), "PLC Based Automatic Bottle Filling and Capping System", International Journal of Trend in Scientific Research and Development (IJTSRD), Volume 3 Issue 6, October 2019

[4] E. R. Alphonsus & M. O. Abdullah (2016), "A review on the applications of programmable logic controllers (PLCs)", Renewable and Sustainable Energy Reviews, 60, 1185–1205

[5] N. Nadgauda, Senthil Arumugam Muthukumaraswamy, S. U. Prabha (2020), "Smart Automated Processes for Bottle-Filling Industry Using PLC-SCADA System", Chapter in Intelligent Manufacturing and Energy Sustainability, pp.693-702

[6] S. V. Viraktamath, A. S. Umarfarooq, Vinayakgouda Yallappagoudar and Abhilash P. Hasankar (2020), "Implementation of Automated Bottle Filling System Using PLC", Chapter in Inventive Communication and Computational Technologies, pp.33-41

[7] K. Chakraborty, M.G. Choudhury, S. Dasc and S. Paul (2020), "Development of PLC-SCADA based control strategy for water storage in a tank for a semi-automated plant", Published by IOP Publishing for Sissa Medialab, Journal of Instrumentation, Volume 15, April 2020

[8] Nadgauda, N., & Muthukumaraswamy, S. A. (2019), "Design and Development of Industrial Automated System using PLC-SCADA", 2019 IEEE 10th GCC Conference & Exhibition (GCC).

[9] Mini Sreejeth and Shilpa Chouhan (2016), "PLC based Automated Liquid Mixing and Bottle Filling System", 1st IEEE International Conference on Power Electronics. Intelligent Control and Energy Systems (ICPEICES-2016)

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:06:18 UTC from IEEE Xplore. Restrictions apply.

Journal of Polymer Research (2021) 28: 7
https://doi.org/10.1007/s10965-020-02362-4

ORIGINAL PAPER

Polyacrylonitrile/clay nanofibrous nanocomposites for efficient adsorption of Cr (VI) ions

Priya Bansal[1] · **Roli Purwar**[1]

Received: 26 July 2020 / Accepted: 26 November 2020 / Published online: 5 January 2021
© The Polymer Society, Taipei 2021

Abstract

This paper highlights the efficient adsorption of Cr (VI) from wastewater using polyacrylonitrile (PAN)/montmorillonite (Mt) and polyacrylonitrile/zinc oxide nanoparticle-modified montmorillonite (ZnO-Mt) nanofibrous nanocomposites as adsorbents. Mt was modified using zinc oxide nanoparticles (ZnO) by an adsorption process. The particle size of ZnO-Mt was found to be 18 ± 2 nm. Field-emission scanning electron microscopy (FESEM) and high-resolution transmission electron microscopy (HRTEM) images showed the presence of ZnO on and within the layers of Mt. Mt and ZnO-Mt were used to prepare PAN/clay nanofibrous nanocomposites which were further used as adsorbents for Cr (VI) ions. X-ray diffraction (XRD) analysis showed that Mt and ZnO-Mt were exfoliated in the PAN nanofibrous matrix. Mt, ZnO-Mt, PAN/Mt and PAN/ZnO-Mt nanofibrous nanocomposites were analyzed as adsorbents for the adsorption of Cr (VI) ions. The effect of pH, concentration, time and amount of adsorbent on the removal of Cr (VI) ions from water were investigated. Adsorption of Cr (VI) ions over the adsorbents follows pseudo-second-order kinetics, and Langmuir and Freundlich isotherms. PAN/ZnO-Mt and PAN/Mt nanofibrous nanocomposites were found to be more effective adsorbents for Cr (VI) ions compared to ZnO-Mt and Mt.

Keywords Montmorillonite · Zinc oxide nanoparticle-modified montmorillonite · Adsorption · Cr (VI) ions · Nanofibrous nanocomposite

Introduction

One of the major challenges currently is the availability of clean drinking water. Non-biodegradable toxic metal ions such as chromium (Cr) released from industries, mainly in its hexavalent state, are injurious to human health and aquatic environments when their tolerance limits are exceeded in water bodies. Cr (VI) is a highly toxic metal ion due to its carcinogenicity, mutagenicity and teratogenicity in living organisms and is 100 times as toxic as in its trivalent state. According to US Environmental Protection Agency (USEPA) regulations, the maximum concentration of chromium must be lower than 100 µg/L, while the World Health Organization (WHO) suggests a maximum of 50 µg/L [1].

Adsorption is one of the most commonly used techniques for the removal of heavy metal ions and dyes from wastewater. Adsorption is preferred over other methods such as liquid–liquid extraction, membrane filtration, ion exchange, coagulation, electrolysis and precipitation due to its higher efficiency, energy savings, broad operational range of pH and temperature, easy recycling of adsorbent, and selectivity at lower contaminant concentrations [2, 3]. Various hydrogels [4–7], biopolymers [8–10], microspheres [11] and nanofibrous membranes [12, 13] have been used for adsorption of metal ions. Adsorption of metal ions through montmorillonite (Mt) clay and its modified forms has been widely studied in the past two decades. Mt has a 2:1-layer structure consisting of an octahedral alumina sheet sandwiched between two opposing tetrahedral silica sheets. The bonding between the two silica sheets is very weak, which permits water and exchangeable ions to enter. The efficiency to exchange cations, high swelling ability, mechanical and chemical stability, and dispersion properties make them suitable adsorbent for metal ion removal. The adsorption efficiency of clay can be improved by modifying it with organic

✉ Roli Purwar
roli.purwar@dtu.ac.in

[1] Discipline of Polymer Science and Chemical Technology, Department of Applied Chemistry, Delhi Technological University, Shahbad, Daulatpur Delhi-110042, India

materials and nanoparticles through intercalation, pillaring, acid/base activation and organo-functionalization [14, 15].

Zinc oxide (ZnO) nanoparticles having varying morphologies show unique properties such as photocatalytic [16–19], photoluminescence [20], optical [18, 21] and antimicrobial activity [22–25], and anticancer drug delivery [26, 27]. The efficiency of ZnO nanoparticles can be enhanced by addition of transition metals (such as Mn, Fe, Cu) and biomolecules, which has expanded its use in biosensors, antimicrobial agents, antioxidants, water disinfectants, drug delivery systems and bio-imaging materials [28, 29]. Metal oxide nanoparticles, especially ZnO-supported Mt, provide reactive absorbance with high surface area, improved chemical reactivity and display added physicochemical functions of both the metal oxide nanoparticles and Mt [19, 30]. ZnO/clay nanocomposites have been studied as good adsorbents for metal ion removal [31] and can photocatalytically degrade cationic and anionic dyes [32] and pharmaceuticals present in wastewater [33]. The major constraint of simple clay minerals is the aggregation of the platelets under use conditions, reducing the surface area and therefore the removal efficiency. Clay/polymer composites have improved the adsorption capacity of bare clay minerals [34].

Electrospinning is one of the widely used technologies to produce polymeric fibers in nanometer to micrometer range using electrostatic force. The electrospun nanofibers, due to their smaller pore size, higher surface area and controllable pore structure, are used as adsorbents for heavy metal ions from aqueous solutions [35]. The nanofibrous membranes have been investigated as adsorbents for heavy metal ions such as Cu (II), Ni (II), Pb (II), Fe (II), U (VI), Cr (III) and Ag (I) [36–38].

Electrospun polyacrylonitrile (PAN) nanofibrous membranes can be used for removal of heavy metal ions from wastewater due to their higher thermal and chemical resistance and good mechanical properties [39]. However, it has been observed that the pure nanofibrous membranes are not well suited for removal of heavy metal ions from wastewater. Thus, studies have been conducted on surface modification of PAN nanofibrous membranes to increase the adsorption capacity. The fabrication of ZnO nanoparticle on the surface of PAN nanofibers increases the adsorption efficiency of Pb (II) and Cd (II) metal [12].

The adsorption capacity of clay and ZnO-modified clay depends on the surface area available for adsorption of metal ions. If clay breaks down from a stacked structure to a two-dimensional nanostructure, a larger surface will be available for the adsorption of metal ions. Through electrospinning, clay and ZnO-modified clay can be exfoliated. In our previous study, we observed that clay was completely exfoliated in polymer/clay nanocomposite nanofibrous mats [40]. Thus, the incorporation of ZnO-modified clay in nanofibrous membranes can enhance the adsorption capacity. With the above hypothesis, Mt was modified with ZnO nanoparticles. ZnO-Mt was characterized by a particle size analyzer, field-emission scanning electron microscopy (FESEM) and X-ray diffraction (XRD). Both Mt and ZnO-Mt were incorporated in PAN and electrospun in the form of nanofibrous nanocomposites. The structural and morphological characterization of the PAN/Mt and PAN/ZnO-Mt nanofibrous nanocomposites was carried out using XRD and FESEM, respectively. The adsorption behavior of the PAN/ZnO-Mt and PAN/Mt nanofibrous nanocomposites for Cr (VI) ions was evaluated and compared with ZnO-Mt and Mt.

Experimental

Materials

Zinc oxide nanopowder having an average particle size of approximately 30 nm and clay nanopowder (sourced from montmorillonite clay, composition: SiO_2 65%, Al_2O_3 15%, Fe_2O_3 1.3%, Ti_2O 0.19%, K_2O 1.3%, Na_2O 0.65%, CaO 3.31%, MgO 0.99%) having an average particle size of 80–150 nm and cation exchange capacity of 3–10 meq/100 g were purchased from SRL, India. Potassium dichromate ($K_2Cr_2O_7$) used for adsorption studies was purchased from Merck Specialties. PAN with average molecular weight of 150,000 was purchased from Sigma-Aldrich. *N*,*N*-Dimethylformamide (DMF) was purchased from Merck. Distilled water was used to carry out the experiment.

Methods

Preparation of zinc oxide nanoparticle-modified montmorillonite (ZnO-Mt)

To prepare ZnO-Mt, 3 g of Mt was added to 30 ml of 1 mM ZnO nanoparticle dispersion in water and stirred constantly for 1 week at room temperature. The dispersion was ultrasonicated and centrifuged. The supernatant solution was discarded, and the clay was dried at 80 °C. The dried sample was ground using a mortar and pestle and characterized.

Preparation of nanofibrous nanocomposites

PAN/Mt and PAN/ZnO-Mt nanofibrous nanocomposites were prepared using electrospinning. PAN dope solution (7% w/v) was prepared by dissolving PAN powder in DMF at 70 °C with continuous stirring for 30 min until a homogeneous solution was obtained. Mt and ZnO-Mt (0.75% w/w wrt PAN) were ultrasonicated for 30 min and added to the dope solution, resulting in PAN/Mt and PAN/ZnO-Mt solutions, respectively. PAN/Mt and PAN/ZnO-Mt solutions

were electrospun at a voltage of 18 kV and flow rate of 6 ml h^{-1}, and the distance between the collector and spinneret was 12 cm. The thickness, average pore size, grams per square meter (GSM) and average diameter of the PAN/Mt nanofibrous nanocomposites were found to be 0.074 mm, 0.134 µm, 4.25 gm^{-2} and 210 ± 30 nm, respectively, while the thickness, average pore size, GSM and average diameter of the PAN/ZnO-Mt nanofibrous nanocomposites were found to be 0.125 mm, 0.158 µm, 6.475 gm^{-2} and 237 ± 40 nm, respectively.

Characterization

The particle size of the ZnO-Mt was determined using a particle size analyzer (Microtrac S3500). Aqueous solution of ZnO-Mt was added to the quartz cell of the analyzer. The size of the particles was analyzed at 90° to the incident light. The surface morphology of the ZnO-Mt and nanofibrous nanocomposites was analyzed using a field-emission scanning electron microscope with an environmental FEI Quanta 200F scanning electron microscope with an Oxford EDS detector (IE 250X Max 80, Netherlands). For analysis, the surface was coated with platinum. A transmission electron microscope (JEOL JEM-2100F) was used to analyze the morphology of the ZnO-Mt. A drop of sonicated aqueous solution of ZnO-Mt was cast on a carbon-coated copper HRTEM grid. The exfoliation of Mt with ZnO nanoparticles and clay with PAN was studied using XRD (D8 Advance, Bruker) by varying 2θ from 5° to 40°. The BET surface area was analyzed using Quantachrome ASiQwin software. Thermogravimetric analysis (TGA) was carried out on a Perkin Elmer TGA instrument by varying the temperature from 30 °C to 900 °C at a heating rate of 10 °C min^{-1} under a nitrogen gas flow rate of 20 ml min^{-1}.

Point of zero charge

A solid addition method [4, 41] was used to analyze the zero surface charge of the adsorbents, Mt, ZnO-Mt, PAN/Mt nanofibrous nanocomposites and PAN/ZnO-Mt nanofibrous nanocomposite. Closed conical flasks containing 50 ml of KNO_3 solution were taken, and the pH of the solution was varied from 2 to 8 using 0.1 M HCl and 0.1 M NaOH solution. In each solution, 0.2 g of adsorbent was added. The flasks were sealed and stirred at 150 rpm for 24 h, and the final pH of the solution was recorded.

Adsorption of Cr (VI) ions

PAN/Mt nanofibrous nanocomposites, PAN/ZnO-Mt nanofibrous nanocomposites, ZnO-Mt and Mt were tested for adsorption of Cr (VI) ions. Stock solution of $K_2Cr_2O_7$ (200 ppm) was used to prepare diluted solutions of 50 ppm to 150 ppm for adsorption study at $\lambda_{max} = 350$ nm. The minimum detectable concentrations for Cr (VI) ions in a UV–visible spectrophotometer were used to study the adsorption process of metal ions. The effect of pH, initial concentration, time and amount of adsorbent were analyzed. The experiments were performed in triplicate with a known amount of adsorbent (Mt and ZnO-Mt); i.e., 60 mg of the adsorbent was added to 25 ml of 100 ppm $K_2Cr_2O_7$ solution at pH 4.75. In the case of the nanofibrous nanocomposite, 3 mg of adsorbent was added to 25 ml of 100 ppm $K_2Cr_2O_7$ solution at pH 4.75. The solutions were uniformly stirred in a magnetic shaker bath for 120 min to attain equilibrium. The solutions were further filtered, and the amount of metal ions adsorbed was analyzed using a UV–visible spectrophotometer. The percentage removal of metal ions and adsorption capacity at equilibrium, q_e (mgg^{-1}), were calculated using the following equations

$$\% \, Removal = \frac{C_0 - C_e}{C_0} 100 \tag{1}$$

$$q_e = \frac{(C_0 - C_e)V}{M} \tag{2}$$

where C_0 and C_e are the initial and equilibrium concentration of metal ions in parts per million, respectively, V is the volume of the solution and M is the mass of the adsorbent (g).

Results and Discussion

Characterization of ZnO-Mt

A diffraction peak at 2θ equal to 7.12° for corresponding d-spacing of 1.24 nm appears in the X-ray diffractogram of Mt (Fig. 1). In the case of ZnO-Mt, the X-ray diffractogram shows a minor shift in the 2θ value. The 2θ value shifts to 7.32° (d spacing = 1.21 nm) from 7.12°. Broadening of the peak appears in ZnO-Mt in comparison to Mt, which indicates that the ZnO nanoparticles are inserted within the layers of Mt with a negligible shift in 2θ value. The particle size of Mt and ZnO nanoparticles was analyzed using a particle size analyzer. The average particle size of Mt was found to be 95 nm. On modification of Mt with ZnO nanoparticles, the size was reduced to 17 nm. Hadjltaief et al. [32] and Akkari et al. [19] calculated the size of ZnO on clay using the Scherrer equation, and it was found to be around 13.65 nm and in the range of 7 to 10 nm, respectively.

Mt was observed to be of a flaky shape under a scanning electron microscope (Fig. 2a). The cuboid-shaped ZnO nanoparticles are uniformly distributed over the surface of Mt (Fig. 2b). This change in the surface of Mt demonstrates

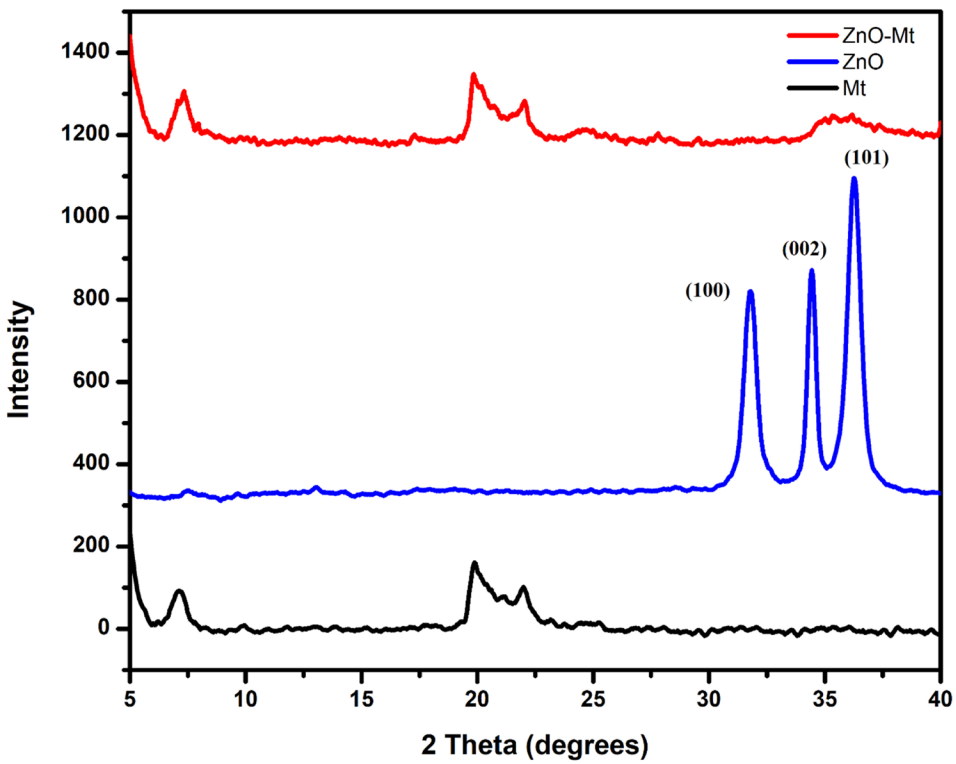


Fig. 1 XRD diffractogram of Mt, ZnO nanoparticle and ZnO-Mt

adsorption of ZnO nanoparticles over Mt surface. The morphology of ZnO-Mt was further analyzed using high-resolution transmission electron microscopy (HRTEM). The HRTEM image of ZnO-Mt (Fig. 3) shows that the ZnO nanoparticles are uniformly present within the layers of Mt. Based on XRD and HRTEM results, it was observed that on modification of Mt by ZnO nanoparticles, the spacing between the layers of Mt decreases, and the ZnO nanoparticles are inserted between the layers and are also uniformly distributed over the layers of Mt (Fig. 4). This insertion of ZnO nanoparticles over the surface and between the layers of Mt results in a change in the size of ZnO-Mt in comparison to ZnO nanoparticles and Mt. Akkari et al. [19] observed homogeneous distribution of ZnO nanoparticles prepared in situ along sepiolite fibers. Sani et al. [31] demonstrated a flower-like shape of a ZnO/Mt nanocomposite.

Characterization of PAN/clay nanofibrous nanocomposites

The nanofibrous nanocomposites show a bead-less morphology with a rough surface under FESEM (Fig. 4). The PAN/ZnO-Mt nanofibrous nanocomposites have a rougher surface than the PAN/Mt nanofibrous nanocomposites. The average diameter of the PAN/Mt nanofibrous nanocomposites was found to be 210 ± 30 nm, while the average diameter of the PAN/ZnO-Mt nanofibrous nanocomposites was found to be 237 ± 40 nm. Higher-magnification FESEM images of the PAN/Mt and PAN/ZnO-Mt nanofibrous nanocomposites suggest a porous structure of the nanofibrous nanocomposites. The porous structure of the nanofibers is responsible for efficient adsorption of metal ions over the surface of the membranes. Haddad and Alharbi [12] have prepared PAN/ZnO nanofibrous mats with average diameter 240 ± 25 nm. They also observed that incorporation of ZnO increases the surface roughness of PAN nanofibers.

The exfoliation or intercalation of clay into nanofibrous matrix was evaluated by XRD. It was observed that when Mt was added to PAN nanofibers, Mt was completely exfoliated and the peak at $2\theta = 7.12°$ completely disappeared in XRD images. Similarly, on addition of ZnO-Mt to PAN nanofibers, ZnO-Mt was exfoliated and the peak at $2\theta = 7.32°$ disappeared in the PAN/ZnO-Mt nanofibrous nanocomposite (Fig. 5). In our previous study, sericin/PVA/clay nanofibrous nanocomposite-organo-modified clay was completely exfoliated in the sericin/PVA matrix [40]. Kahraman et al. [42] observed that organo-modified clay intercalated in the PAN nanofibrous matrix and increased the interlayer spacing in the clay platelets.

BET surface area

The BET surface area of Mt was found to be 6.910 $m^2 g^{-1}$. On modification of Mt with ZnO nanoparticles, the surface area (11.165 $m^2 g^{-1}$) was increased 1.6-fold. The addition of Mt and ZnO-Mt into PAN fibers resulted in three-fold and

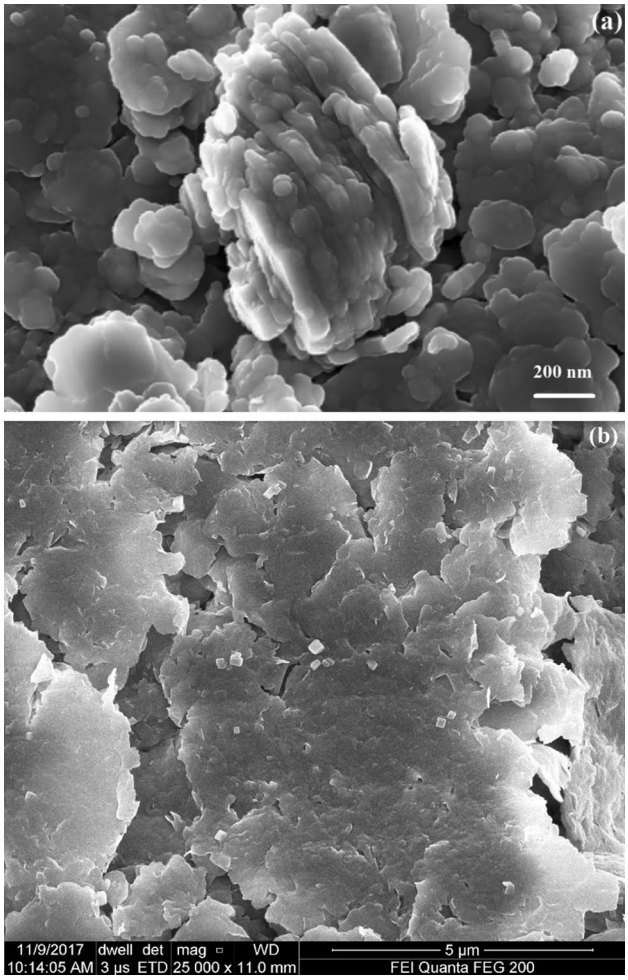


Fig. 2 FESEM images of (a) Mt and (b) ZnO-Mt

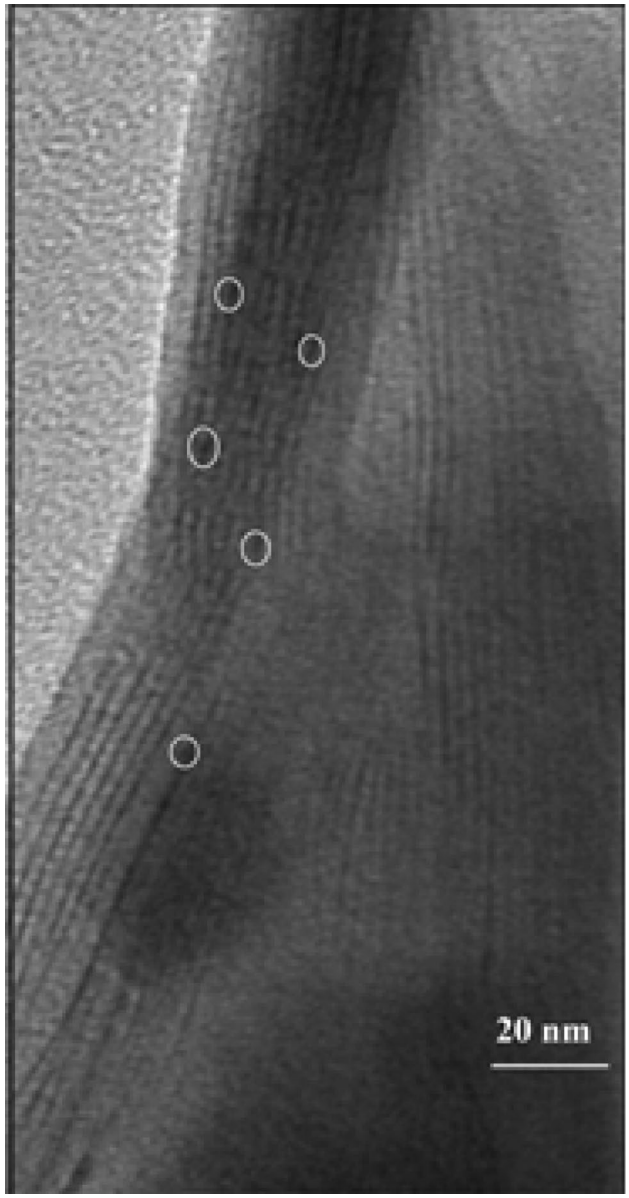


Fig. 3 HRTEM image of ZnO-Mt

four-fold increase, respectively, for BET surface area, and the surface area of the PAN/Mt and PAN/ZnO-Mt nanofibrous nanocomposites was found to be 23.665 $m^2 g^{-1}$ and 46.567 $m^2 g^{-1}$, respectively. Li et al. [43] prepared chitosan nanofibers of enhanced specific BET surface area of 6.692 $m^2 g^{-1}$ as compared to chitosan (0.489 $m^2 g^{-1}$). The higher surface area of chitosan nanofibers resulted in higher adsorption capacity for Cr (VI) ions in comparison to the powder.

Thermal analysis

Mt, ZnO-Mt, PAN/Mt nanofibrous nanocomposites and PAN/ZnO-Mt nanofibrous nanocomposite were analyzed for their thermal stability (Fig. 6). It was found that the first thermal degradation temperature of ZnO-Mt was 235 °C, whereas for Mt it was 203 °C. The residual mass obtained at 900 °C increased from 55% in Mt to 58% for ZnO-Mt. This implies an increase in the thermal stability of Mt with addition of ZnO to Mt. Three-step thermal degradation was observed in the case of nanofibrous nanocomposites. The initial thermal degradation was observed at 76 °C and 74 °C with weight loss of 1.3% and 1.6% for PAN/Mt and the PAN/ZnO-Mt nanofibrous nanocomposites, respectively, due to loss of bound water. The second thermal degradation temperature of the PAN/Mt nanofibrous nanocomposites was found to be 265 °C with weight loss of 33%, and increased to 272 °C with weight loss of 36.4% for the PAN/ZnO-Mt nanofibrous nanocomposites. During the third transition, the weight loss was 8% and 13% for the PAN/Mt and PAN/ZnO-Mt nanofibrous nanocomposites, respectively. The residual mass increased from 31% for PAN/Mt to 34% for the PAN/ZnO-Mt nanofibrous nanocomposites, respectively, at 900 °C, suggesting an increase in the thermal stability of

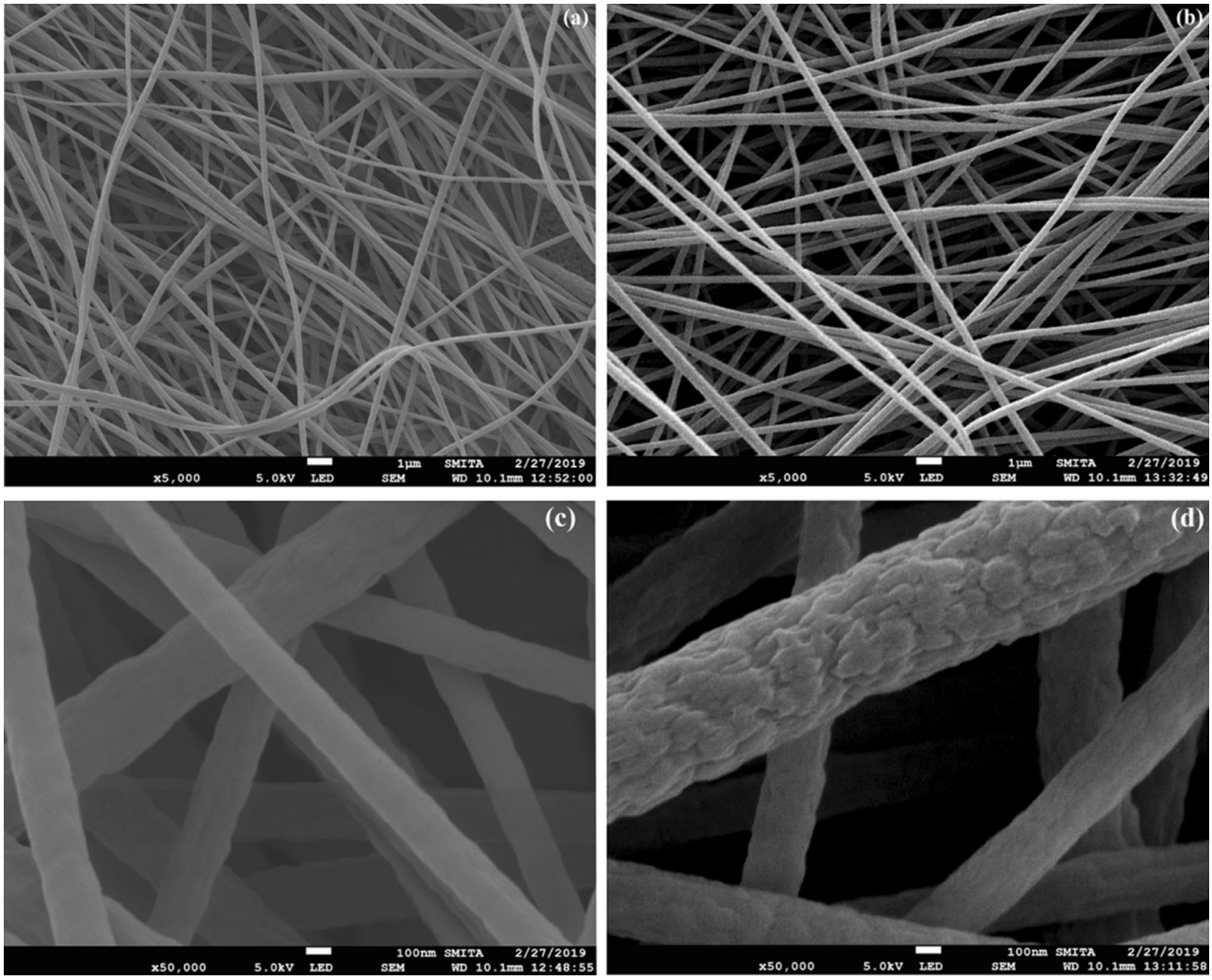


Fig. 4 FESEM images of (a) PAN/Mt, (b) PAN/ZnO-Mt nanofibrous nanocomposites, higher-magnification FESEM images of (c) PAN/Mt and (d) PAN/ZnO-Mt nanofibrous nanocomposites

the PAN nanofibers on addition of ZnO-Mt in comparison to Mt.

Effect of pH

The effect of pH on the adsorption of Cr (VI) ions using Mt, ZnO-Mt, PAN/Mt nanofibrous nanocomposites and PAN/ZnO-Mt nanofibrous nanocomposites is shown in Fig. 7. The maximum adsorption of Cr (VI) ions was found to occur at a pH of 4.5 for Mt and ZnO-Mt, whereas for nanofibrous nanocomposites it was observed at pH 4.75. The amount of Cr (VI) ions adsorbed using Mt, ZnO-Mt, PAN/Mt nanofibrous nanocomposites and PAN/ZnO-Mt nanofibrous nanocomposites was 14.10 mgg^{-1}, 52.61 mgg^{-1}, 216.21 mgg^{-1} and 350.16 mgg^{-1}, respectively. The adsorption of Cr (VI) ions increased initially with an increase in pH, and after attaining an optimum value, adsorption decreased with increasing pH. The acidity of the solution is one of the important factors governing the adsorption of Cr (VI) ions. The point of zero charge of Mt, ZnO-Mt, PAN/Mt nanofibrous nanocomposite and PAN/ZnO-Mt nanofibrous nanocomposite was found to be 5.9, 6.7, 6.4 and 7.2, respectively. The point of zero charge of PAN/ZnO fibers has been reported to be 6.4 [44]. According to the point of zero charge, the surface of the adsorbents will be positively charged below pH 5.9 for Mt, pH 6.7 for ZnO-Mt, pH 6.4 for PAN/Mt and pH 7.2 for the PAN/ZnO-Mt nanofibrous nanocomposites and negatively charged above this pH. The adsorption of Cr (VI) ions will be higher at pH below the point of zero charge as the anionic form of Cr (VI) ion will bind with the positively charged surface of the adsorbent in acidic media. The difference in the adsorption of Cr (VI) ions at different pH is attributed to the ability of the adsorbent to bind with the

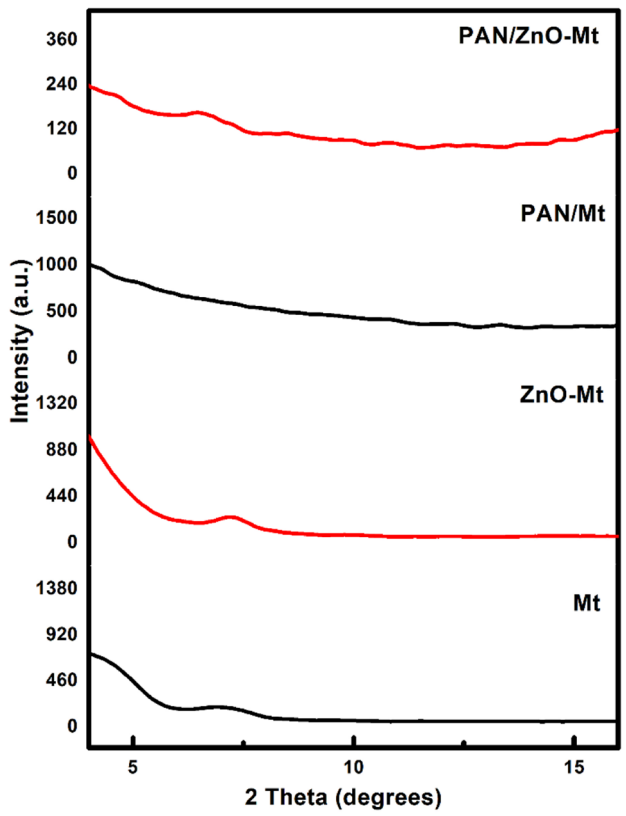


Fig. 5 XRD of (a) Mt, (b) ZnO-Mt, (c) PAN/Mt nanofibrous nanocomposites and (d) PAN/ZnO-Mt nanofibrous nanocomposites

various Cr (VI) ion species (H_2CrO_4, $Cr_2O_7^{2-}$, CrO_4^{2-} and $HCrO^{4-}$) present in the solution at different pH values [44, 45].

Adsorption kinetics

The metal ion adsorption of the PAN/ZnO-Mt nanofibrous nanocomposites, PAN/Mt nanofibrous nanocomposites, ZnO-Mt and Mt was studied for hexavalent Cr (VI) ions, which are known to be the most toxic form and are reported as carcinogenic and mutagenic to living organisms. Mt can adsorb metal ions via two different ways: (i) through cation exchange at the interlayers resulting from interactions between ions and a negative permanent charge, and (ii) through the formation of complexes with Si–O and Al–O groups at the clay particle edge. Both the mechanisms are pH-dependent [46, 47]. In the case of the ZnO nanoparticles, the surface area of the particle is responsible for adsorption of metal ions and forms a metal ion/ZnO hybrid which is dependent on the pH of the solution. The predicted mechanism of Cr (VI) adsorption on ZnO or clay in the acidic pH (pH = 4.75) is associated with electrostatic attraction between the surface hydroxyl groups of metal oxides ($ZnOH^{2+}$) or protonated silanol and aluminol groups at the edge of the clay and CrO_4^{2-} or $Cr_2O_7^{2-}$. Therefore, the present adsorption study was carried out at pH 4.75 for Cr (VI) ions. In the case of nanofibrous nanocomposites, both Mt and ZnO-Mt are exfoliated during electrospinning, providing greater surface area for removal of Cr (VI) ions.

A rapid and stable kinetics of adsorption is necessary for an adsorbent to determine its efficiency in removal of heavy metal ions from aqueous solutions. The adsorption of Cr (VI) ions with respect to time of contact with adsorbent shows that as the time of contact increases, adsorption of metal ions initially increases rapidly up to 60 min, beyond which a slower increase in adsorption of the metal ion is observed (Fig. 8). The addition of ZnO to Mt enhanced the adsorption of Cr (VI) ions by two-fold. However, for PAN/Mt and the PAN/ZnO-Mt nanofibrous nanocomposites, the amount of Cr (VI) adsorbed was found to be 401 mgg^{-1} and 421 mgg^{-1}, respectively, in comparison to 2.06 mgg^{-1} for Mt and 4.11 mgg^{-1} for ZnO-Mt, suggesting an enhanced adsorption for Cr (VI) ions in the case of the nanofibrous nanocomposites. This difference in behavior may be attributed to the higher specific surface area and porosity of the nanofibrous nanocomposites as compared to ZnO-Mt and Mt, providing a greater number of active sites for adsorption and possibly partial access to the inter- and intra-fibrous network, resulting in better adsorption of metal ions, while as time increases, the diffusion rate is reduced. The addition of Mt and ZnO-Mt to the nanofibrous nanocomposites substantially improved the adsorption of Cr (VI) ions over the adsorbents. The BET surface area of the adsorbents was found to increase in the order Mt (6.910 $m^2 g^{-1}$) < ZnO-Mt (11.165 $m^2 g^{-1}$) < PAN/Mt (23.665 $m^2 g^{-1}$) < PAN/ZnO-Mt (46.567 $m^2 g^{-1}$), which was another major factor governing the adsorption. Amongst the various adsorbents studied, PAN/ZnO-Mt nanofibrous nanocomposites with the highest specific surface area achieved the highest adsorption (421 mgg^{-1}) for Cr (VI) ions.

One of the important factors in determining the applicability of the adsorption process is predicting the rate of adsorption. To examine the reaction kinetics, pseudo-first-order and pseudo-second-order models were studied. Various factors, including the morphology and type of adsorbent and the physicochemical properties of adsorbent and adsorbate, affect the reaction kinetics and the process performance.

Pseudo-first-order and pseudo-second-order kinetic models are represented using Eqs. 3 and 4, respectively,

$$\frac{dq_t}{dt} = k_1\left(q_e - q_t\right) \tag{3}$$

$$\frac{dq_t}{dt} = k_2\left(q_e - q_t\right)^2 \tag{4}$$

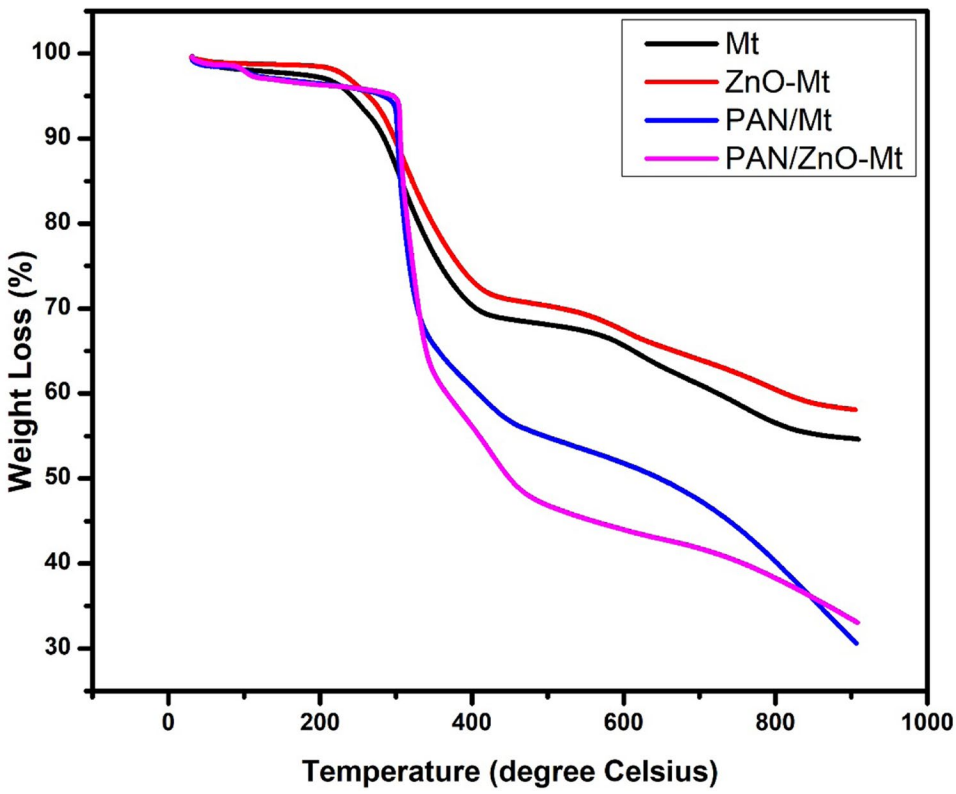


Fig. 6 TGA thermograms of Mt, ZnO-Mt, PAN/Mt nanofibrous nanocomposites and PAN/ZnO-Mt nanofibrous nanocomposites

The linear form of the pseudo-first-order and pseudo-second-order reaction kinetic models can be represented using Eqs. 5 and 6, respectively.

$$\ln\left(q_e - q_t\right) = \ln q_e - k_1 t \tag{5}$$

$$\frac{t}{q_t} = \frac{1}{k_2 q_e^2} + \frac{t}{q_e} \tag{6}$$

where t, q_e, q_t, k_1 and k_2 are time, amount of Cr (VI) ion adsorbed at equilibrium (mgg^{-1}) and time t (min), and rate constant of the first-order (min^{-1}) and second-order ($lmin^{-1}$), respectively.

From the linear plot of $\ln(q_e - q_t)$ versus t (Fig. 9 a and b), the value of rate constant k_1 was found to be 0.0401 min^{-1}, 0.0138 min^{-1}, 0. 0757 min^{-1} and 0.1071 min^{-1} for Cr (VI) for Mt, ZnO-Mt, PAN/Mt and PAN/ZnO-Mt nanofibrous nanocomposites, respectively. The value of second-order rate constant k_2 was calculated from the linear plot of t/q_t versus t (Fig. 9 c and d) and was found out to be 5.67×10^{-4} $lmin^{-1}$, 75.12×10^{-4} $lmin^{-1}$, 38.40×10^{-4} $lmin^{-1}$ and 0.76×10^{-4} $lmin^{-1}$ for Cr (VI) for Mt, ZnO-Mt, PAN/Mt and PAN/ZnO-Mt nanofibrous nanocomposites, respectively.

The value of q_e was also calculated from the plots and was found to be in good agreement with the experimental value. The various parameters calculated from the pseudo-first-order and pseudo-second-order kinetic models are summarized in Table 1.

The value of R^2 ($R^2 > 95\%$), i.e., the correlation coefficient, was used to fit the models to the experimental data. The average absolute relative deviation (AARD) $= \frac{1}{n}\Sigma(\frac{y_i - y_{fit,i}}{y_i})$, where y_i and $y_{fit,i}$ are the left-hand side of Eqs. 5 and 6). An error metric was also used for evaluation of the kinetic model to best fit the experimental data. The higher values of R^2 revealed the best fit between experimental data and the model. Mt best fitted both the pseudo-first-order and pseudo-second-order kinetic model, but in the case of the pseudo-first-order kinetic model, the value of q_e was not in good agreement with experimental data, which was also indicated from the higher value of AARD. The experimental data of the kinetic study suggest that the adsorption of Cr (VI) ions over adsorbent best fits the pseudo-second-order kinetic model (low AARD and high R^2), which suggests that the rate-controlling step is chemically governed or involves chemisorption involving shared or exchanged electrons between adsorbent and adsorbate. Similar results were obtained in amidoximated polyacrylamide/organobentonite composite for adsorption of copper ions [48]. The adsorption of lead ion on clay/poly(methoxyethyl)acrylamide composite [49] and Cr (VI) ions on aromatic polyamide and *s*-triazine core silane coupling agent-modified ZnO nanocomposite [50] also follow the pseudo-second-order kinetic model. PAN/ZnO nanofibrous nanocomposites follow the second-order kinetic model for Cd (II) and Pb (II) ions [12].

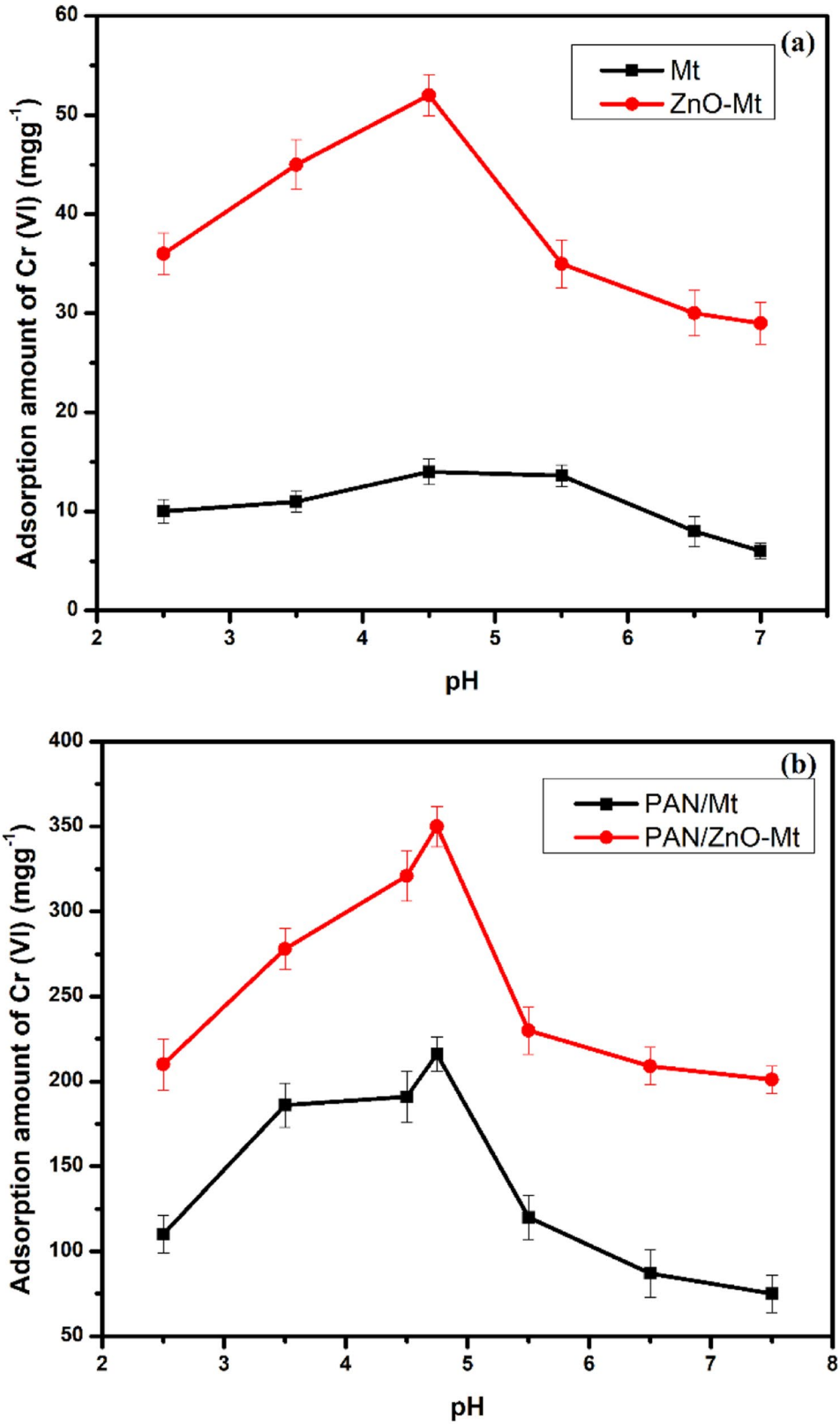


Fig. 7 Effect of pH on adsorption of Cr (VI) for (a) Mt and ZnO-Mt and (b) PAN/Mt and PAN/ZnO-Mt nanofibrous nanocomposite

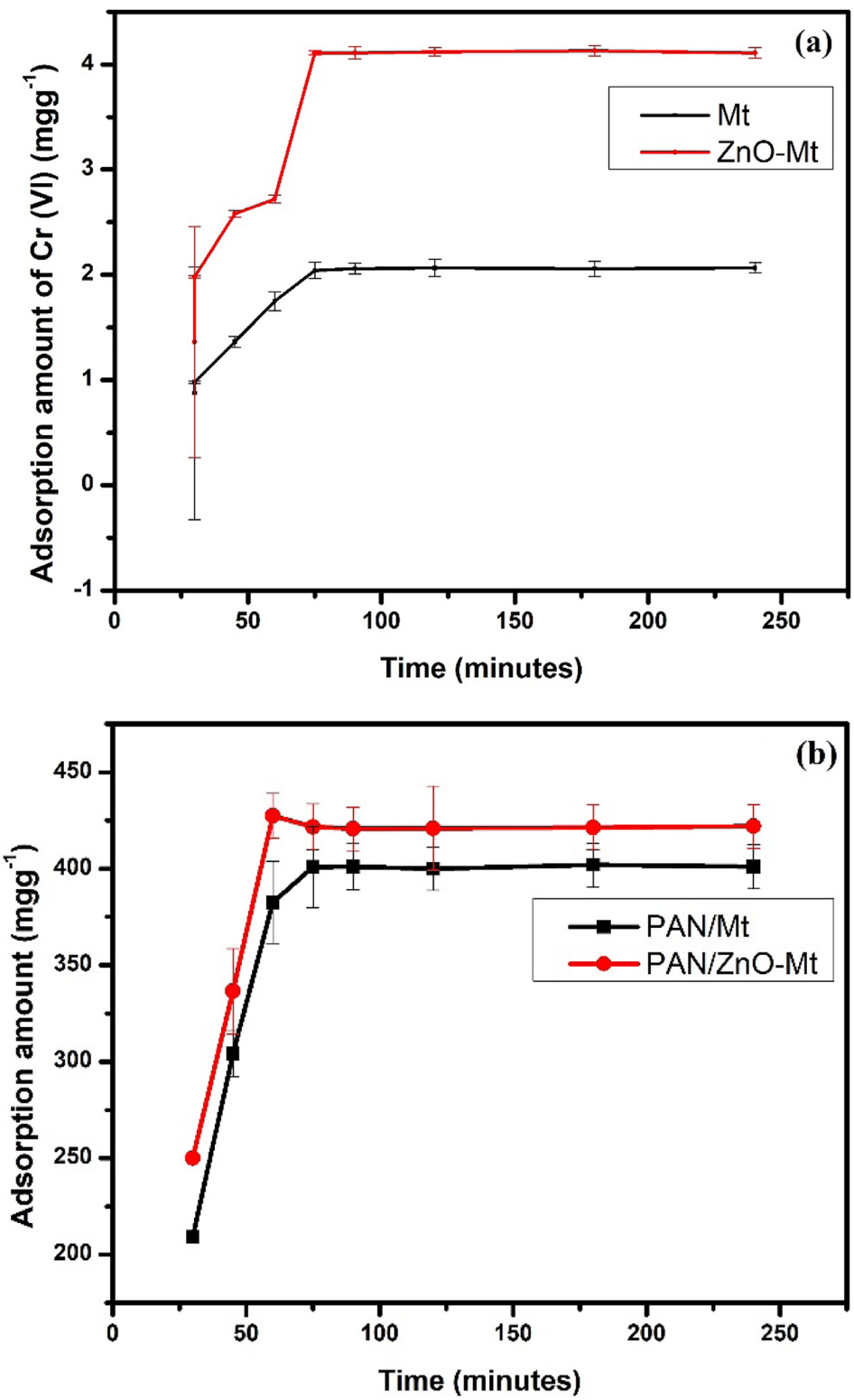


Fig. 8 Effect of contact time on removal of Cr (VI) (experimental conditions: concentration of Cr (VI): 100 ppm, amount of Mt and ZnO-Mt 60 mg in Cr (VI), amount of PAN/Mt and PAN/ZnO-Mt: 3 mg)

The effect of amount of adsorbent on removal of Cr (VI) ions was studied, and it was observed that the increase in the amount of adsorbent increases the adsorption of Cr (VI) ions (Fig. 10). The amount of adsorbent affects the removal efficiency of Cr (VI) ion as it determines the number of available active sites. On increasing the amount of adsorbent in the same volume of solution, a greater number of active sites are available for binding to metal ions. The percentage removal

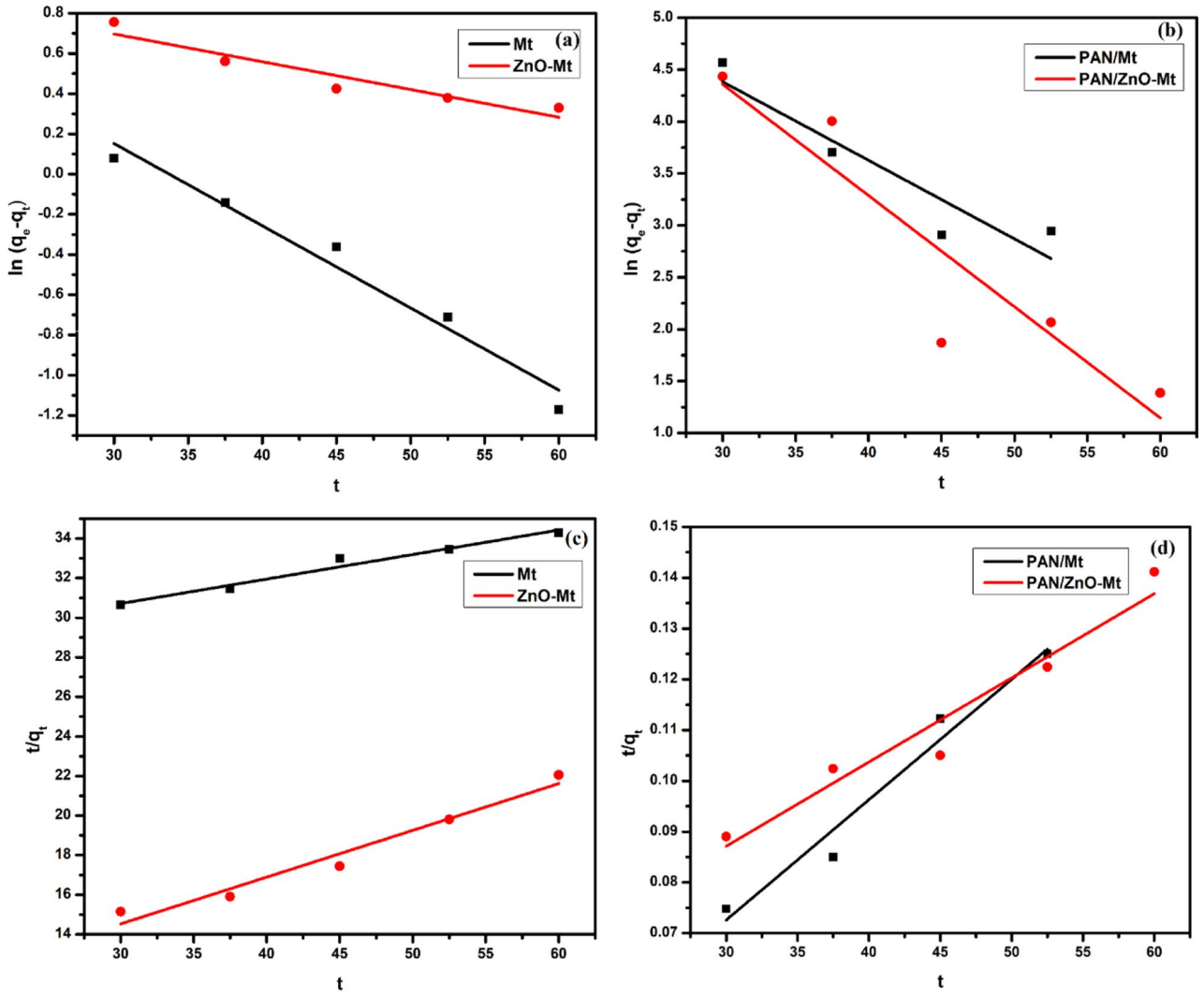


Fig. 9 (a) Pseudo-first-order kinetic plot, pseudo-second-order kinetic plot of (c) Mt and ZnO-Mt and (d) PAN/Mt and PAN/ZnO-Mt nanofibrous nanocomposite

of metal ions increases as the surface area of the adsorbent increases on increasing the amount of adsorbent [51]. Initially, Cr (VI) ions are adsorbed over and between the layers of adsorbent resulting in rapid increase of removal of Cr (VI) ions. After attaining an optimum value, the adsorption of Cr (VI) ions increases at a slower rate due to the formation of

Table 1 Model parameters of pseudo-first-order and pseudo-second-order kinetic models for Cr (VI)

		Mt	ZnO-Mt	PAN/Mt nanofibrous nanocomposites	PAN/ZnO-Mt nanofibrous nanocomposites
Pseudo-first-order	q_e(experimental)	2.06	4.11	401	421
Pseudo-first-order	k_1 X 10^{-2} (min^{-1})	4.01	1.38	7.57	10.71
	q_e (mgg^{-1})	3.97	3.04	777.51	1945.16
	R^2	0.9709	0.9073	0.8766	0.8597
	AARD (%)	31.77	9.4	6.28	16.17
Pseudo-second-order	k_2 X 10^{-4} ($lmin^{-1}$)	5.67	75.12	38.40	0.76
	q_e (mgg^{-1})	8.08	4.23	416.67	488.23
	R^2	0.9718	0.9650	0.9686	0.9507
	AARD (%)	0.51	2.49	3.43	3.35

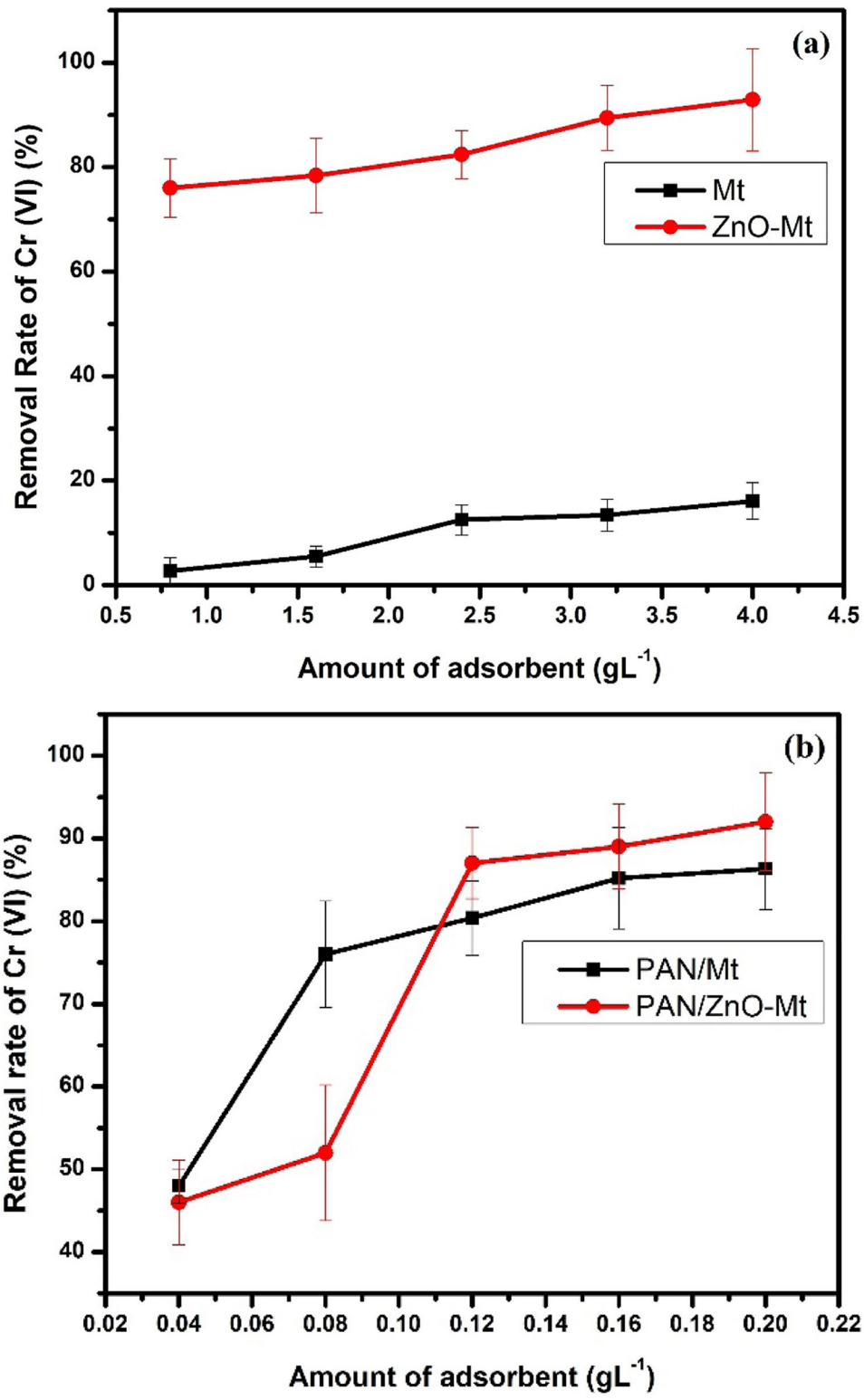


Fig. 10 Effect of amount of adsorbent on removal of (a and b) Cr (VI) (experimental conditions: concentration of Cr (VI): 100 ppm, contact time: 120 min)

clusters over the surface of the adsorbent as an increasing number of metal ions are adsorbed over the surface of the adsorbent, preventing the Cr (VI) ions from being adsorbed between the layers of adsorbent.

It is observed that the incorporation of Mt and ZnO-Mt into the PAN nanofibers reduces the amount of adsorbent required for removal of metal ions. On incorporation of Mt and ZnO-Mt into the nanofibers, the amount of adsorbent required for removal of Cr (VI) ions is decreased

by 20-fold. The optimum dose of adsorbent required for removal of metal ions is lower in the case of ZnO-Mt in comparison to Mt. This is because the smaller size of ZnO-Mt provides a greater number of active sites for binding of metal ions. Cr (VI) is adsorbed through electrostatic interaction with clay as well as ZnO. When Mt and ZnO-Mt are incorporated into the PAN nanofibers, the surface area of the adsorbent for the metal ion increases, and the pore size of the nanofibrous nanocomposites decreases, resulting in a lower amount of adsorbent required for adsorption of metal ions compared to Mt and ZnO-Mt. Haddad and Alharbi [12] found a 123% increase in the adsorption of Cd (II) ions on PAN/ZnO nanofibrous mat as compared to PAN nanofibrous mat, and adsorption equilibrium was reached within 60 min. When kaolinite and kaolinite modified with acid were used as an adsorbent for the adsorption of Cr (VI) ions, it was reported that the adsorption increased with an increase in the amount of adsorbent added [52].

The effect of the initial concentration of Cr (VI) ions on the adsorbent shows that as the concentration of Cr (VI) ions increases, the amount of adsorbed metal ions increases rapidly without attaining equilibrium (Fig. 11). It was observed that the adsorption of metal ions was higher in the case of the PAN/ZnO-Mt nanofibrous nanocomposite and ZnO-Mt as compared with the PAN/Mt nanofibrous nanocomposite and Mt, respectively. This increase in the removal of metal ions using nanofibrous nanocomposites as an adsorbent is observed due to an increased surface-to-volume ratio and surface area of Mt and ZnO-Mt during fabrication of the nanofibrous nanocomposites. ZnO was immobilized over Tunisian clay, and a decrease was found in the degradation ratio of dyes with the initial dye concentration [32]. It has been reported that kaolinite and acid-modified kaolinite decrease the extent of adsorption and increase the amount adsorbed per unit mass as the concentration of metal ions increases [52].

Adsorption Isotherm

Equilibrium adsorption isotherms are promising tools to understand the mechanism of adsorption. The Langmuir and Freundlich isotherms are represented using Eqs. 7 and 8, respectively.

$$q_e = \frac{bq_m C_e}{1 + bC_e} \tag{7}$$

$$q_e = k_f C_e^{\frac{1}{n}} \tag{8}$$

The linear forms of the Langmuir and Freundlich isotherms were analyzed using Eqs. 9 and 10, respectively.

$$\frac{C_e}{q_e} = \frac{1}{bq_m} + \frac{C_e}{q_m} \tag{9}$$

$$\ln q_e = \ln k_f + \frac{1}{n}\ln C_e \tag{10}$$

in which C_e and q_e are the adsorbate concentration at equilibrium in $mg\,L^{-1}$ and equilibrium amount of heavy metal adsorbed in mgg^{-1}, respectively, b and q_m are adsorption energy and saturated monolayer sorption capacity, respectively, related to the Langmuir constant, and k_f and n are the intensity of adsorption and Freundlich constant related to adsorption capacity, respectively. The Langmuir adsorption model depicts the adsorption of a gas–solid phase and is used to study the adsorptive efficiency of adsorbents. It describes the surface coverage by the study of the dynamic equilibrium between adsorption and desorption. The Langmuir constant related to adsorption capacity is related to surface area and the porosity of the adsorbent, which suggests that greater surface area and pore size increase adsorption efficiency. The separation factor R_L was calculated to determine the favorable conditions for adsorption using Eq. 11. The adsorption is said to be favorable when $0 < R_L < 1$ and unfavorable when $R_L > 1$.

$$R_L = \frac{1}{1 + bC_0} \tag{11}$$

The adsorption process occurring over heterogeneous surfaces is described by the Freundlich isotherm. For the Langmuir isotherm of Cr (VI), a plot of C_e/q_e versus C_e was constructed, representing a straight line (Fig. 12a and b). A plot of lnq_e and lnC_e was constructed to study the Freundlich isotherm of Cr (VI). A straight line was obtained in the case of the Freundlich isotherm (Fig. 12 c and d). The values of b, q_m, R_L, n and k_f for Cr (VI) ions are tabulated in Table 2.

By linear curve fitting, it was found that the adsorption of Cr (VI) follows Langmuir and Freundlich isotherms. The error metrics (AARD) were used for fitting of experimental data. It was observed that R^2 was higher in the case of the Langmuir model. The Langmuir model is based on the assumption that none of the interactions are present between adjacent adsorbed species and all the available sites of adsorption are equivalent in energy. The fitting for the Langmuir adsorption isotherm suggested homogeneously distributed adsorption sites on the surface of the adsorbent [12]. The Freundlich isotherm assumes that adsorption occurs on a heterogeneous surface, with the stronger binding sites occupied first, and that the binding strength decreases with an increasing degree of site

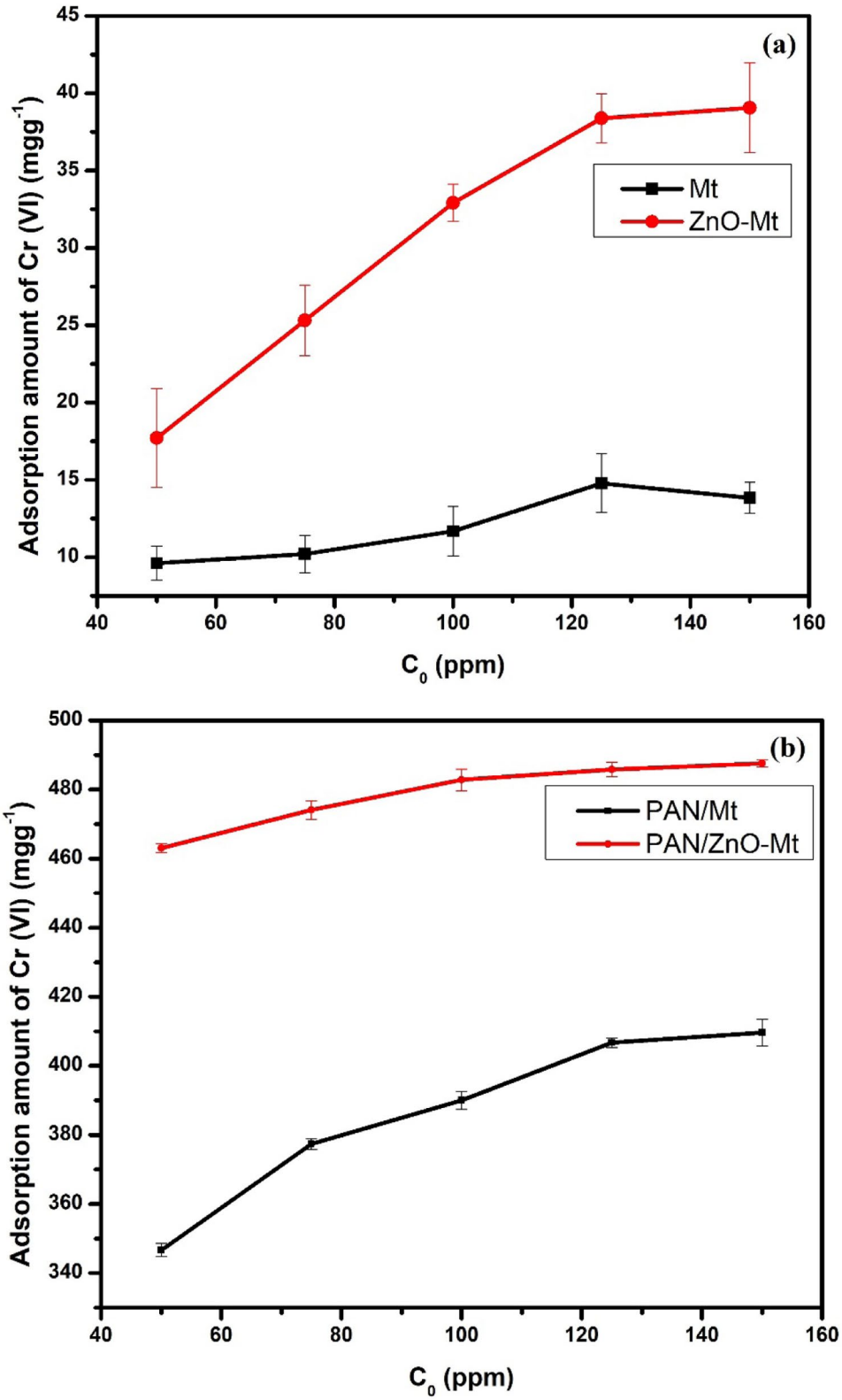


Fig. 11 Effect of concentration of adsorbent on percent removal of Cr (VI) metal ion (experimental conditions: amount of Mt and ZnO-Mt is 60 mg, amount of PAN/Mt and PAN/ZnO-Mt is 3 mg, contact time is 120 min)

occupation [53]. For the Freundlich model, the value of k_f at equilibrium is 3.16 to 456 for different adsorbent and adsorbate systems. The Freundlich constant n is a measure of the deviation from linearity of the adsorption. If the value for n is below unity, it implies that the adsorption process is governed by a chemical mechanism, whereas

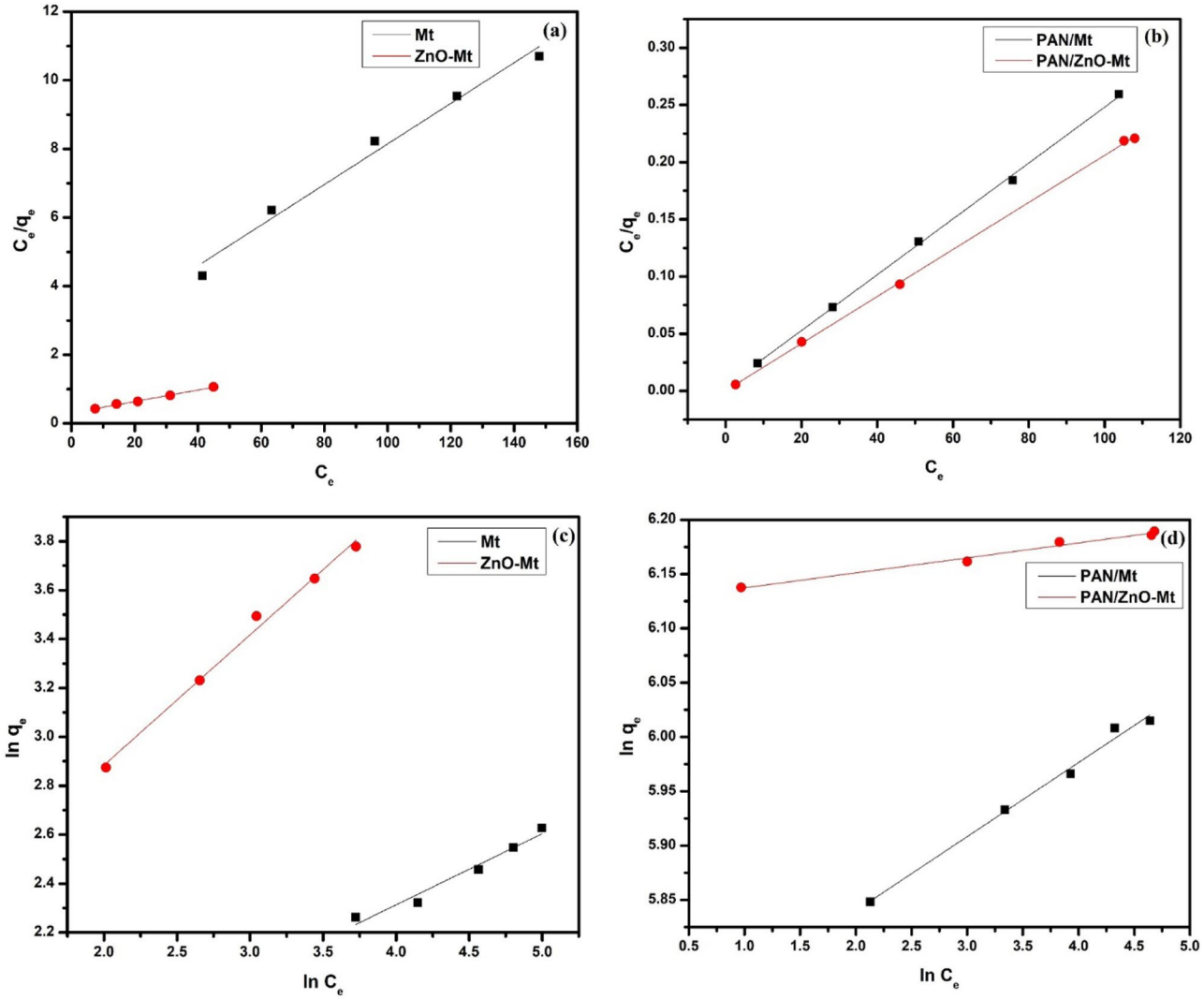


Fig. 12 (a) Langmuir isotherm and (b) Freundlich isotherm for adsorption of Cr (VI) ions

a value for n above unity suggests that the adsorption is a favorable physical process [49]. The value of n at equilibrium lies between 2.87 and 4.54, suggesting a physical mechanism of adsorption. From the linear plots, the R^2 and AARD values were lower in the case of the Freundlich model as compared to Langmuir model. Lower values of

Table 2 Various parameters of Langmuir and Freundlich isotherm for adsorption of Cr (VI) ions

Isotherm		Mt	ZnO-Mt	PAN/Mt nanofibrous nanocomposites	PAN/ZnO-Mt nanofibrous nanocomposites
Langmuir	q_m (mgg^{-1})	16.89	59.88	416.67	476.19
	b	0.03	0.06	0.60	7.00
	R^2	0.9855	0.9962	0.9991	0.9997
	AARD (%)	4.13	0.53	1.06	0.63
	R_L	0.25	0.14	0.02	0.001
Freundlich	k_f	3.16	6.16	300..01	456.46
	n	3.44	2.87	3.46	4.54
	R^2	0.9671	0.9613	0.9919	0.9847
	AARD (%)	1.15	0.51	0.26	0.21

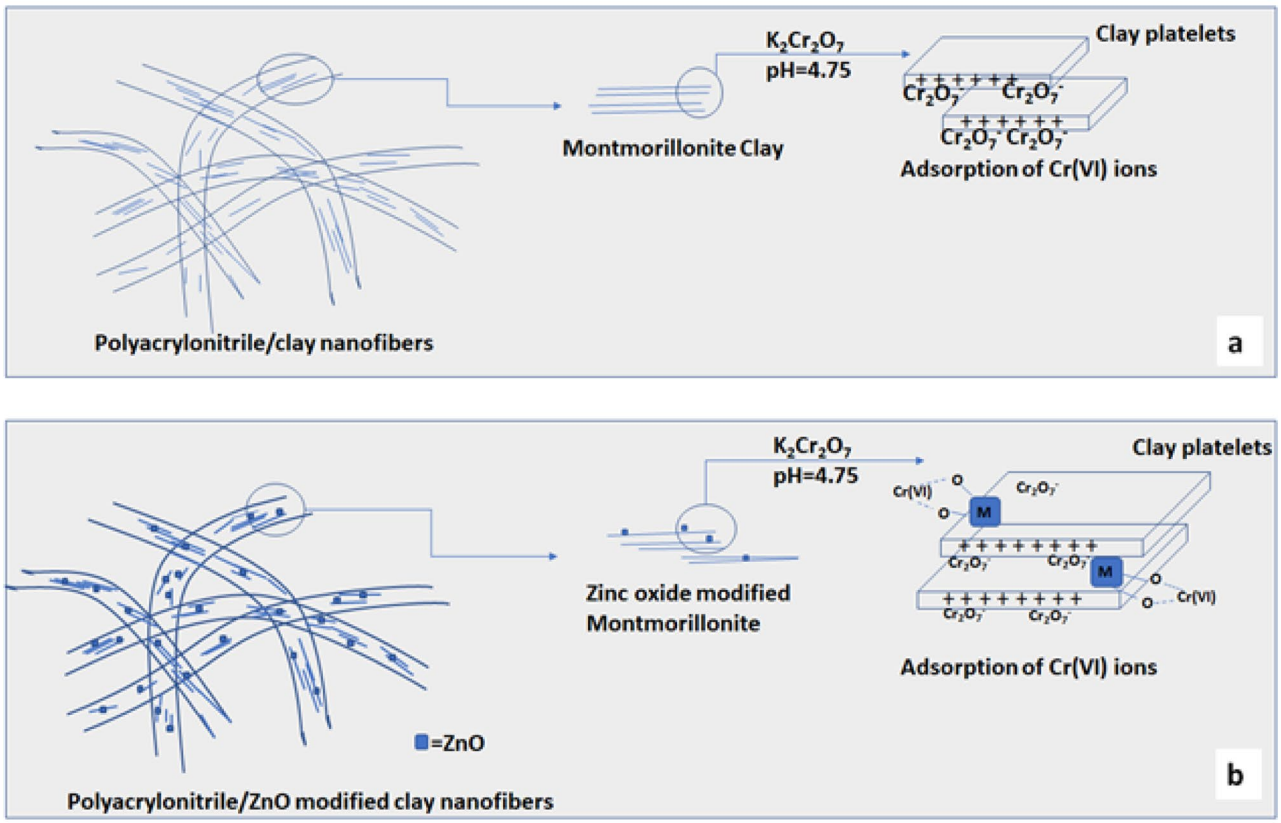


Fig. 13 Mechanism of adsorption of Cr (VI) ions on (a) PAN/ Mt (b) PAN/ZnO-Mt. nanofibrous nanocomposites

AARD suggest that the adsorbents best fit the Freundlich adsorption isotherm, indicating heterogeneous adsorption of Cr (VI) ions on the surface of Mt, ZnO-Mt, PAN/Mt and PAN/ZnO-Mt nanofibrous nanocomposites. The possible interaction of Cr (VI) with the nanocomposite nanofibers is depicted in Fig. 13. The PAN/ZnO-Mt nanofibrous nanocomposites were found to be most efficient in adsorption of Cr (VI) ions, which is attributed to the larger number of surface active sites for metal ion adsorption because of its porous structure and surface polar nature of positively (Zn) and negatively charged (O) ions [54].

The heterogeneous adsorption in the presence of ZnO-Mt can also be explained by XRD (Fig. 1). The XRD analysis of ZnO-Mt suggests that the spacing between the layers of Mt decreases on addition of ZnO nanoparticles, and ZnO is inserted between the layers and is also uniformly distributed over the layers of Mt (Fig. 1). This insertion of ZnO nanoparticles over the surface and between the layers of Mt results in a change in the size of ZnO-Mt in comparison to ZnO and Mt. The reduction in size of clay after modification with ZnO enhances the surface area, which further helps to improve the adsorption efficiency

Table 3 Cr (VI) ion adsorption capacity of various adsorbents

Adsorbent material	Amount of adsorbent ($g L^{-1}$)	Initial Cr (VI) concentration ($mg L^{-1}$)	q_m ($mg\ g^{-1}$)
HCl-activated Akadama Clay [55]	5.00	112.5	7.47
Acid-activated montmorillonite clay [56]	10.00	250	4.51
Cerium oxide nanoparticles [57]	0.32	80	121.95
Chitosan/nylon 6 nanofibers [58]	1.00	30–200	23.90
Tannin/nylon 6 nanofibers [58]	1.00	30–200	62.70
Mt (this study)	2.40	100	16.89
ZnO-Mt (this study)	2.40	100	59.88
PAN/Mt (this study)	0.12	100	416.67
PAN/ZnO-Mt (this study)	0.12	100	476.19

of ZnO-Mt. When Mt is added to PAN nanofibers, Mt is completely exfoliated, and on addition of ZnO-Mt to PAN nanofibers, ZnO-Mt is exfoliated. The porous structure of the nanofibers also supports the heterogeneous adsorption of metal ions on nanofibrous nanocomposites. Cr (VI) ion adsorption on kaolinite clay [52] and aromatic polyamide/s-triazine heterocyclic ring-modified ZnO nanocomposite [50] follows both the Langmuir and Freundlich isotherms. The value of *n* lies between 2 and 9. Amidoximated polyacrylonitrile/organobentonite composite [48] shows the best correlation with the Langmuir isotherm for removal of Cu (II) ions. In the case of the PAN/ZnO nanofibrous nanocomposite, the Langmuir isotherm is best fit for adsorption of Cd (II) and Pb (II) ions [12]. The addition of ZnO-Mt to PAN nanofibers enhances the adsorption efficiency for Cr (VI). Various adsorbents and their adsorption capacities for Cr (VI) are tabulated in Table 3.

Conclusions

The effect of concentration, amount of adsorbent and time on the removal of metal ions shows that when Mt and ZnO-Mt are incorporated in PAN nanofibers, better adsorption of metal ions is observed owing to the increase in the specific surface area available for adsorption and the uniform distribution of pore size compared to Mt and ZnO-Mt. The intercalation of ZnO-Mt over PAN nanofibers resulted in a decrease in the amount of adsorbent and contact time for removal of metal ions compared to ZnO-Mt. PAN/Mt nanofibrous nanocomposites are found to be more effective adsorbents for heavy metal ions compared to Mt. The adsorption studies suggest that PAN/Mt and PAN/ZnO-Mt nanofibrous nanocomposites can be used as adsorbent for removal of Cr (VI) ions from aqueous solutions. The adsorption of Cr (VI) ions over Mt, ZnO-Mt, PAN/Mt nanofibrous nanocomposites and PAN/ZnO-Mt nanofibrous nanocomposites follows pseudo-second-order kinetics, suggesting chemisorption between adsorbate and adsorbent. The adsorption isotherm analysis suggested that homogeneous and heterogeneous adsorption of Cr (VI) ions occurs over the adsorbent and thus best fit the Langmuir and Freundlich adsorption isotherms. The addition of Mt and ZnO-Mt to electrospun PAN nanofibers enhances the adsorption of Cr (VI) ions over the adsorbent site.

Acknowledgements The authors thankfully acknowledge the financial support received from the Science and Engineering Research Board (SERB), Department of Science and Technology (DST), Govt. of India (EMR/2017/002833).

Compliance with ethical standard

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Wang G, Hua Y, Su X et al (2016) Cr(VI) adsorption by montmorillonite nanocomposites. Appl Clay Sci 124–125:111–118. https://doi.org/10.1016/j.clay.2016.02.008
2. Upadhyay RK, Soin N, Roy SS (2013) Role of graphene/metal oxide composites as photocatalysts, adsorbents and disinfectants in water treatment: a review. RSC Advances 4:3823–3851. https://doi.org/10.1039/C3RA45013A
3. Ngwenya N, Ncube EJ, Parsons J (2013) Recent Advances in Drinking Water Disinfection: Successes and Challenges. Rev Environ Contam Toxicol 222:111–170. https://doi.org/10.1007/978-1-4614-4717-7_4
4. Ahmad M, Zhang B, Wang J et al (2019) New method for hydrogel synthesis from diphenylcarbazide chitosan for selective copper removal. Int J Biol Macromol 136:189–198. https://doi.org/10.1016/j.ijbiomac.2019.06.084
5. Yi J, Li Y, Yang L, Zhang LM (2019) Kinetics and thermodynamics of adsorption of Cu2+ and methylene blue to casein hydrogels. J Polym Res 26. https://doi.org/10.1007/s10965-019-1870-x
6. Li L, Wang Z, Ma P et al (2015) Preparation of polyvinyl alcohol/chitosan hydrogel compounded with graphene oxide to enhance the adsorption properties for Cu(II) in aqueous solution. J Polym Res 22:1–10. https://doi.org/10.1007/s10965-015-0794-3
7. Li T, Liu X, Li L, et al (2019) Polydopamine-functionalized graphene oxide compounded with polyvinyl alcohol/chitosan hydrogels on the recyclable adsorption of cu(II), Pb(II) and cd(II) from aqueous solution. J Polym Res 26. https://doi.org/10.1007/s10965-019-1971-6
8. Manzoor K, Ahmad M, Ahmad S, Ikram S (2019a) Synthesis, Characterization, Kinetics, and Thermodynamics of EDTA-Modified Chitosan-Carboxymethyl Cellulose as Cu(II) Ion Adsorbent. ACS Omega 4:17425–17437. https://doi.org/10.1021/acsomega.9b02214
9. Manzoor K, Ahmad M, Ahmad S, Ikram S (2019b) Removal of Pb(ii) and Cd(ii) from wastewater using arginine cross-linked chitosan-carboxymethyl cellulose beads as green adsorbent. RSC Advances 9:7890–7902. https://doi.org/10.1039/C9RA00356H
10. Cheng R, Ou S, Xiang B et al (2009) Adsorption behavior of hexavalent chromium on synthesized ethylenediamine modified starch. J Polym Res 16:703–708. https://doi.org/10.1007/s10965-009-9276-9
11. Li P, Dai X, Yan Q, et al (2020) A novel modification method for polystyrene microspheres with dithizone and the adsorption properties for Pb2+. J Polym Res 27. https://doi.org/10.1007/s10965-020-02180-8
12. Haddad MY, Alharbi HF (2019) Enhancement of heavy metal ion adsorption using electrospun polyacrylonitrile nanofibers loaded with ZnO nanoparticles. J Appl Polym Sci 136. https://doi.org/10.1002/app.47209
13. Cai J, Lei M, Zhang Q et al (2017) Composites: Part A Electrospun composite nanofiber mats of Cellulose@Organically modified montmorillonite for heavy metal ion removal: Design, characterization, evaluation of absorption performance. Compos A 92:10–16. https://doi.org/10.1016/j.compositesa.2016.10.034

14. sen Gupta S, Bhattacharyya KG, (2012) Adsorption of heavy metals on kaolinite and montmorillonite: a review. Phys Chem Chem Phys 14:6698. https://doi.org/10.1039/c2cp40093f
15. Uddin MK (2017) A review on the adsorption of heavy metals by clay minerals, with special focus on the past decade. Chem Eng J 308:438–462. https://doi.org/10.1016/j.cej.2016.09.029
16. Wang Y, Zhao X, Duan L et al (2015) Structure, luminescence and photocatalytic activity of Mg-doped ZnO nanoparticles prepared by auto combustion method. Mater Sci Semicond Process 29:372–379. https://doi.org/10.1016/j.mssp.2014.07.034
17. Deng Z, Chen M, Gu A, Wu L (2008) A facile method to fabricate ZnO hollow spheres and their photocatalytic property. J Phys Chem B 112:16–22. https://doi.org/10.1021/jp077662w
18. Rezapour M, Talebian N (2011) Comparison of structural, optical properties and photocatalytic activity of ZnO with different morphologies: Effect of synthesis methods and reaction media. Mater Chem Phys 129:249–255. https://doi.org/10.1016/j.matchemphys.2011.04.012
19. Akkari M, Aranda P, ben Rhaiem H, et al (2016) ZnO/clay nanoarchitectures: Synthesis, characterization and evaluation as photocatalysts. Appl Clay Sci 131:131–139. https://doi.org/10.1016/j.clay.2015.12.013
20. Yang SJ, Park CR (2008) Facile preparation of monodisperse ZnO quantum dots with high quality photoluminescence characteristics. Nanotechnol 19. https://doi.org/10.1088/0957-4484/19/03/035609
21. Talebian N, Amininezhad SM, Doudi M (2013) Controllable synthesis of ZnO nanoparticles and their morphology-dependent antibacterial and optical properties. J Photochem Photobiol, B 120:66–73. https://doi.org/10.1016/j.jphotobiol.2013.01.004
22. Yamamoto O (2001) Influence of particle size on the antibacterial activity of zinc oxide. Int J Inorg Mater 3:643–646. https://doi.org/10.1016/S1466-6049(01)00197-0
23. Raghupathi KR, Koodali RT, Manna AC (2011) Size-dependent bacterial growth inhibition and mechanism of antibacterial activity of zinc oxide nanoparticles. Langmuir 27:4020–4028. https://doi.org/10.1021/la104825u
24. Zarrindokht Emami-Karvani (2012) Antibacterial activity of ZnO nanoparticle on Gram-positive and Gram-negative bacteria. Afr J Microbiol Res 5. https://doi.org/10.5897/AJMR10.159
25. Sirelkhatim A, Mahmud S, Seeni A et al (2015) Review on zinc oxide nanoparticles: Antibacterial activity and toxicity mechanism. Nano-Micro Lett 7:219–242
26. Upadhyaya L, Singh J, Agarwal V, et al (2014) In situ grafted nanostructured ZnO/carboxymethyl cellulose nanocomposites for efficient delivery of curcumin to cancer. J Polym Res 21. https://doi.org/10.1007/s10965-014-0550-0
27. Upadhyaya L, Singh J, Agarwal V et al (2015) Efficient water soluble nanostructured ZnO grafted O-carboxymethyl chitosan/curcumin-nanocomposite for cancer therapy. Process Biochem 50:678–688. https://doi.org/10.1016/j.procbio.2014.12.029
28. Dimapilis EAS, Hsu C-S, Mendoza RMO, Lu M-C (2018) Zinc oxide nanoparticles for water disinfection. Sustain Environ Res 28:47–56. https://doi.org/10.1016/J.SERJ.2017.10.001
29. Ong CB, Ng LY, Mohammad AW (2018) A review of ZnO nanoparticles as solar photocatalysts: Synthesis, mechanisms and applications. Renew Sustain Energy Rev 81:536–551. https://doi.org/10.1016/J.RSER.2017.08.020
30. Kumar JP, Ramacharyulu PVRK, Prasad GK, Singh B (2015) Montmorillonites supported with metal oxide nanoparticles for decontamination of sulfur mustard. Appl Clay Sci 116–117:263–272. https://doi.org/10.1016/j.clay.2015.04.007
31. Sani HA, Ahmad MB, Hussein MZ et al (2017) Nanocomposite of ZnO with montmorillonite for removal of lead and copper ions from aqueous solutions. Process Saf Environ Prot 109:97–105. https://doi.org/10.1016/j.psep.2017.03.024
32. Bel Hadjltaief H, ben Ameur S, da Costa P, et al (2018) Photocatalytic decolorization of cationic and anionic dyes over ZnO nanoparticle immobilized on natural Tunisian clay. Appl Clay Sci 152:148–157. https://doi.org/10.1016/j.clay.2017.11.008
33. Akkari M, Aranda P, Belver C et al (2018) ZnO/sepiolite heterostructured materials for solar photocatalytic degradation of pharmaceuticals in wastewater. Appl Clay Sci 156:104–109. https://doi.org/10.1016/j.clay.2018.01.021
34. Unuabonah EI, Taubert A (2014) Clay–polymer nanocomposites (CPNs): Adsorbents of the future for water treatment. Appl Clay Sci 99:83–92. https://doi.org/10.1016/J.CLAY.2014.06.016
35. Patel S, Hota G (2016) Iron oxide nanoparticle-immobilized PAN nanofibers: synthesis and adsorption studies. RSC Advances 6:15402–15414. https://doi.org/10.1039/C5RA20345G
36. Hallaji H, Keshtkar AR, Moosavian MA (2015) A novel electrospun PVA/ZnO nanofiber adsorbent for U(VI), Cu(II) and Ni(II) removal from aqueous solution. J Taiwan Inst Chem Eng 46:109–118. https://doi.org/10.1016/j.jtice.2014.09.007
37. Kampalanonwat P, Supaphol P (2010) Preparation and adsorption behavior of aminated electrospun polyacrylonitrile nanofiber mats for heavy metal ion removal. ACS Appl Mater Interfaces 2:3619–3627. https://doi.org/10.1021/am1008024
38. Makaremi M, Lim CX, Pasbakhsh P et al (2016) Electrospun functionalized polyacrylonitrile-chitosan Bi-layer membranes for water filtration applications. RSC Advances 6:53882–53893. https://doi.org/10.1039/c6ra05942b
39. Bode-Aluko CA, Pereao O, Ndayambaje G, Petrik L (2017) Adsorption of Toxic Metals on Modified Polyacrylonitrile Nanofibres: A Review. Water Air Soil Pollut 228:35. https://doi.org/10.1007/s11270-016-3222-3
40. Purwar R, Goutham KS, Srivastava CM (2016) Electrospun Sericin/PVA/Clay Nanofibrous Mats for Antimicrobial Air Filtration Mask. Fibers Polym 17:1206–1216. https://doi.org/10.1007/s12221-016-6345-7
41. Kumar A, Prasad B, Mishra IM (2008) Adsorptive removal of acrylonitrile by commercial grade activated carbon: Kinetics, equilibrium and thermodynamics. J Hazard Mater 152:589–600. https://doi.org/10.1016/j.jhazmat.2007.07.048
42. Kahraman HT, Yar A, Avcı A, Pehlivan E (2018) Preparation of nanoclay incorporated PAN fibers by electrospinning technique and its application for oil and organic solvent absorption. Sep Sci Technol (Philadelphia) 53:303–311. https://doi.org/10.1080/01496395.2017.1384018
43. Li L, Li Y, Cao L, Yang C (2015) Enhanced chromium (VI) adsorption using nanosized chitosan fibers tailored by electrospinning. Carbohyd Polym 125:206–213. https://doi.org/10.1016/j.carbpol.2015.02.037
44. Parlayıcı Ş, Yar A, Pehlivan E, Avcı A (2019) ZnO-TiO2 doped polyacrylonitrile nano fiber-Mat for elimination of Cr (VI) from polluted water. J Anal Sci Technol 10:0–11. https://doi.org/10.1186/s40543-019-0183-3
45. Pradhan D, Sukla LB, Mishra BB, Devi N (2019) Biosorption for removal of hexavalent chromium using microalgae Scenedesmus sp. J Clean Prod 209:617–629. https://doi.org/10.1016/j.jclepro.2018.10.288jornal
46. Abollino O, Aceto M, Malandrino M et al (2003) Adsorption of heavy metals on Na-montmorillonite. Effect of pH and organic substances. Water Res 37:1619–1627. https://doi.org/10.1016/S0043-1354(02)00524-9
47. El-Bayaa AA, Badawy NA, AlKhalik EA (2009) Effect of ionic strength on the adsorption of copper and chromium ions by vermiculite pure clay mineral. J Hazard Mater 170:1204–1209. https://doi.org/10.1016/J.JHAZMAT.2009.05.100
48. Anirudhan TS, Ramachandran M (2008) Synthesis and Characterization of Amidoximated Polyacrylonitrile/Organobentonite Composite for Cu(II), Zn(II), and Cd(II) Adsorption from

Aqueous Solutions and Industry Wastewaters. Ind Eng Chem Res 47:6175–6184. https://doi.org/10.1021/ie070735d
49. Şölener M, Tunali S, Özcan AS et al (2008) Adsorption characteristics of lead(II) ions onto the clay/poly(methoxyethyl)acrylamide (PMEA) composite from aqueous solutions. Desalination 223:308–322. https://doi.org/10.1016/J.DESAL.2007.01.221
50. Dinari M, Haghighi A (2018) Ultrasound-assisted synthesis of nanocomposites based on aromatic polyamide and modified ZnO nanoparticle for removal of toxic Cr(VI) from water. Ultrason Sonochem 41:75–84. https://doi.org/10.1016/j.ultsonch.2017.09.023
51. Tangaromsuk J, Pokethitiyook P, Kruatrachue M, Upatham ES (2002) Cadmium biosorption by Sphingomonas paucimobilis biomass. Biores Technol 85:103–105. https://doi.org/10.1016/S0960-8524(02)00066-4
52. Bhattacharyya KG, Gupta S, sen, (2006) Adsorption of chromium(VI) from water by clays. Ind Eng Chem Res 45:7232–7240. https://doi.org/10.1021/ie060586j
53. Khan TA, Nazir M, Ali I, Kumar A (2017) Removal of Chromium(VI) from aqueous solution using guar gum–nano zinc oxide biocomposite adsorbent. Arabian J Chem 10:S2388–S2398. https://doi.org/10.1016/J.ARABJC.2013.08.019
54. Singh S, Barick KC, Bahadur D (2013) Functional oxide nanomaterials and nanocomposites for the removal of heavy metals and dyes. Nanomater Nanotechnol 3:1–19. https://doi.org/10.5772/57237
55. Zhao Y, Qi W, Chen G et al (2015) Behavior of Cr(VI) removal from wastewater by adsorption onto HCl activated Akadama clay. J Taiwan Inst Chem Eng 50:190–197. https://doi.org/10.1016/j.jtice.2014.12.016
56. Akar ST, Yetimoglu Y, Gedikbey T (2009) Removal of chromium (VI) ions from aqueous solutions by using Turkish montmorillonite clay: effect of activation and modification. Desalination 244:97–108. https://doi.org/10.1016/j.desal.2008.04.040
57. Recillas S, Colón J, Casals E et al (2010) Chromium VI adsorption on cerium oxide nanoparticles and morphology changes during the process. J Hazard Mater 184:425–431. https://doi.org/10.1016/j.jhazmat.2010.08.052
58. Kummer G, Schonhart C, Fernandes MG et al (2018) Development of Nanofibers Composed of Chitosan/Nylon 6 and Tannin/Nylon 6 for Effective Adsorption of Cr(VI). J Polym Environ 26:4073–4084. https://doi.org/10.1007/s10924-018-1281-9

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

Circuits, Systems, and Signal Processing
https://doi.org/10.1007/s00034-020-01603-7

Quadrature Sinusoidal Oscillators Using CDBAs: New Realizations

Ram Bhagat[1] · **D. R. Bhaskar**[1] · **Pragati Kumar**[1]

Received: 15 March 2020 / Revised: 10 November 2020 / Accepted: 13 November 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract

Four new circuits of fully uncoupled quadrature sinusoidal oscillators, using two current differencing buffered amplifiers (CDBAs), four/five resistors and two capacitors have been presented. In contrast to all previously published CDBA-based, fully decoupled, quadrature sinusoidal oscillators in which one of the input terminals is left unutilized, the presented circuits are realized by utilizing the intrinsic current differencing property of CDBAs, thus utilizing all the four terminals of the CDBA. All these quadrature oscillator (QO) circuits have an inherent feature of amplitude control of output voltages without using external control circuitry. Out of the four proposed QO circuits, two circuits have the additional feature of generating low-frequency oscillations, thus making them capable of generating wide range of frequency waveforms. The proposed oscillator circuits also possess additional functionalities not available in other CDBA-based quadrature sinusoidal oscillator circuits presented previously. The workability of the proposed circuits has been confirmed through experimental results where CDBAs are implemented using commercially available current feedback operational amplifiers (AD844).

Keywords Current differencing buffered amplifiers · Quadrature sinusoidal oscillators · Voltage-mode · Signal processing · Analog filters

✉ Ram Bhagat
rambhagat@dtu.ac.in

D. R. Bhaskar
drbhaskar@dtu.ac.in

Pragati Kumar
pragatikumar@dtu.ac.in

[1] Department of Electrical Engineering, Delhi Technological University, New Delhi, Delhi 110042, India

Published online: 03 January 2021

1 Introduction

Sinusoidal oscillators play an important part in various measurements, instrumentation, control, communication and other related electronic systems [29]. Quadrature oscillators, which provide two sinusoids with 90° phase difference, find applications in communication systems wherein they are used in quadrature mixers, single-sideband generators, and in measurement systems, used in vector generators or selective voltmeters. In past, a number of active building blocks (ABBs) have been used to realize QOs such as current feedback operational amplifiers (CFOAs) [30, 36], current differencing buffered amplifiers (CDBAs) [5, 11, 15, 16, 20, 23, 37–39], voltage differencing transconductance amplifiers (VDTAs) [25], voltage differencing buffered amplifiers (VDBAs) [21], current differencing transconductance amplifiers (CDTAs) [6, 10, 12, 17, 35, 40], voltage differencing inverting voltage buffered amplifiers (VDIBAs) [1, 3, 28], voltage differencing current conveyors (VDCCs) [31, 33], modified current differencing transconductance amplifiers (MCDTAs) [18], current follower transconductance amplifiers (CFTAs) [13, 14, 22, 24], zero copy current follower transconductance amplifiers (ZC-CFTAs) [19, 41], current gain current follower transconductance amplifiers (CG-CFTAs) [27], differential input buffered transconductance amplifier (DBTAs) [4], current conveyor transconductance amplifiers (CCTAs) [7, 34], multiple-output current controlled current differencing transconductance amplifiers (MO-CCCDTAs) [26], current controlled current conveyors transconductance amplifiers (CCCCTAs) [8], current follower current conveyor (CFCC) [32] and commercially available IC LT 1228 s [9].

Out of the various new building blocks mentioned above, CDBA has received comparatively more prominent attention in analog signal processing and signal generation circuits due to its important property of providing at its input terminals, virtual ground, which eliminates the effect of parasitic at the input terminals. Apart from this, its low output impedance enables the CDBA-based circuit configurations to be cascaded easily. Since the presented paper deals with new realization of quadrature sinusoidal oscillators with CDBAs, it is worthwhile to review various QO circuits realized by CDBAs.

A catalogue of 12 QO circuits was presented in [23]. These circuits were realized using a methodology involving cascading of non-inverting low pass filter with inverting integrator in a loop or by cascading inverting low pass filter with non-inverting integrator circuit in a loop realized with two CDBAs, four resistors and two grounded capacitors. It may be noted that in all the 12 QO circuits, one of the two input terminals (p/n) was left unused. An open input terminal of CDBA makes the circuit vulnerable to parasitic noise and parasitic terminal impedances. A fully uncoupled QO circuit was reported in [5] using two CDBAs, four resistors and two capacitors. In this circuit also, n-terminal of one of the CDBAs was left open. The QO circuit presented in [11] was realized by cascading two all-pass filters in a loop employing two CDBAs, four resistors and four capacitors. However, the circuits presented therein do not provide independent control of CO and FO. In [15], four fully uncoupled QO circuits using two CDBAs, four resistors

and two grounded capacitors were reported. However, in all four QO circuits one of the terminals (p/n) of the CDBA was grounded. In [16], three low-frequency sinusoidal quadratures oscillator were reported using two CDBAS, four resistors and two grounded/virtually grounded capacitors. However, out of these three proposed QO circuits, one QO circuit uses all the terminals of the CDBAs and in the remaining two QO circuits, one of the n-terminal of the CDBA was grounded. Another QO circuit with independent control of CO and FO was reported in [20], realized with two CDBAs, three resistors and two capacitors. In [37], a fully uncoupled QO circuit derived from a multiple-input–single-output type universal biquad filter was reported. The circuit employed two CDBAs, four resistors and two grounded capacitors. In this circuit also, p-terminal of one of the CDBAs has been grounded. QO circuit reported in [38] has been developed by cascading an all-pass filter with an integrator in a loop. The circuit has independent control of CO and FO and was realized with two CDBAs, three resistors and three capacitors. However, in this circuit also, n-terminal of one of the CDBAs was grounded. Lastly, another QO circuit with independent control of CO and FO reported in [39] uses two CDBAs, three resistors and two grounded capacitors. Like all other CDBA-based oscillator circuits, one of the terminals of the CDBAs was left open.

From the above description, it emerges that none of the CDBA-based QO circuits (except first QO circuit in [16]) reported earlier have utilized the intrinsic (current differencing) property of the CDBA as either n-terminal or p-terminal of one of the CDBAs has been left open. Also, from open literature it has been found that the references cited in [9] and [16] have described the amplitude adjustment of the output voltages. In [16], the author has described the amplitude adjustment of the output voltages by using oscillator internal circuit parameters. However, in [9], besides the oscillator circuit parameters, additional four resistors are required to achieve the amplitude control of the quadrature output voltages.

Therefore, the main aim of this paper is to introduce, four new, fully uncoupled amplitude adjustable (without using external circuitry) QO circuits in which, both the input terminals of the CDBAs are fully utilized. As these QO circuits have been realized using the method proposed in [23], and thus, add four more circuits to the twelve QO circuits reported therein. All the four QO circuits proposed in this paper use two CDBAs, four/five resistors and two capacitors and utilize the intrinsic (current differencing) property of the CDBAs. Out of these four QO circuits, two circuits have the feature of generating low-frequency oscillation (LFO). In addition, these circuits also realize low pass filter (LPF) and band pass filter (BPF) functions simultaneously at the output terminals of each CDBA (if we break the feedback and apply signal to p-terminal of the first CDBA through a resistor). Lastly, the technique used in this paper can also be applied to convert all twelve QO circuits reported in [23] utilizing the intrinsic (current differencing) property of CDBA.

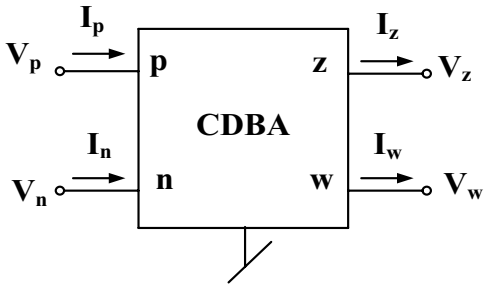


Fig. 1 Symbolic representation of CDBA

2 Proposed Circuits

The symbolic representation of the CDBA is shown in Fig. 1, where p and n are current input terminals, w is the voltage output terminal, and z is the current output terminal.

The characterizing terminal equations of an ideal CDBA are given by:

$$I_z = \left(I_p - I_n\right),\ V_p = 0,\ V_n = 0 \text{ and } V_w = V_z \tag{1}$$

The proposed QO circuits with four/five resistors and two capacitors are shown in Fig. 2.

Using straightforward circuit analysis, the characteristic equation (CE), condition of oscillation (CO) and frequency of oscillation (FO) for the proposed fully

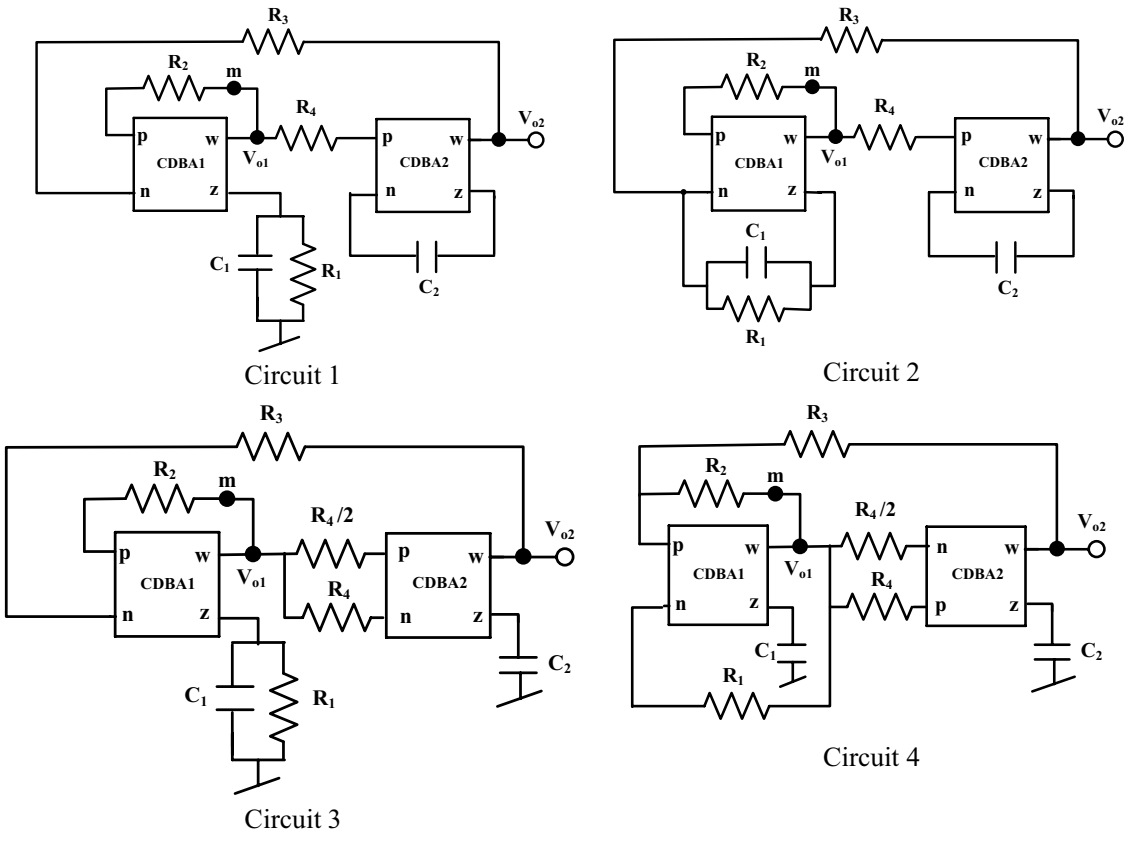


Fig. 2 Proposed fully uncoupled quadrature oscillator circuits

Table 1 CE, CO and FO for proposed quadrature oscillator circuits

Circuit	CE	CO	FO
Circuit 1	$s^2+\left(\frac{1}{R_1C_1}-\frac{1}{R_2C_1}\right)s+\frac{1}{2R_3R_4C_1C_2}$	$R_2 \le R_1$	$\frac{1}{\sqrt{2R_3R_4C_1C_2}}$
Circuit 2	$s^2+\left(\frac{1}{R_1C_1}-\frac{1}{2R_2C_1}\right)s+\frac{1}{4R_3R_4C_1C_2}$	$2R_2 \le R_1$	$\frac{1}{2\sqrt{R_3R_4C_1C_2}}$
Circuit 3	$s^2+\left(\frac{1}{R_1C_1}-\frac{1}{R_2C_1}\right)s+\frac{1}{R_3R_4C_1C_2}$	$R_2 \le R_1$	$\frac{1}{\sqrt{R_3R_4C_1C_2}}$
Circuit 4	$s^2+\left(\frac{1}{R_1C_1}-\frac{1}{R_2C_1}\right)s+\frac{1}{R_3R_4C_1C_2}$	$R_2 \le R_1$	$\frac{1}{\sqrt{R_3R_4C_1C_2}}$

Table 2 Quadrature relationship between voltages V_{01} and V_{02}

Circuit 1	Circuit 2	Circuit 3	Circuit 4
$\frac{V_{02}(j\omega)}{V_{01}(j\omega)}=\left\lvert\frac{1}{\omega R_4C_2}\right\rvert e^{-j90^\circ}$	$\frac{V_{02}(j\omega)}{V_{01}(j\omega)}=\left\lvert\frac{1}{\omega R_4C_2}\right\rvert e^{-j90^\circ}$	$\frac{V_{02}(j\omega)}{V_{01}(j\omega)}=\left\lvert\frac{1}{\omega R_4C_2}\right\rvert e^{-j90^\circ}$	$\frac{V_{02}(j\omega)}{V_{01}(j\omega)}=\left\lvert\frac{1}{\omega R_4C_2}\right\rvert e^{j90^\circ}$
$=\left\lvert\sqrt{\frac{2R_3C_1}{R_4C_2}}\right\rvert e^{-j90^\circ}$	$=2\left\lvert\sqrt{\frac{R_3C_1}{R_4C_2}}\right\rvert e^{-j90^\circ}$	$=\left\lvert\sqrt{\frac{R_3C_1}{R_4C_2}}\right\rvert e^{-j90^\circ}$	$=\left\lvert\sqrt{\frac{R_3C_1}{R_4C_2}}\right\rvert e^{j90^\circ}$

Table 3 Transfer functions of filters realized from proposed circuits

Circuit	Transfer function	
	LPF	BPF
Circuit 1	$\frac{V_{02}(s)}{V_{\text{in}}(s)}=\frac{\left(\frac{1}{2R_2R_4C_1C_2}\right)}{s^2+\left(\frac{1}{R_1C_1}\right)s+\frac{1}{2R_4R_3C_1C_2}}$	$\frac{V_{01}(s)}{V_{\text{in}}(s)}=\frac{\left(\frac{1}{R_2C_1}\right)s}{s^2+\left(\frac{1}{R_1C_1}\right)s+\frac{1}{2R_4R_3C_1C_2}}$
Circuit 2	$\frac{V_{02}(s)}{V_{\text{in}}(s)}=\frac{\left(\frac{1}{4R_2R_4C_1C_2}\right)}{s^2+\left(\frac{1}{R_1C_1}\right)s+\frac{1}{4R_4R_3C_1C_2}}$	$\frac{V_{01}(s)}{V_{\text{in}}(s)}=\frac{\left(\frac{1}{2R_2C_1}\right)s}{s^2+\left(\frac{1}{R_1C_1}\right)s+\frac{1}{4R_4R_3C_1C_2}}$
Circuit 3	$\frac{V_{02}(s)}{V_{\text{in}}(s)}=\frac{\left(\frac{1}{R_2R_4C_1C_2}\right)}{s^2+\left(\frac{1}{R_1C_1}\right)s+\frac{1}{R_4R_3C_1C_2}}$	$\frac{V_{01}(s)}{V_{\text{in}}(s)}=\frac{\left(\frac{1}{R_2C_1}\right)s}{s^2+\left(\frac{1}{R_1C_1}\right)s+\frac{1}{R_4R_3C_1C_2}}$
Circuit 4	$\frac{V_{02}(s)}{V_{\text{in}}(s)}=\frac{\left(\frac{1}{R_2R_4C_1C_2}\right)}{s^2+\left(\frac{1}{R_1C_1}\right)s+\frac{1}{R_4R_3C_1C_2}}$	$\frac{V_{01}(s)}{V_{\text{in}}(s)}=\frac{\left(\frac{1}{R_2C_1}\right)s}{s^2+\left(\frac{1}{R_1C_1}\right)s+\frac{1}{R_4R_3C_1C_2}}$

uncoupled QO circuits have been determined and are given in Table 1. From Table 1, it can be observed that both CO (with R_1 and R_2) and FO (with R_3 and R_4) enjoy fully uncoupled independent tunability.

The steady state relationship between output voltages V_{01} and V_{02} for the circuits of Fig. 2 is given in Table 2.

From Table 2, it may be observed that for $C_1=C_2$, the amplitude ratio of output voltages depends upon R_3 and R_4. If R_3 increases, then the magnitude of V_{01} decreases and V_{02} increases and at the same time if R_4 increases, then the magnitude of V_{01} increases and magnitude of V_{02} decreases. Therefore, from here, it

Table 4 Cut-off frequency (ω_o), bandwidth (BW), quality factor (Q) and gain (H_0) for filters of Table 3

Circuit	Filter	ω_o	BW/Q	H_0
Circuit 1	LPF	$\frac{1}{\sqrt{2R_4R_3C_1C_2}}$	$Q=\frac{R_1C_1}{\sqrt{2R_4R_3C_1C_2}}$	$\frac{R_3}{R_2}$
	BPF	$\frac{1}{\sqrt{2R_4R_3C_1C_2}}$	$\mathrm{BW}=\frac{1}{R_1C_1}$	$\frac{R_1}{R_2}$
Circuit 2	LPF	$\frac{1}{2\sqrt{R_4R_3C_1C_2}}$	$Q=\frac{R_1C_1}{2\sqrt{R_4R_3C_1C_2}}$	$\frac{R_3}{R_2}$
	BPF	$\frac{1}{2\sqrt{R_4R_3C_1C_2}}$	$\mathrm{BW}=\frac{1}{R_1C_1}$	$\frac{R_1}{R_2}$
Circuit 3	LPF	$\frac{1}{\sqrt{R_3R_4C_1C_2}}$	$Q=\frac{R_1C_1}{\sqrt{R_4R_3C_1C_2}}$	$\frac{R_3}{R_2}$
	BPF	$\frac{1}{\sqrt{R_3R_4C_1C_2}}$	$\mathrm{BW}=\frac{1}{R_1C_1}$	$\frac{R_1}{R_2}$
Circuit 4	LPF	$\frac{1}{\sqrt{R_3R_4C_1C_2}}$	$Q=\frac{R_1C_1}{\sqrt{R_4R_3C_1C_2}}$	$\frac{R_3}{R_2}$
	BPF	$\frac{1}{\sqrt{R_3R_4C_1C_2}}$	$\mathrm{BW}=\frac{1}{R_1C_1}$	$\frac{R_1}{R_2}$

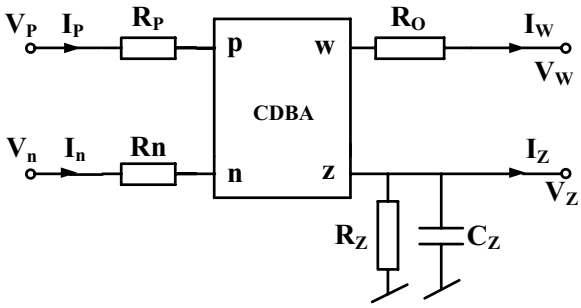


Fig. 3 Non-ideal CDBA with parasitic elements

can be concluded that the amplitude of output voltages of the oscillator circuits can be controlled by varying R_3 and R_4. Equal voltage amplitudes can be achieved in all the four proposed QO circuits if frequency is varied by changing R_3 and R_4 simultaneously, holding the condition $R_4=2R_3$ for Circuit 1, $R_4=4R_3$ for Circuit 2 and $R_3=R_4$ for both circuit 3 and circuit 4.

Furthermore, if we break oscillator feedback loop at point 'm' and apply input voltage to p-terminal of the first CDBA through resistor R_2, then the open-loop transfer functions of the proposed circuits thus obtained are given in Table 3. From Table 3, it is seen that all the proposed circuits realize low pass filter (LPF) and band pass filter (BPF) simultaneously. Table 4 contains the characterizing parameters for the realized filters.

From Table 4, it is observed that in the case of the LPFs, the different parameters, namely pole frequency, pole quality factor and DC gain, may be tuned independently by varying R_3 or R_4, R_1 and R_2 respectively, whereas in the case of the BPFs, for a given bandwidth $(1/R_1C_1)$, the gain and the pole frequency can be tuned independently by varying R_3 or R_4, respectively.

3 Non-ideal Analysis

The non-ideal representation of the CDBA is shown in Fig. 3. A practical CDBA can be described [23] by the following relationships that consider the non-idealities of the device:

Table 5 Non-ideal FO of fully uncoupled QOs of Fig. 2

Circuit	Non-ideal CO	Non-ideal FO $(\overline{\omega})$
Circuit 1	$\frac{2C_2}{R_1}+\frac{C_1}{R_z}+\frac{2C_z}{R_1}+2\frac{C_2}{R_z}\le\frac{2C_2}{R_2}+\frac{C_Z}{R_2}$	$\omega\frac{\left(1+\sqrt{\frac{R_3R_4}{R_z^2}}\right)}{\sqrt{\left(1+\frac{C_z}{C_1}\right)}}$
Circuit 2	$\frac{4C_2}{R_1}+\frac{2C_1}{R_z}+\frac{2C_z}{R_1}+\frac{2C_2}{R_z}\le\frac{2C_2}{R_2}+\frac{C_z}{R_2}$	$\omega\frac{\left(1+\sqrt{\frac{R_3R_4}{R_1R_z}}\right)}{\sqrt{\left(1+\frac{C_z}{2C_1}\right)}}$
Circuit 3	$\frac{1}{R_1}\left(1+C_z\right)+\frac{(1+2C_Z)}{R_z}\le\frac{1}{R_2}$	$\omega\frac{\left(1+\sqrt{\frac{R_3R_4}{R_z^2}}\right)}{\sqrt{\left(1+\frac{C_z}{C_2}\right)\left(1+\frac{C_z}{C_1}\right)}}$
Circuit 4	$\frac{C_z}{R_z}+\frac{(1+C_z)}{R_1}\le\frac{1}{R_2}+\frac{1}{R_z}$	$\omega\frac{\left(1+\sqrt{\frac{R_3R_4}{R_z^2}}\right)}{\sqrt{\left(1+\frac{C_z}{C_2}\right)\left(1+\frac{C_z}{C_1}\right)}}$

$$V_p=0,\ V_n=0,\ I_z=\left(\beta_P I_P-\beta_n I_n\right),\ V_w=\alpha V_z \tag{2}$$

where $\beta_P=\left(1-\varepsilon_P\right);\left(\left|\varepsilon_p\right|<<1\right)$ is the current tracking error from p-terminal to z-terminal,$\beta_n=\left(1-\varepsilon_n\right)$; is the current tracking error from n-terminal to z-terminal, and $\alpha=\left(1-\varepsilon_v\right);\left(\left|\varepsilon_v\right|<<1\right)$ is the voltage tracking error from z-terminal to w-terminal of the CDBA. As the parasitic resistances R_p and R_n associated with the input terminals of the CDBA and R_o associated with its output terminal are negligibly small, we have neglected these impedances in our analysis but considered the parasitic resistance R_z and capacitance C_z associated with the z terminal of the CDBA. In addition to the above, we have ignored the current and voltage tracking errors, in order to focus more clearly on the effects of the parasitic resistance and capacitance associated with the z-terminal of the CDBA. Now, if we reanalyse the circuits of Fig. 2, assuming identical CDBAs, the non-ideal expressions of CO and FO of proposed QO circuits presented in Table 5 are obtained.

To examine the effect of various non-idealities on the performance of the QOs presented in Fig. 2, we have calculated the FO in all the four circuits for a design frequency of 500 kHz by taking $C_1=C_2=100$ pF, $R_1=R_2=1$ kΩ, $R_4=3.18$ kΩ. The

Table 6 Non-ideal phase relationship between V_{01} and V_{02} in circuits of Fig. 2

Circuit 1	Circuit 2	Circuit 3	Circuit 4
$-\tan^{-1}\left(2\omega C_2R_z\right)$	$-\tan^{-1}\left(2\omega C_2R_z\right)$	$-\tan^{-1}\left(\omega C_2\left(1+\frac{C_{z2}}{C_2}\right)R_z\right)$	$\tan^{-1}\left(\omega C_2\left(1+\frac{C_{z2}}{C_2}\right)R_z\right)$

value of R_3 was chosen as 1.59 kΩ, 0.795 kΩ, 3.18 kΩ and 3.18 kΩ for the circuits of oscillator 1, 2, 3 and 4, respectively. For the chosen implementation of the CDBA, using AD844 in this paper, the value of $C_z = 4.5$ pF, $R_z = 3$ MΩ were taken [23]. The values of FO as obtained from Table 5 have been found to be 495 kHz, 492 kHz, 485 kHz and 484 kHz, respectively. These values are in close agreement with the experimental results. It may also be noted that the feature of fully decoupled tuning of the FO and CO is preserved even under the presence of the dominant parasitic influences because the CO depends only on R_1 and R_2 while the FO depends on R_3 and R_4 even under non-ideal condition. In view of this, no adjustment of R_1 or R_2 was needed to readjust the CO when FO was varied by R_3 and R_4.

Non-ideal expressions for quadrature relationship between voltages V_{01} and V_{02} along with phase angle between them are given in Table 6.

From Table 6, it may be observed that the phase difference between the voltages V_{01} and V_{02} would be close to 90° for all the four circuits because of very large value of R_z. The calculated phase difference between two V_{01} and V_{02} has been found to be to 89.9°, 89.9°, 89.93° and 89.93°, respectively, for Circuit1, Circuit 2, Circuit 3 and Circuit 4 of Fig. 2 at a design frequency of 500 kHz.

A similar non-ideal analysis has also been performed for all the filter circuits derived from the oscillators presented in Fig. 2. However, to conserve space, we present the results for the filter derived from the Circuit 3 only. The non-ideal transfer functions of the LPF and BPF are given by:

$$\left.\frac{V_{02}(s)}{V_{\text{in}}(s)}\right|_{\text{LPF}} = \frac{\frac{1}{R_2R_4}}{As^2 + Bs + C}$$

$$\text{where } A = C_zC_z + C_1C_2 + C_1C_z + C_zC_2, B = \left(\frac{C_2}{R_1} + \frac{C_z}{R_z} + \frac{C_z}{R_z} + \frac{C_2}{R_z} + \frac{C_1}{R_z} + \frac{C_z}{R_1}\right) \text{ and}$$

$$C = \left(\frac{1}{R_1R_z} + \frac{1}{R_zR_z} + \frac{2}{R_zR_4} + \frac{1}{R_3R_4}\right) \tag{3}$$

$$\left.\frac{V_{01}(s)}{V_{\text{in}}(s)}\right|_{\text{BPF}} = \frac{C_zC_2s^2 + \left(\frac{C_2}{R_2} + \frac{C_2}{R_z}\right)s}{As^2 + Bs + C}$$

$$\text{where } A = C_zC_z + C_1C_2 + C_1C_z + C_zC_2, \ B = \left(\frac{C_2}{R_1} + \frac{C_z}{R_z} + \frac{C_z}{R_{z1}} + \frac{C_2}{R_z} + \frac{C_1}{R_z} + \frac{C_z}{R_1}\right) \text{ and}$$

$$C = \left(\frac{1}{R_1R_z} + \frac{1}{R_zR_z} + \frac{2}{R_zR_4} + \frac{1}{R_3R_4}\right) \tag{4}$$

The frequency responses of the LPF and BPF as obtained from Eqs. (3)–(4) have been plotted in MATLAB for a cut-off frequency of 159 kHz. The frequency responses for ideal and non-ideal transfer functions of the filters derived for Circuit 3 are plotted and which are shown in Fig. 4.

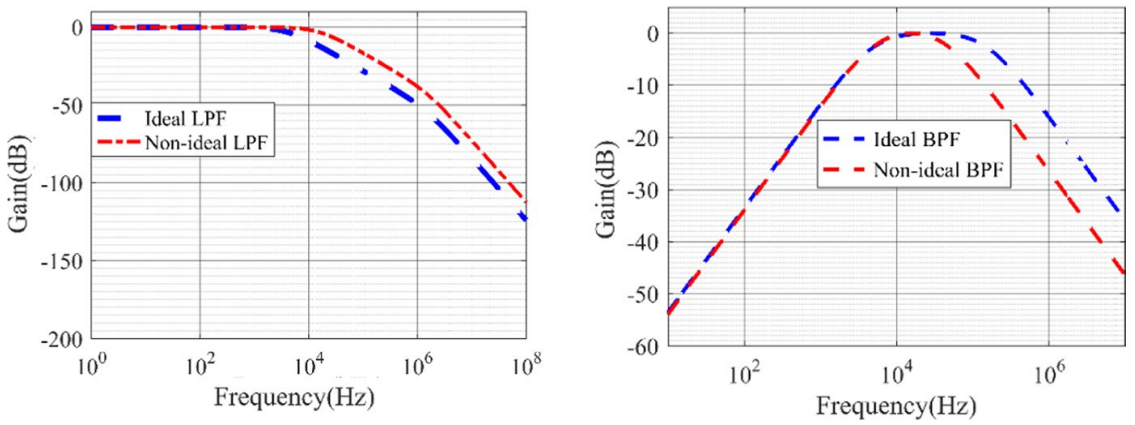


Fig. 4 Non-ideal and ideal frequency responses of LPF and BPF derived from Circuit 3

From frequency responses, it is observed that there is a little deviation in the non-ideal frequency response from the ideal frequency response, indicating that the effect of the dominant non-idealities of the CDBAs on the performance of the circuits is minimal.

Also, we have evaluated the sensitivity of FO with respect to various active and passive parameters using the classical definition of sensitivity and found them all to be no more than half. Similarly, for the realized filters, the various sensitivity coefficients have also been found to be no more than half. Thus, the proposed circuits have low active and passive sensitivities.

4 Frequency Stability

Frequency stability may be considered as one of the important figures of merit, based on which various oscillators can be compared. Using the definition of the frequency stability factor (S^F) as given in [2] $S^F = \frac{\mathrm{d}\varphi(\delta)}{\mathrm{d}\delta}$ where $\delta = \frac{\omega}{\omega_0}$ is the normalized frequency and $\varphi(\delta)$ is the phase of the open-loop transfer function of the oscillator circuit, the expressions for the frequency stability factors for the proposed new quadrature oscillators have been tabulated in Table 7.

Table 7 Frequency stability factor for proposed quadrature oscillator circuits

Circuit	Stability factor $\left(S^F\right)$ $R_1=R_2=R$, $R_3=R_4=R/n$ and $C_1=C_2=C$
Circuit 1	$-n\sqrt{2}$
Circuit 2	$-n$
Circuit 3	$-2n$
Circuit 4	$-2n$

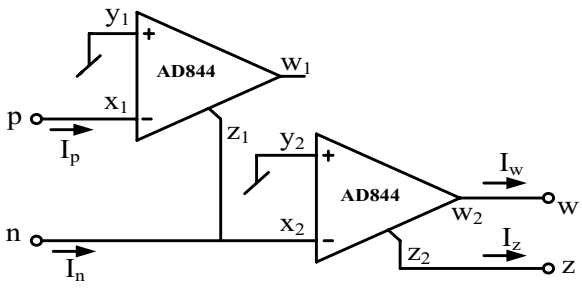
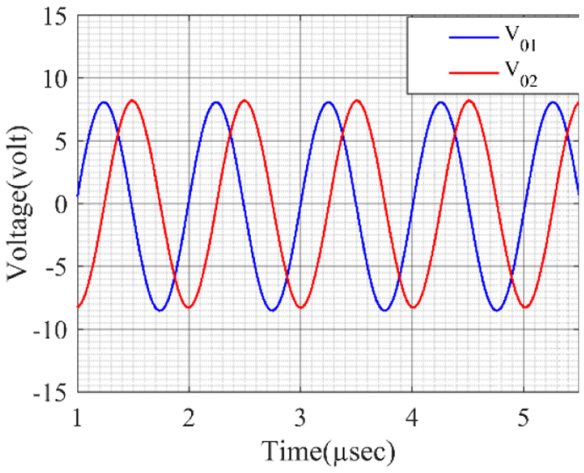
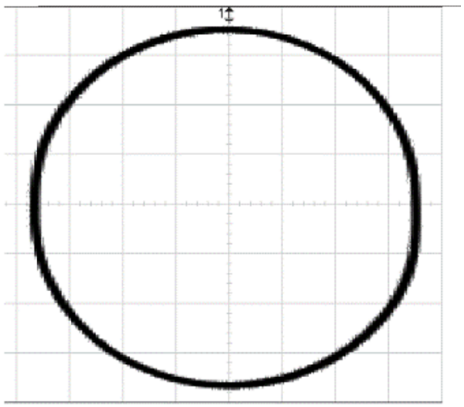


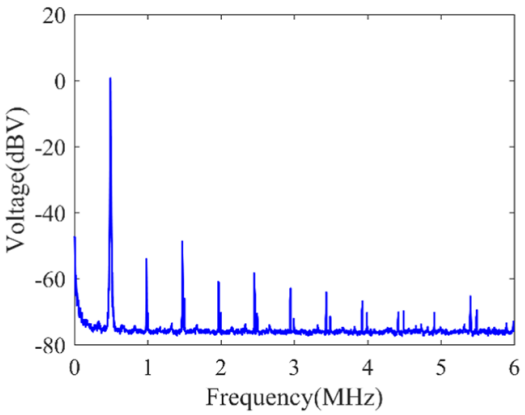
Fig. 5 CDBA realization using AD844



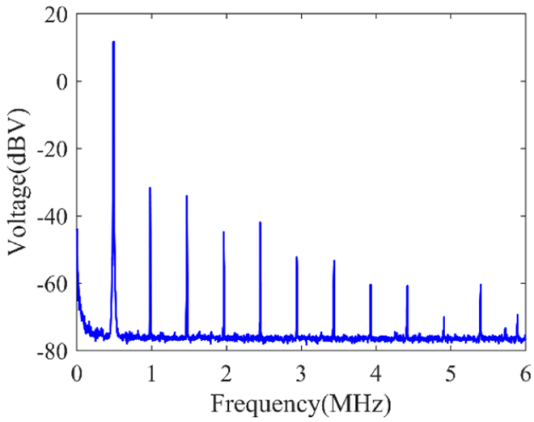
(i)Output waveform at 500kHz



(ii) Lissajous Pattern



(iii) Frequency spectrum of V_{01}



(iv) Frequency spectrum of V_{02}

Fig. 6 Experimental results of proposed QO for Circuit 1

From Table 7, it has been observed that the frequency stability factors for Circuit 3 and Circuit 4 turn out to be $2n$. Thus, excellent frequency stability is obtainable by selecting larger values of n [2].

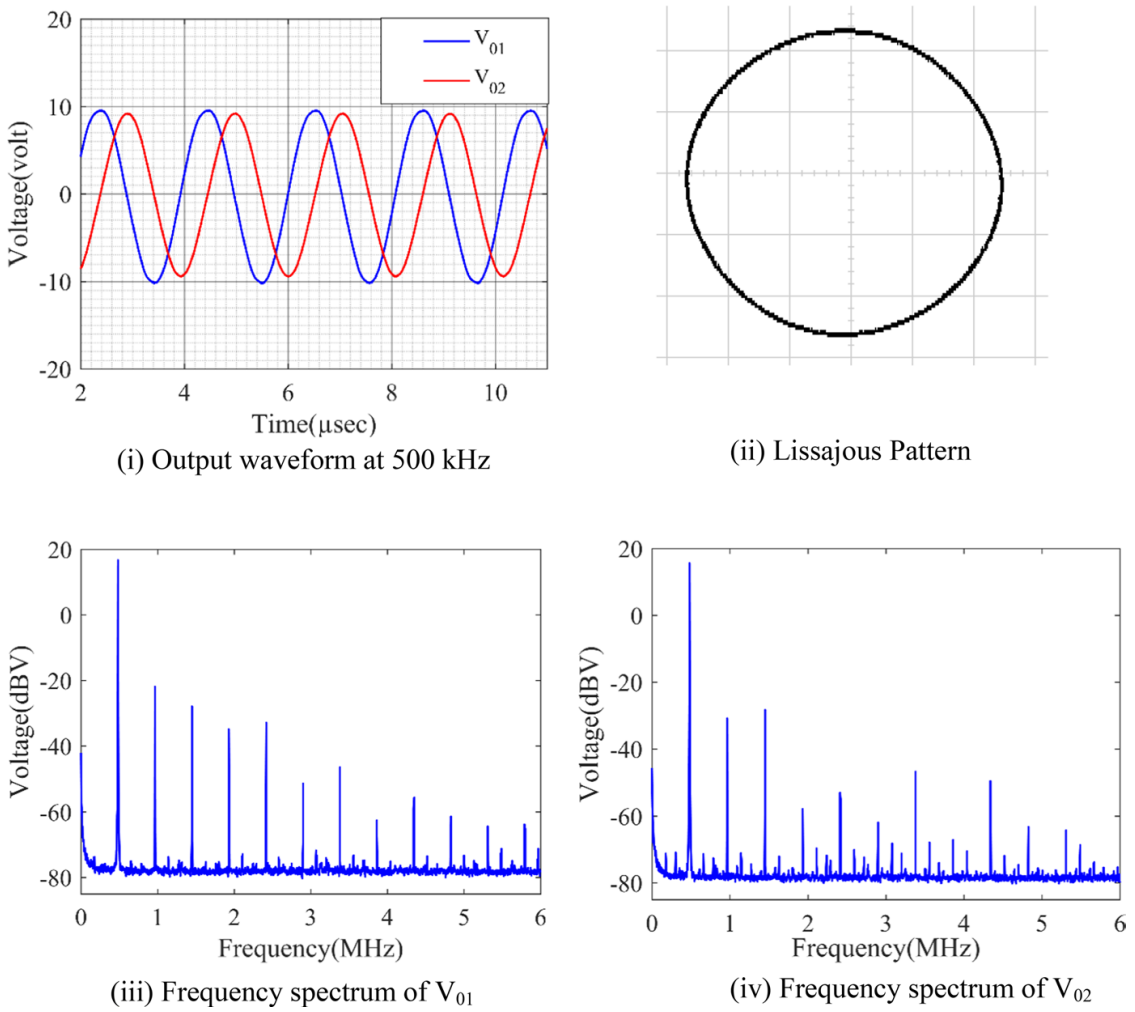


Fig. 7 Experimental results of proposed QO for Circuit 2

5 Experimental Results

For experimental verification of the proposed circuits, we have used a CDBA realization, using off-the-shelf available IC AD844 [23], as shown in Fig. 5, biased with $\pm$ 12 V.

All the circuits were bread-boarded using 5% tolerance resistors and 10% tolerance capacitors and tested for experimental verification using KEYSIGHT DSO-X3034T. The QOs were designed for a frequency of 500 kHz with following component values: $C_1 = C_2 = 100$ pF, $R_1 = 1$ kΩ, $R_2 = 1$ kΩ (100 Ω in series with 10 kΩ pot), $R_3 = 1.59$ kΩ for Circuit 1, 0.795 kΩ for Circuit 2, 3.18 kΩ for Circuit 3 and Circuit 4 and $R_4 = 3.18$ kΩ. The experimentally obtained output waveforms, Lissajous pattern and frequency spectrum for all the proposed QO circuits are displayed in Figs. 6, 7, 8 and 9, respectively.

Table 8 shows the FO, percentage total harmonic distortion (THD), phase angle and amplitude ratio of the proposed QOs. From Table 8, it may be observed that the percentage THD is less than 1% (except for circuit 1 where %THD corresponding to V_{02} is 1.23), the error in phase angle is less than 1. 5%, while the frequency error

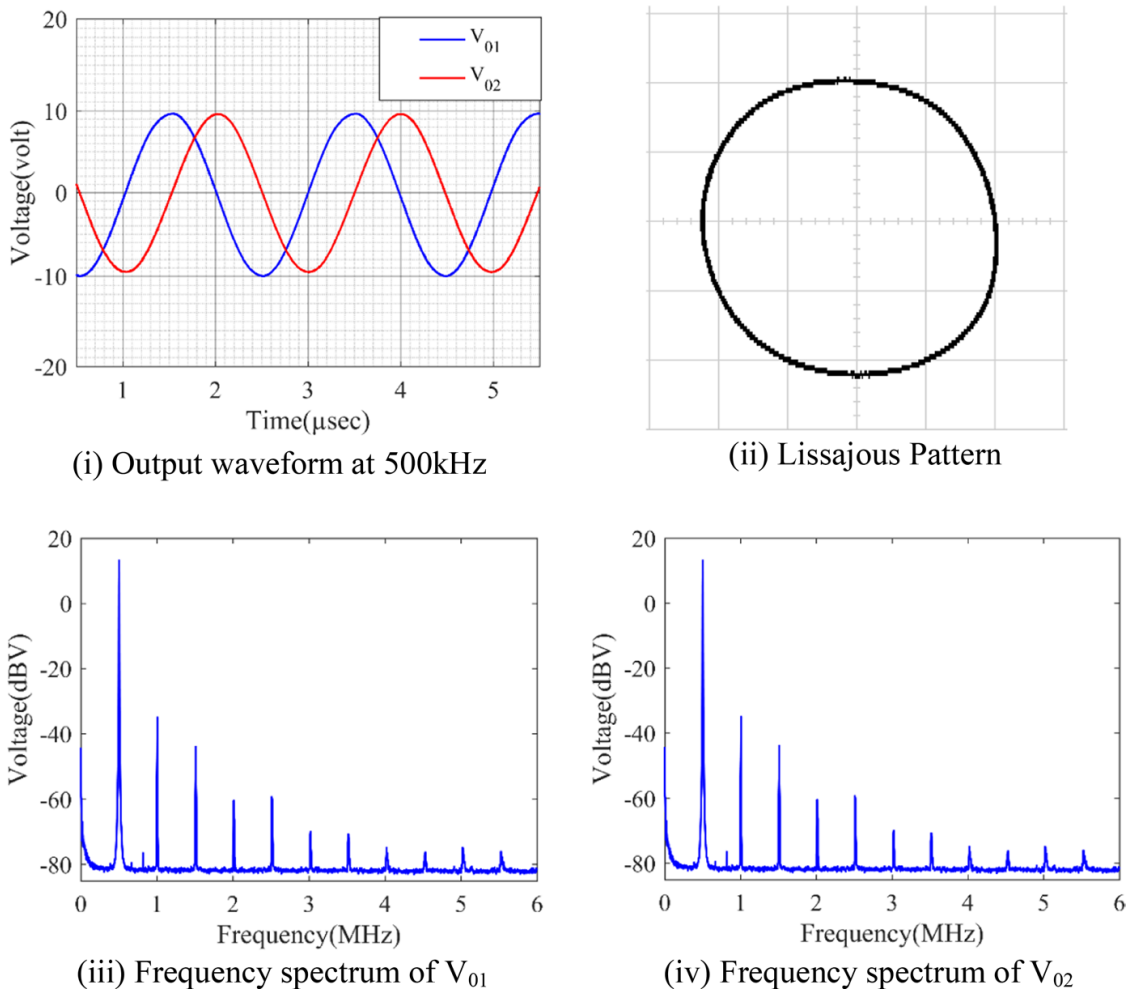


Fig. 8 Experimental results of proposed QO for Circuit 3

is less than 4%. The amplitude ratio is very close to the theoretical value for all QO circuits. These results, thus, establish the workability of the proposed QO circuits.

Figure 10 shows the variation of FO and percentage frequency error when the FO was changed by varying R_3 and R_4 simultaneously such that $R_3 = R_4$ for Circuit 3 and Circuit 4 ($R_4 = 2R_3$ for Circuit 1 and $R_4 = 4R_3$ for Circuit 2). However, it must be pointed out that if these resistors are replaced by identical MOS voltage controlled resistors (VCRs) by a common control voltage (V_c), then these oscillator circuits will provide a linear variation of FO with respect to V_c. The maximum frequency error is less than 5% for all QO circuits for the variation of frequency from 100 kHz to 1 MHz. The percentage frequency error is more for QOs proposed in Circuit 1 and Circuit 2 due to parasitic capacitances and resistors of z-terminals of the CDBAs, as these capacitors and resistors are left unabsorbed.

Figure 11 shows the variation in phase error with frequency, from where it is observed that the phase error in all proposed QOs is more for higher range of frequency of oscillations. Phase error for QOs proposed in Circuit 1 and Circuit 2 is more than 5%, when frequency of oscillation is more than 800 kHz; however, for QOs in Circuit 3 and Circuit 4, the phase error is within a limit of 2.5% up to 1 MHz.

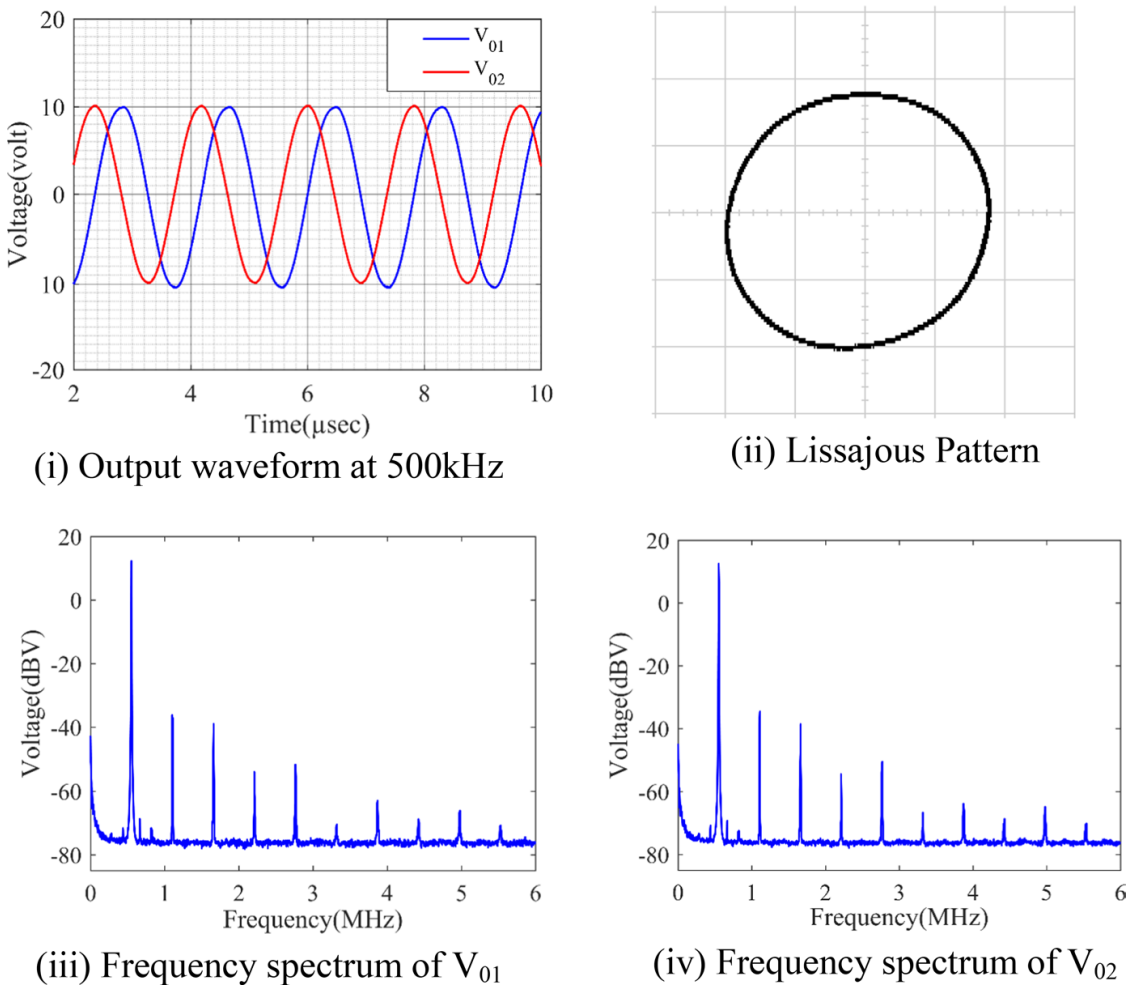


Fig. 9 Experimental results of proposed QO for Circuit 4

Table 8 FO, %THD, phase angle and amplitude ratio of output voltages of the proposed QOs

Circuit	Frequency(kHz)		%THD		Phase angle between V_{01} and V_{02}	Amplitude ratio of output voltages $\left(\frac{V_{02}}{V_{01}}\right)$	
	Theoretical (kHz)	Experimental (kHz)	V_{01}	V_{02}		Theoretical	Experimental
Circuit 1	500	486	0.83	1.23	88.65°	1	0.972
Circuit 2	500	484	1.4	1.34	88.80°	1	0.92
Circuit 3	500	489	0.51	0.52	89.78°	1	0.972
Circuit 4	500	487	0.5	0.58	89.6°	1	0.985

Figure 12 shows the variation of percentage total harmonic distortion (THD) of controlled output voltages V_{01} and V_{02} with frequency. From THD plot, it can be observed that %THD for all QOs is less than 3% up to around 800 kHz. The %THD increases if one moves beyond 800 kHz (except for Circuit 4).

All the proposed QOs have the tendency of amplitude control with frequency tuning. To keep constant equal voltage, the frequency is tuned by varying R_3 and

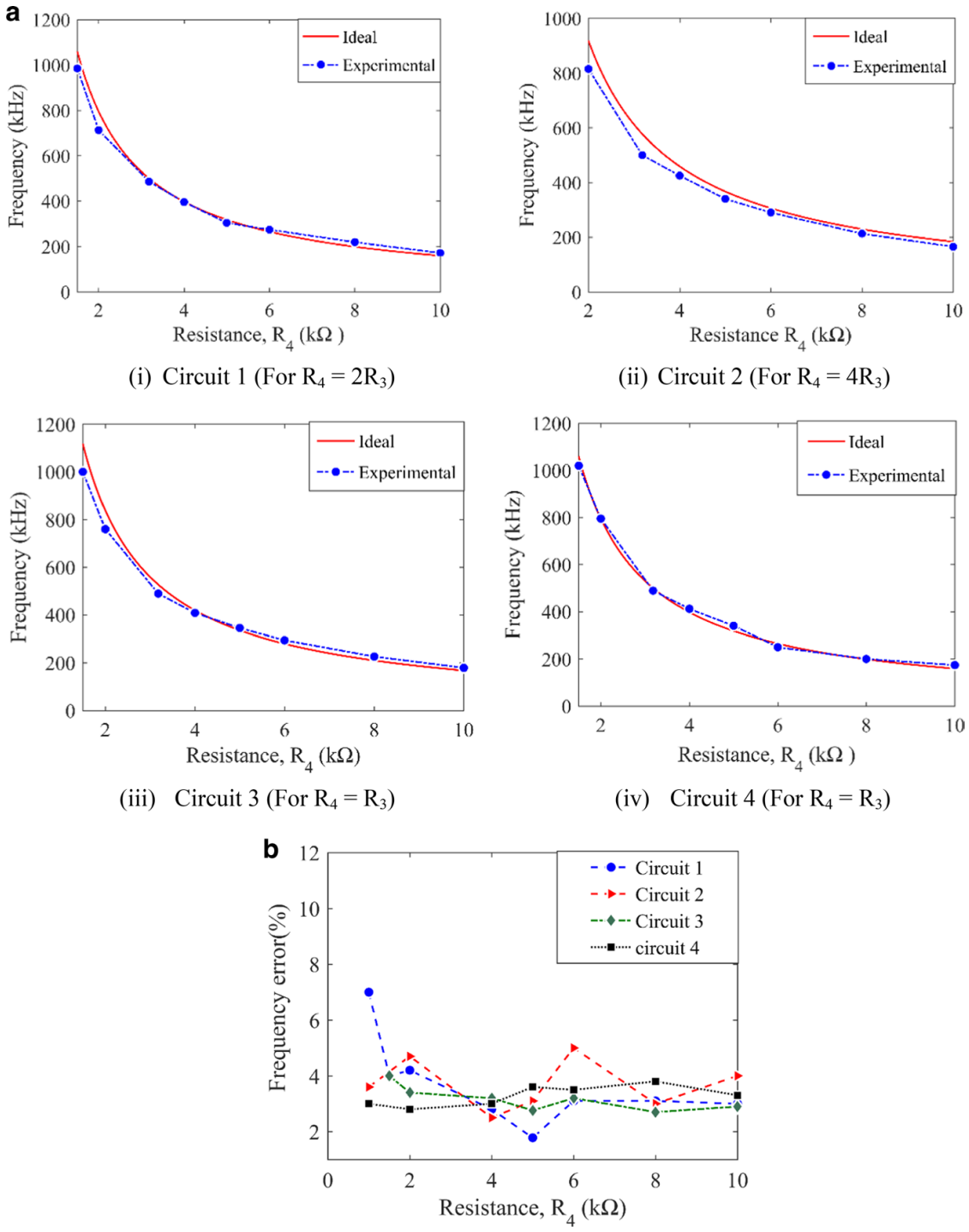


Fig. 10 Variation of FO and frequency error with R_4

R_4 simultaneously, keeping the amplitude ratio of the QOs equal to unity as per the relations described in Table 2. Figure 13 shows the variation of peak to peak output voltages with frequency when (i) frequency is tuned by varying resistance R_3 (keeping R_4 constant), (ii) frequency is tuned by varying resistance R_4 (keeping R_3 constant), and (iii) frequency is tuned by varying R_3 and R_4 simultaneously, as per desired gain given in Table 2. Under amplitude control conditions,

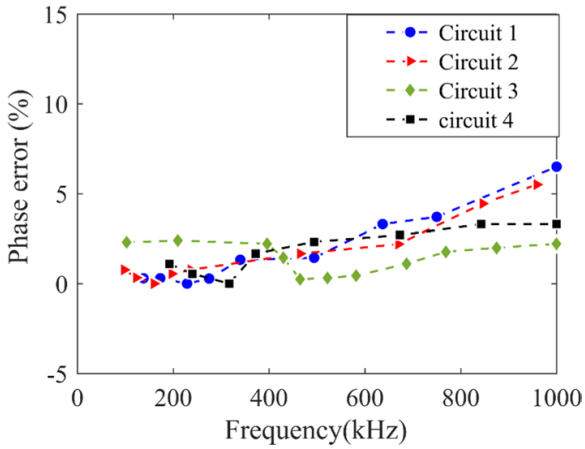


Fig. 11 Variation of phase error with frequency

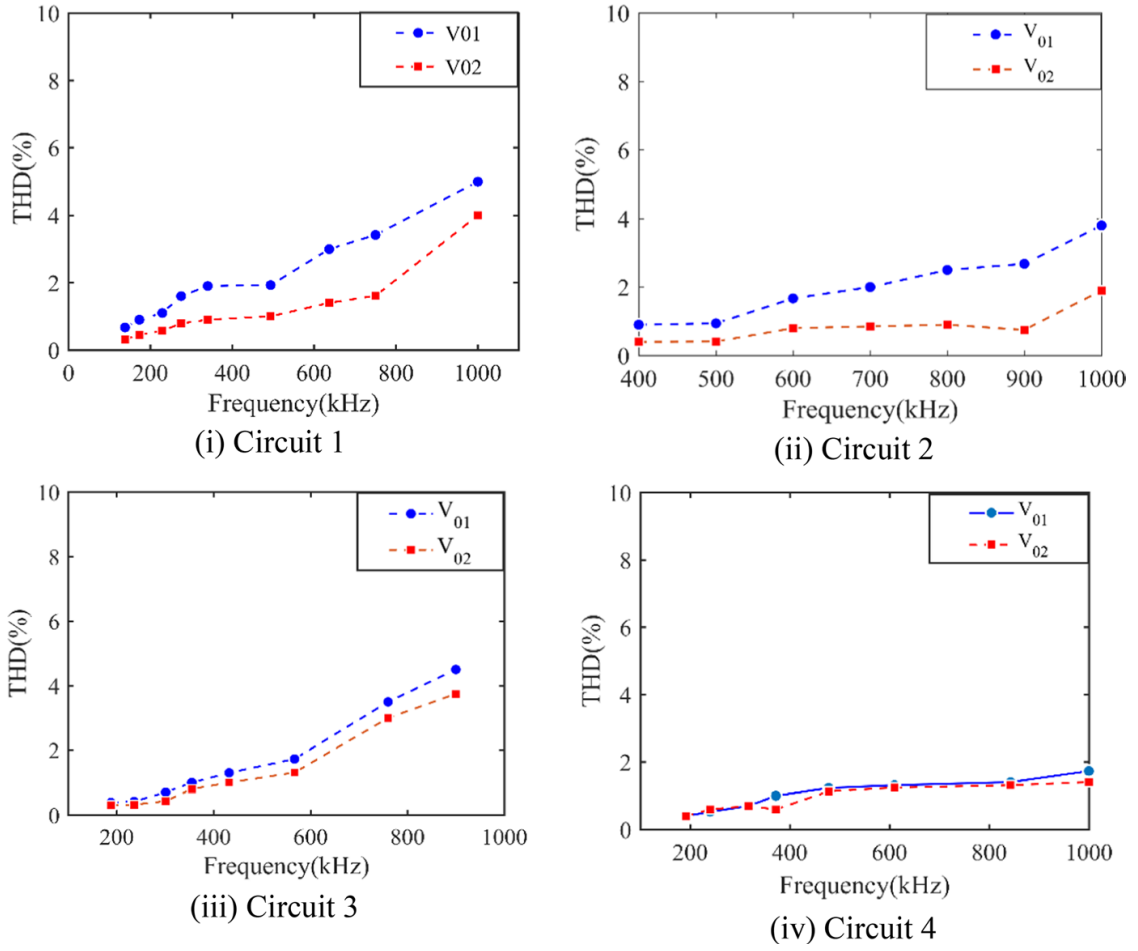


Fig. 12 Total harmonic distortion of the controlled quadrature output voltages

the two voltages V_{01} and V_{02} remain nearly equal and constant with respect to frequency; however, without the control, V_{01} is almost constant, while V_{02} changes from around 20 V to 5 V.

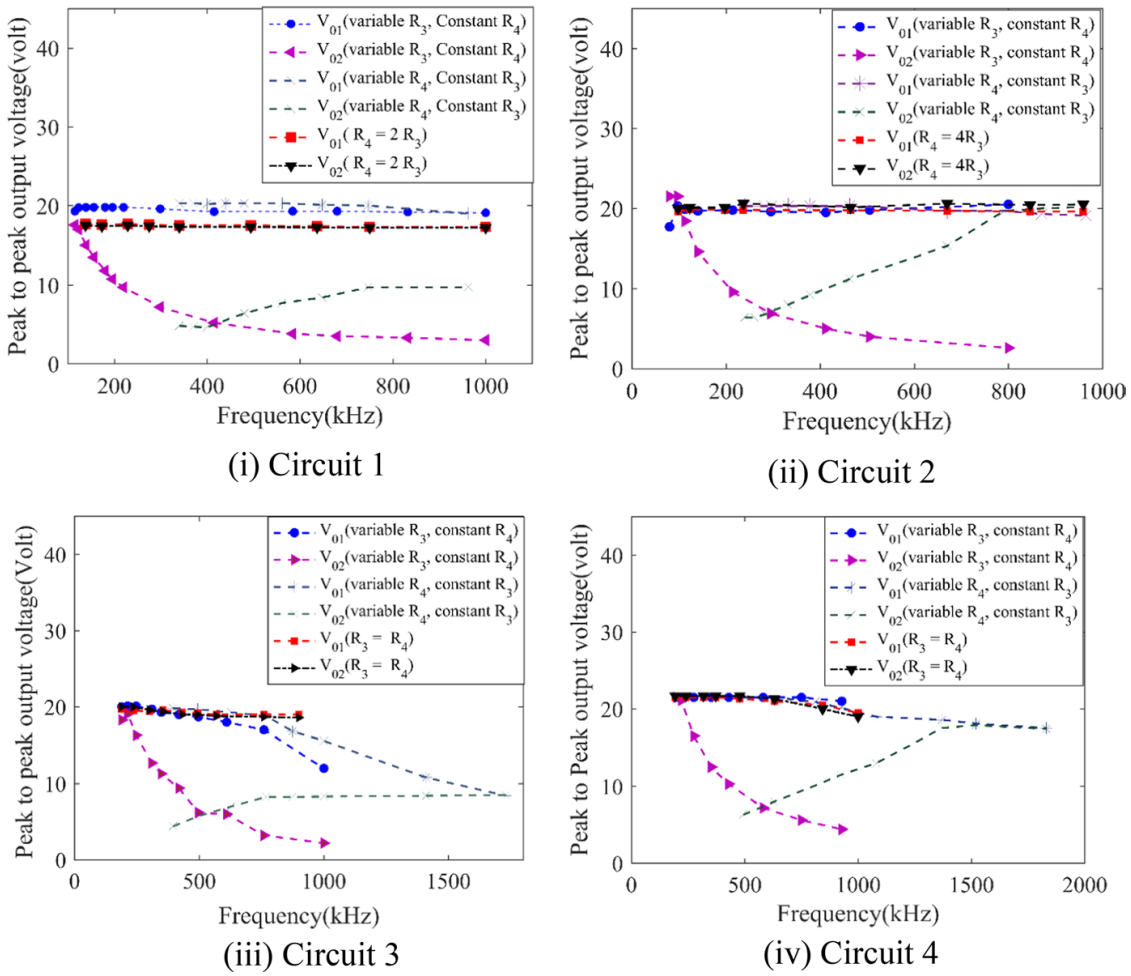


Fig. 13 Peak-to-peak amplitude of oscillations of quadrature oscillators under frequency tuning

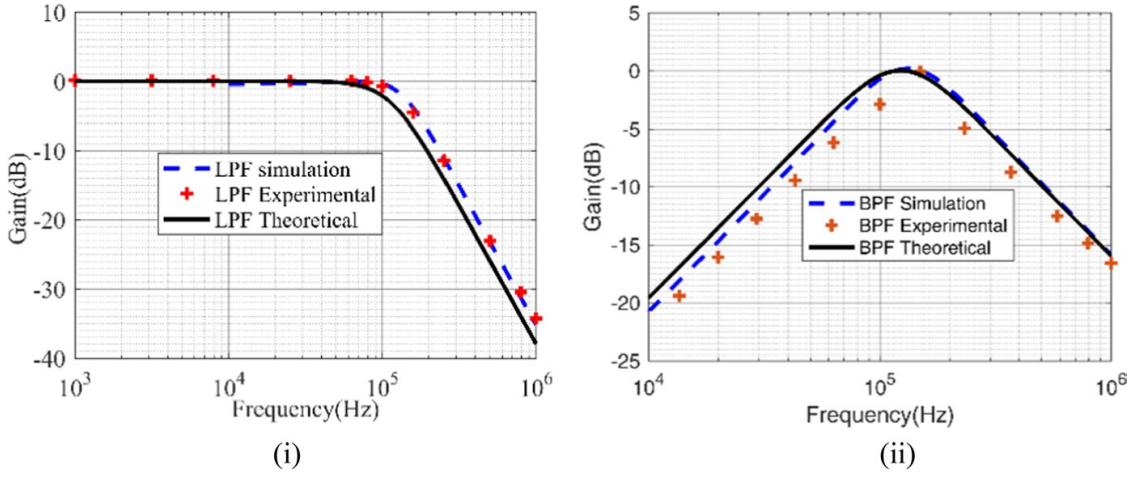


Fig. 14 Experimental frequency response (i) LPF (ii) BPF

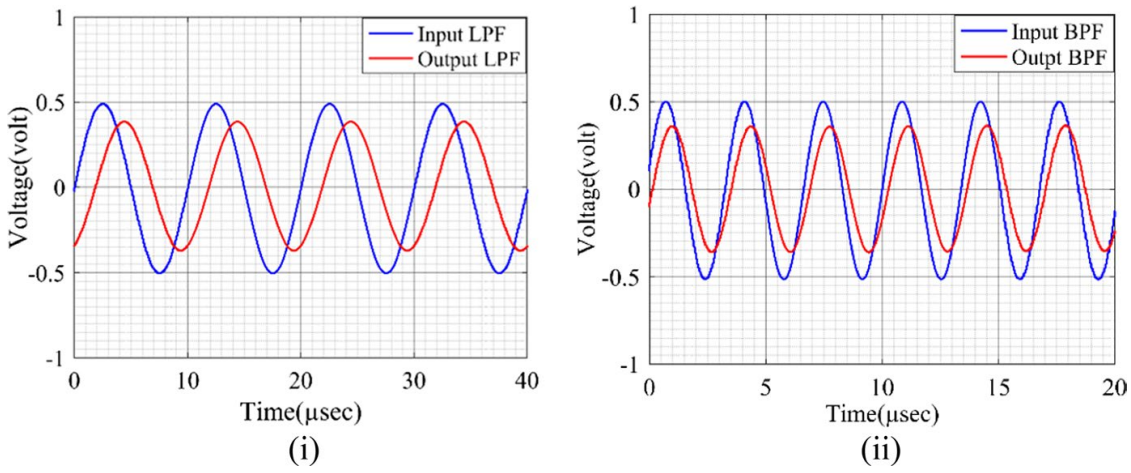


Fig. 15 Experimental transient response (i) LPF at 100 kHz and (ii) BPF at 148 kHz

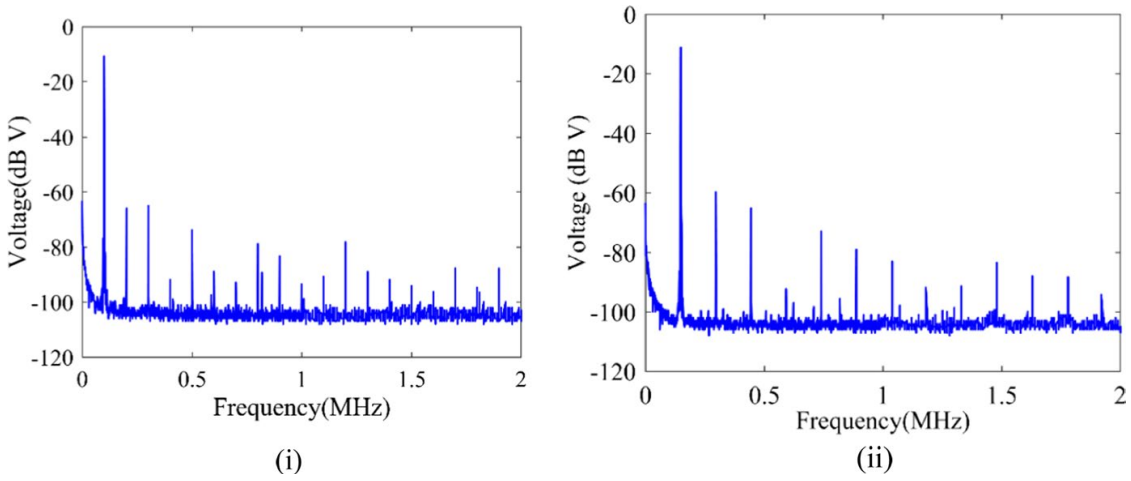


Fig. 16 Experimental frequency spectrum of (i) LPF and (ii) BPF

In the present investigation, no attempt was made to incorporate any scheme for amplitude stabilization/control. In fact, the authors believe that the development of an entirely CDBA-based amplitude control scheme constitutes an interesting problem for further research work.

A perusal of Figs. 6, 7, 8, 9, 10, 11, 12 and 13 reveals that, out of the four circuits proposed in this paper, the QO proposed in Circuit 4 offers minimum frequency error, minimum phase error, minimum %THD, better frequency stability and equal constant output voltage in the frequency range of 100 kHz to 1 MHz.

Some sample experimental results to validate the realization of LPF and BPF of the proposed quadrature oscillator circuits have also been presented for filters derived from Circuit 3. A sinusoid input signal of 1 V peak to peak was applied through R_2. The filters were designed for a cut-off frequency of 159 kHz and Q-factor of 1 with passive elements as $C_1 = C_2 = 1$ nF, $R_1 = R_2 = R_3 = R_4 = 1$ kΩ. Experimental frequency responses, transient responses and frequency spectrums of LPF and BPF are demonstrated in Figs. 14, 15 and 16, respectively. The

Table 9 Comparative study with previously reported QOs realized with modern active building blocks

References	ABB used	No. of ABBs used	No. of passive components used		Full decoupling of CO and FO	Inherent Amplitude control possible
			Resistor	Capacitor		
[1]	VD-DIBA	2	0	2 GC	No (only independent control of CO and FO)	No
[3]	VDIBA	2	0	2 GC	No (only independent control of CO and FO)	No
[4]	DBTA	1	2	3 GC	No (only independent control of CO and FO)	No
[5]	CDBA	2	4	2 GC	Yes	No
[6]	CDTA	1	1	2 (1 GC, 1 VGC)	No	No
[7]	CCTA	1	1	2 GC	No	No
[8]	CCCCTA	4	0	2 GC	Yes	No
[9]	LTI228	3	1	2 GC	Yes	Yes (using additional 4 resistors)
[10]	CDTA	2	0	2 (1 GC, 1 VGC)	No	No
[11]	CDBA	2	4	4 (2 GC, 2 VGC)	No	No
[12]	CDTA	2	4	2 VGC	Yes	No
[13]	CFTA	2	1	2 GC	No	No
[14]	CFTA	2	1	2 GC	No (only independent control of CO and FO)	No
[15]	CDBA	2	4 (ckt. A, B) 3 (ckt C, D)	2 GC	Yes, (ckt. A, B) independent control of CO and FO for ckt. C, D	No
[16]	CDBA	2	4	2 GC (For Fig. 3 1 GC, 1 VGC)	No (only independent control of CO and FO)	Yes
[17]	CDTA	2	1	2 GC	No (only independent control of CO and FO)	No
[18]	MCDTA	1	0	2 GC	No	No
[19]	ZC-CFTA	3	1	2 GC	No (only independent control of CO and FO)	No
[20]	CDBA	2	3	2 GC	No (only independent control of CO and FO)	No

Table 9 (continued)

References	ABB used	No. of ABBs used	No. of passive components used		Full decoupling of CO and FO	Inherent Amplitude control possible
			Resistor	Capacitor		
[21]	VDBA	3	2	3 GC	No	No
[22]	CFTA	2	1	2 GC	No (only independent control of CO and FO)	No
[23]	CDBA	2	4	2 GC/2 VGC	Yes	No
[24]	CFTA	3	0	2 GC	No (only independent control of CO and FO)	No
[25]	VDTA	2	0	2 GC	Yes	No
[26]	MO-CCDTA	1	1	2 GC	No (only independent control of CO and FO)	No
[27]	ZG-CFTA	2	0	2 GC	No	No
[28]	VD-DIBA	2	0	2 GC	No (only independent control of CO and FO)	No
[32]	CFCC	2	5	2 GC	No (only independent control of CO and FO)	No
[33]	VDCC	1	2	2 GC	No (only independent control of CO and FO)	No
[34]	CCTA	2	1	2 GC	No	No
[35]	CDTA	3	0	2 GC	No (only independent control of CO and FO)	No
[36]	CFOA	2	3	2 GC	No (only independent control of CO and FO)	No
[37]	CDBA	2	4	2 GC	Yes	No
[38]	CDBA	2	3	3 (2 GC, VGC)	No (only independent control of CO and FO)	No
[39]	CDBA	2	3	2 GC	No (only independent control of CO and FO)	No
[40]	CDTA	3	0	2 GC	No (only independent control of CO and FO)	No
[41]	ZC-CFTA	4	0	2 GC	No (only independent control of CO and FO)	No
Proposed QOs	CDBA	2	4 (ckt. 1,2)/5 (ckt. 3, 4)	2 (1 GC, 1 VGC for ckt. 1/2 VGC for ckt. 2/2 GC for ckt. 3–4)	Yes	Yes

GC grounded capacitor, *VGC* virtually grounded capacitor, *ckt* circuit

percentage THD obtained for LPF and BPF is 0.58 and 0.64, respectively, which are quite low.

6 Comparison with the Existing Quadrature Oscillator Circuits

Over the past two decades, several QOs designs employing different modern active building blocks have been reported. It is worthwhile to compare the proposed QO circuits with the existing QO circuits realized with modern active building blocks (excluding, the classical operational amplifiers, operational transconductance amplifier and the current conveyor) to put the QO circuits presented in this paper in proper perspective. Table 9 summarizes the salient features of these QOs.

From the above table, it is observed that apart from the QOs reported in [9], no configuration(s) of QOs realized with modern active building blocks, reported earlier in the open literature, possess the advantageous features of full decoupling of FO and CO and inherent amplitude control of the quadrature output voltage without external adjustment, a feature which is simultaneously available in all the four circuits presented in this paper.

7 Concluding Remarks

Four new, fully uncoupled QO circuits have been presented which possess several additional features and functionalities in contrast to the quoted set of 12 QOs published earlier in [23]. The results of experimental investigations have also been included to support the theory. The attractive features of the QO circuits are: (i) decoupled control of CO and FO through separate resistors, (ii) low harmonic distortion, (iii) high-frequency stability factor, (iv) use of grounded/virtually grounded capacitors, (v) low phase error, (vi) amplitude adjustment of output voltages without using any external control circuitry and (vii) possibility of very low-frequency operation in some of the circuits. The development of an entirely CDBA-based amplitude control scheme for quadrature oscillators has not been attempted in this paper and thus is an interesting research problem open for investigation. In addition to this, the proposed circuits can also realize tuneable LPF and BPF at the output terminals of each CDBA (if we break feedback loop at point 'm' and apply sinusoidal input at terminal-p through a resistor R_2). The proposed filter outputs are supported by sample experimental frequency responses, transient responses and frequency spectrums. The presented QO circuits thus enrich and enlarge the repertoire of the fully decoupled quadrature sinusoidal oscillators existing in the open literature. Lastly, we would like to comment that all the circuits presented in [23] can also be converted easily to facilitate the utilization of intrinsic property (current differencing) of CDBA as demonstrated by an example in Appendix.

Acknowledgement The authors gratefully acknowledge Professor Raj Senani, Department of Electronics and Communication Engineering, Netaji Subhash University of Technology, New Delhi, India, for his help and valuable suggestions in the preparation of this manuscript. The authors also like to thank the anonymous reviewers for their constructive comments/suggestions in improving the quality of the manuscript. The experimental work reported in this paper was carried out in the Analog Signal Processing Laboratory, Department of Electrical Engineering, Delhi Technological University, New Delhi-110042, India. All data generated during this study are included in this article and have no further associated data with it.

Appendix

Consider an exemplary Circuit 2 from Fig. 5 of [23] which is reproduced here in Fig. 17.

The CO and FO of this circuit are given by:

$$\text{CO: } 2R_2 \leq R_6 \text{ and FO: } \omega = \frac{1}{\sqrt{R_3 R_9 C_1 C_{10}}}$$

The modified version of this circuit as per the methodology of this paper will take the form as shown in Fig. 17.

An analysis of this circuit reveals its CO and FO as:

$$\text{CO: } 2R_2 \leq R_6 \text{ and FO: } \omega = \sqrt{\frac{R_9 - R_{10}}{C_1 C_2 R_3 R_9 R_{10}}}$$

If we choose $R_{10} = \frac{R_9}{n}$, then $\omega = \sqrt{\frac{n-1}{C_1 C_2 R_3 R_9}}$, where $n > 1$. If we take $n = 2$, then $\omega = \sqrt{\frac{1}{C_1 C_2 R_3 R_9}}$

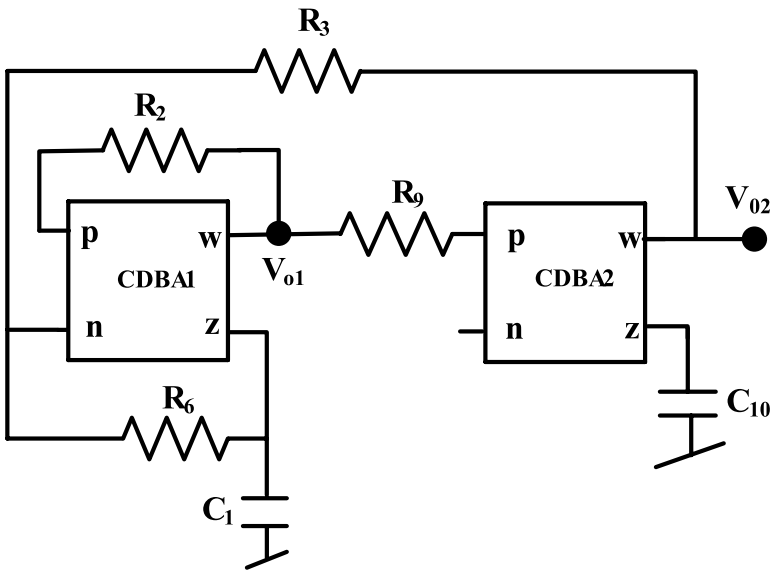


Fig. 17 QO Circuit 2 from Fig. 5 of [23]

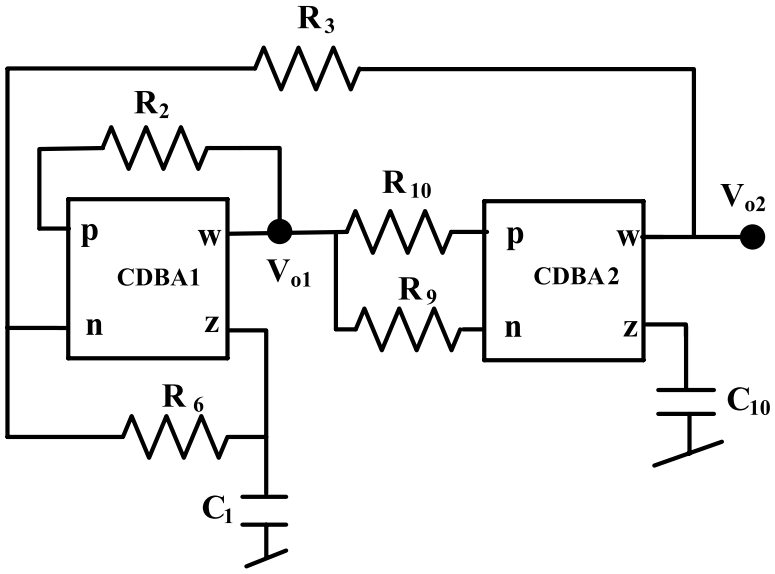


Fig. 18 Modified version of the QO circuit shown in Fig. 17

From where it is clear that (i) the intended property of independent control of both CO and FO remains unaffected even in the modified circuit but now all the terminals of both the CDBAs are effectively utilized.

(ii) Also, as FO can be expressed as $\omega = \sqrt{\frac{\Delta R}{C_1 C_2 R_3 R_9 R_{10}}}$, where $\Delta R = R_9 - R_{10}$. From CO, and FO, it can be observed that the given QO circuit can be used to generate fully uncoupled amplitude adjustable low-frequency sinusoidal oscillation (LFO).

It is thus clear that as many as 12 QO circuits from Fig. 5 of [23] can be similarly modified at the cost of only one additional resistor (Fig. 18).

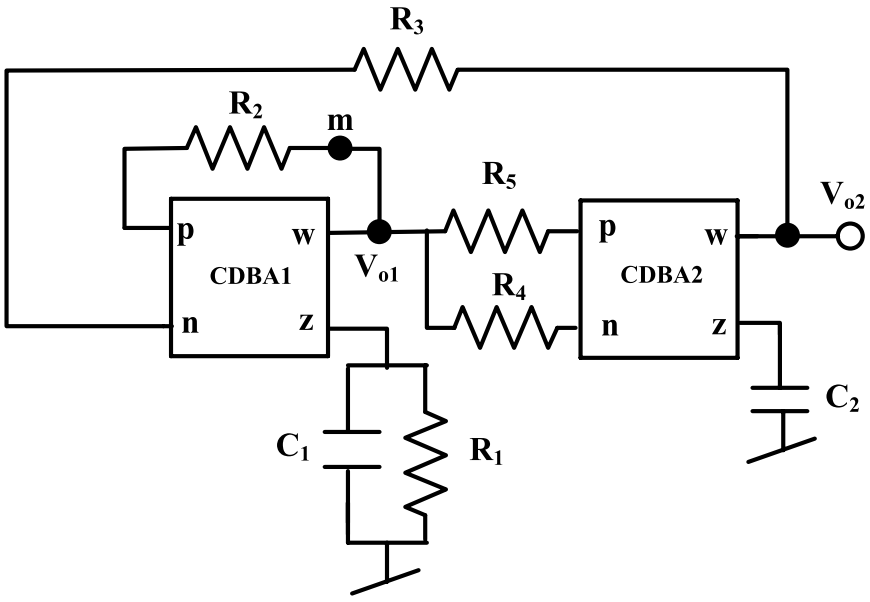


Fig. 19 Modified version of Circuit 3 of Fig. 2

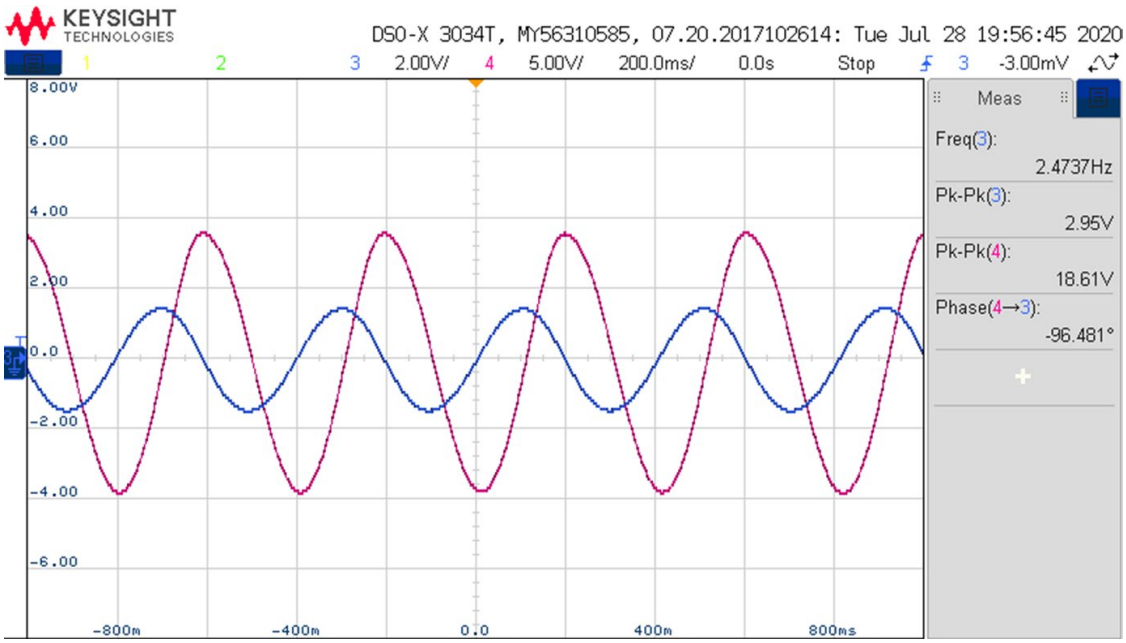


Fig. 20 Experimental output waveforms of 2.47 Hz for Fig. 19

Figure 19 shows the modified version of the circuit 3 given in Fig. 2.
An analysis of this circuit reveals its CO and FO as:

$$\text{CO: } R_2 \leq R_1 \text{ and FO: } \omega = \sqrt{\frac{R_4 - R_5}{C_1 C_2 R_3 R_4 R_5}}$$

For this circuit, we have produced a low-frequency output waveform of 2.47 Hz against a designed frequency of 2.7 Hz. The circuit parameters selected are as $C_1 = C_2 = 100$nF, $R_1 = 10$ kΩ, $R_2 = 10$ k Ω (100Ω in series with 22 kΩ pot), $R_3 = 100$ kΩ, $R_4 = 10.25$ kΩ and $R_5 = 10.22$ kΩ. Experimentally observed waveforms are shown in Fig. 20. The peak-to-peak output voltages V_{01} and V_{02} recorded are as 18.6 V and 2.95 V. The phase difference between V_{01} and V_{02} is 96.48°.

It is thus clear that as many as 12 QO circuits from Fig. 5 of [23] can be similarly modified as LFO QOs with the addition of only one extra resistor.

References

1. J. Bajer, J. Vavra, D. Biolek, Voltage-mode quadrature oscillator using VD-DIBA active elements, in 2014 IEEE Asia Pacific Conference on Circuits and Systems (Ishigaki), pp. 197–200 (2014)
2. D.R. Bhaskar, Realisation of second-order sinusoidal oscillator/filter with non-interacting controls using CFAs. Frequenz **57**(1–2), 12–14 (2003)
3. O. Channumsin, W. Tangsrirat, VDIBA-based sinusoidal quadrature oscillator. Przeglad Electrotechniczny **93**(3), 248–250 (2017)
4. N. Herencsar, J. Koton, K. Vrba, A. Lahiri, New voltage-mode quadrature oscillator employing single DBTA and only grounded passive elements. IEICE Electron. Express **6**(24), 1708–1714 (2009)

5. J.W. Horng, CDBA-based single resistance controlled quadrature oscillator employing grounded capacitors. IEICE Trans. Fundam. Electron. Commun. Comput. Sci. **E-85A**(6), 1416–1419 (2002)
6. W. Jaikla, M. Siripruchyanun, J. Bajer, D. Biolek, A simple current-mode quadrature oscillator using single CDTA. Radioengineering **17**(4), 33–40 (2008)
7. W. Jaikla, P. Silapan, C. Chanapromma, M. Siripruchyanun, Practical implementation of CCTA based on commercial CCII and OTA, in International symposium on intelligent signal processing and communications systems—ISPACS (2009), pp. 1–4
8. W. Jaikla, A. Noppakarn, S. Lawanwisut, New gain controllable resistor-less current-mode first order allpass filter and its application. Radioengineering **21**(1), 312–316 (2012)
9. W. Jaikla, S. Adhan, P.S. Kumngern, Current/voltage controlled quadrature sinusoidal oscillators for phase sensitive detection using commercially available IC. Sensors **20**(5), 1319 (2020)
10. J. Jin, C. Wang, Current-mode universal filter and quadrature oscillator using CDTAs. Turk. J. Electr. Eng. Comput. Sci. **22**(2), 276–286 (2014)
11. A.U. Keskin, C. Aydin, E. Hancioglu, C. Acar, Quadrature oscillator using CDBA. Frequenz **60**, 21–23 (2005)
12. A.U. Keskin, D. Biolek, Current mode quadrature oscillator using current differencing transconductance amplifiers (CDTA). IEE Proc. Circuits Devices Syst. **153**(3), 214–218 (2006)
13. S. Kumari, M. Gupta, Design and analysis of high transconductance current follower transconductance amplifier (CFTA) and its applications. Analog Integr. Circuits Sig. Process **93**(3), 489–506 (2017)
14. M. Kumngern, P. Lamun, S. Junnapiya, CFTA-based electroniccally tunable quadrature sinusoidal oscillator, in Proceeding of the International Electrical Engineering Congress (iEECON), IEEE, pp. 1–4 (2014)
15. A. Lahiri, New realizations of voltage-mode quadrature oscillators using current differencing buffered amplifiers. J. Circuits Syst. Comput. **19**(05), 1069–1076 (2010)
16. A. Lahiri, Low-frequency quadrature sinusoidal oscillators using current differencing buffered amplifiers. Indian J. Pure Appl. Phys. **49**, 423–428 (2011)
17. A. Lahiri, New current-mode quadrature oscillators using CDTA. IEICE Electron. Express **6**(3), 135–140 (2009)
18. Y. Li, Electronically tunable current-mode quadrature oscillator using single MCDTA. Radioengineering **19**(4), 667–671 (2010)
19. Y.A. Li, Electronically tunable current-mode biquadratic filter and four-phase quadrature oscillator. Microelectron. J. **45**(3), 330–335 (2014)
20. S. Maheshwari, I.A. Khan, Novel single resistance controlled quadrature oscillator using two CDBAs. J. Act. Passive Electron. Devices **2**, 137–142 (2007)
21. C. Malhotra, V. Ahalawat, V. Kumar, R. Pandey, N. Pandey, Voltage differencing buffered amplifier based quadrature oscillator, in 2016 IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES), pp. 1–4 (2016)
22. S. Maneewan, B. Sreewirote, W. Jaikla, A current mode quadrature oscillator using a minimum number of active and passive components, in Proceeding of the IEEE International Conference on Vehicular Electronics and Safety (ICVES), pp. 312–315 (2011)
23. J.K. Pathak, A.K. Singh, R. Senani, Systematic realisation of quadrature oscillators using current differencing buffered amplifiers. IET Circuits Devices Syst. **5**(3), 203–211 (2011)
24. P. Phatsornsiri, P. Lamun, Tunable current-mode quadrature oscillator using CFTAs and grounded capacitors, in Proceeding of the 12th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), IEEE, pp. 1–4 (2015)
25. D. Prasad, M. Srivastava, D.R. Bhaskar, Electronically controllable fully-uncoupled explicit current-mode quadrature oscillator using VDTAs and grounded capacitors. Circuits Syst. **4**(2), 169–172 (2013)
26. D. Prasad, D.R. Bhaskar, A.K. Singh, Electronically controllable grounded capacitor current-mode quadrature oscillator using single MO-CCCDTA. Radio Eng. **20**(1), 354–359 (2011)
27. D. Prasertsom, W. Tangsrirat, Current gain controlled CFTA and Its application to resistorless quadrature oscillator, in Proceeding of the 9th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), pp. 1–4 (2012)

28. K.L. Pushkar, Electronically controllable quadrature sinusoidal oscillator using VD-DIBAs. Circuits Syst. **9**(3), 41–48 (2018)
29. R. Senani, D.R. Bhaskar, V.K. Singh, R.K. Sharma, *Sinusoidal Oscillators and Waveform Generators Using Modern Electronic Circuit Building Blocks* (Springer, Berlin, 2016)
30. R. Senani, D.R. Bhaskar, A.K. Singh, V.K. Singh, *Current Feedback Operational Amplifiers and their Applications* (Springer, New York, 2013)
31. R. Senani, D.R. Bhaskar, A.K. Singh, *Current Conveyors: Variants, Applications and Hardware Implementations* (Springer, Berlin, 2014)
32. A.K. Singh, P. Kumar, R. Senani, Current-mode quadrature oscillator using CFCC. Am J Electr Electron Eng **6**(3), 85–89 (2018)
33. R. Sotner, V. Kledrowetz, J. Jerabek, R. Prokop, Simple CMOS voltage differencing current conveyor-based electronically tunable quadrature oscillator. Electron. Lett. **52**(12), 1016–1018 (2016)
34. R. Sotner, J. Slezak, T. Dostal, Tunable oscillator using two CCTA-s and only grounded passive elements, in Proceeding of the 9th International Conference on Environment and Electrical Engineering (EEEIC), pp. 447–450 (2010)
35. S. Summart, C. Thongsopa, W. Jaikla, New current-controlled current-mode sinusoidal quadrature oscillators using CDTAs. AEU Int. J. Electron. Commun. **69**(1), 62–68 (2015)
36. W. Tangsrirat, W. Surakampontorn, Single-resistance-controlled quadrature oscillator and universal biquad filter using CFOAs. AEU Int. J. Electron. Commun. **63**(12), 1080–1086 (2009)
37. W. Tangsrirat, T. Pukkalanun, W. Surakampontorn, CDBA-based universal biquad filter and quadrature oscillator. Act. Passive Electron. Compon. **6**, 12 (2008)
38. W. Tangsrirat, S. Pisitchalermpong, CDBA-based quadrature sinusoidal oscillator. Frequenz **61**, 102–104 (2007)
39. W. Tangsrirat, D. Preasertsom, T. Piyatat, W. Surakampontom, Single resistance controlled quadrature oscillator using current differencing buffered amplifiers. Int. J. Electron. **95**(11), 1119–1126 (2008)
40. W. Tangsrirat, Current differencing transconductance amplifier-based current-mode four-phase quadrature oscillator. Indian J. Eng. Mater. Sci. **14**, 289–294 (2007)
41. W. Tangsrirat, P. Mongkolwai, T. Pukkalanun, Current-mode high-Q bandpass filter and mixed-mode quadrature oscillator using ZC-CFTAs and grounded capacitors. India J. Pure Appl. Phys. **50**, 600–607 (2012)

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

Technical Note

Landslides
DOI 10.1007/s10346-020-01602-4
Received: 14 June 2020
Accepted: 4 December 2020
© The Author(s) 2020

Sansar Raj Meena · Omid Ghorbanzadeh · Cees J. van Westen · Thimmaiah Gudiyangada Nachappa · Thomas Blaschke · Ramesh P. Singh · Raju Sarkar

Rapid mapping of landslides in the Western Ghats (India) triggered by 2018 extreme monsoon rainfall using a deep learning approach

Abstract Rainfall-induced landslide inventories can be compiled using remote sensing and topographical data, gathered using either traditional or semi-automatic supervised methods. In this study, we used the PlanetScope imagery and deep learning convolution neural networks (CNNs) to map the 2018 rainfall-induced landslides in the Kodagu district of Karnataka state in the Western Ghats of India. We used a fourfold cross-validation (CV) to select the training and testing data to remove any random results of the model. Topographic slope data was used as auxiliary information to increase the performance of the model. The resulting landslide inventory map, created using the slope data with the spectral information, reduces the false positives, which helps to distinguish the landslide areas from other similar features such as barren lands and riverbeds. However, while including the slope data did not increase the true positives, the overall accuracy was higher compared to using only spectral information to train the model. The mean accuracies of correctly classified landslide values were 65.5% when using only optical data, which increased to 78% with the use of slope data. The methodology presented in this research can be applied in other landslide-prone regions, and the results can be used to support hazard mitigation in landslide-prone regions.

Keywords Landslides · Convolutional neural network (CNN) · Deep learning · Western Ghats

Introduction

Landslides are one of the most devastating natural disasters in the mountainous regions around the world. Landslides severely damage infrastructure, cause a loss of life and properties, and impact the daily life of people living in the affected regions (Juang et al. 2019). Landslides are of various types, such as debris slides, rock falls, spreads, debris flow, and lahars (Cruden and Varnes 1996). According to Cruden and Varnes (1996), a landslide is 'the movement of a mass of rock, debris or earth down a slope'. The occurrence of landslides depends on the local terrain, geology and geomorphology of the area, soil types, tectonics, land use, and land cover. In the seismically active regions, landslides are commonly triggered by earthquakes, slope deformation, rock mass movements, and extreme rainfall events (Guzzetti et al. 1999). The situation is worsened by human activities, as instance the development of road networks with road cuts is a widely acknowledged predisposing factor in hilly areas (Das et al. 2011; Xu et al. 2017).

Several approaches have been developed for mapping landslides (Guzzetti et al. 2012; Lu et al. 2011; Martha et al. 2010; Meena et al. 2019; Pradhan et al. 2006; Prakash et al. 2020). The rapid mapping of landslides after an event is still a challenge for disaster management despite the availability of high-resolution satellite images and the algorithms for landslide detection.

Duro et al. (2012) used remote sensing data and semi-automated feature extraction of machine learning models in both pixel- and object-based environments for landslide extraction. In the last decade, the object-based approach has become more common (Jin et al. 2019; Liu et al. 2019; Shahabi et al. 2019; Tavakkoli Piralilou et al. 2019). The recent advancements in the performance of computing platforms have resulted in the development of several machine learning models, including deep learning methods (DLMs). Of the developed DLMs, the deep convolutional neural networks (DCNNs) have been especially widely used for classification and segmentation of satellite images and object detection (Du et al. 2019; Qayyum et al. 2019).

The use of CNN models has yielded very promising results for object classification from aerial images, but only a few studies have assessed landslide detection using the CNN model (see Table 1). Chen et al. (2018) used D-CNN (deep convolutional neural networks) for automated landslide detection in mountainous regions using multi-temporal remote sensing data. Lei et al. (2019) optimized FCN-PP (fully convolutional network within pyramid pooling) for landslide inventory mapping and compared the results with other models, such as the ELSE (employed edge-based level set evolution), RLSE (region-based level set evolution), CDMRF (change detection-based on Markov random field), and CDFFCM (change detection-based fast fuzzy c-means clustering). Ye et al. (2019) used hyperspectral data for landslide detection using DLWC (deep learning with constraints), SID (spectral information divergence), SAM (spectral angle match), and SVM (support vector machine) (Eskandari et al. 2020). Ghorbanzadeh et al. (2019a) evaluated the performance of different CNN models for landslide detection and compared these with three other ML models, namely, ANN, SVM, and RF, using the elevation factor coupled with remote sensing data. Recent studies used different elevation factors combined with remote sensing data for landslide detection using deep learning approaches (Liu et al. 2020; Prakash et al. 2020; Sameen and Pradhan 2019).

In this study, we used the CNN model to detect landslides caused by the extreme rainfall event of August 2018 in the Kodagu district of Karnataka state in the Western Ghats of peninsular India. The extreme rainfall caused deadly floods and landslides in the region (Martha et al. 2019), severely impacting the lives of the local population. Martha et al. (2019) carried out a rapid mapping of the landslides in the affected area using OBIA with Resourcesat-2 LISS-IV images (5.8 m spatial resolution) and reported a total of 771 landslides within an area of 7.1 million m^2. In this study, we used remote sensing based on 3-m PlanetScope Dove optical satellite imagery and 12.5-m ALOS PALSAR digital elevation data using the CNN model for landslide detection. We compared the resulting landslide inventory based on the CNN model

Published online: 05 January 2021

Table 1 Overview of some recent published studies on the automated mapping of landslides using deep learning approaches

Study	Main objective	Algorithms used	Topographical feature used	Accuracy evaluation methods
Chen et al. (2018)	An automated approach for landslides detection	DCNN	Slope	CE, DP, QP
Lei et al. (2019)	Optimization of (FCN-PP) for landslide inventory mapping	ELSE, RLSE, CDMRF, CDFFCM, CNN, FCN, U-Net, FCN-PP	N/A	Precision, recall, F1-score, OE, accuracy
Ye et al. (2019)	Landslide detection using hyperspectral remote sensing data and comparison of conventional methods with DLWC	DLWC, SVM, SID, SAM	Slope	Overall accuracy, Kappa coefficient, accuracy
Ghorbanzadeh et al. (2019a)	Comparison of different sample patch sizes for landslide detection using deep learning and machine learning	CNN, SVM, D-CNN, RF, ANN	Plan curvature, slope, slope aspect,	Precision, recall, F1-score, mIOU
Sameen and Pradhan (2019)	Landslide detection using residual networks	ResNet, CNN	Altitude, slope, slope aspect, total curvature	Training accuracy, validation accuracy, F1-score, mIOU
Ghorbanzadeh et al. (2019b)	Landslide detection using UAV-derived VHR imagery and topographical factors	CNN	Slope	PPV, TPR, F1-score, OPR, UPR, mIOU
Prakash et al. (2020)	Comparison of pixel-based, object-based, and deep learning methods for landslide detection	RF, ANN, LR, U-Net + ResNet	Hill shade, slope, slope aspect, terrain roughness, curvature, Valley depth, TWI	Accuracy, F1-score, MCC, POD, POFD
Liu et al. (2020)	Post-earthquake landslide extraction using the U-Net model	U-Net, U-Net + ResNet	DSM, slope, slope aspect	Precision, recall, F1-score, mIOU

UAV unmanned aerial vehicle, *CNN* convolutional neural network, *D-CNN* deep convolution neural network, *FCN* fully convolutional network, *FCN-PP* fully convolutional network within pyramid pooling, *ELSE* employed edge-based level set evolution, *RLSE* region-based level set evolution, *SVM* support vector machine, *LR* logistic regression, *CDMRF* change detection-based on Markov random field, *CDFFCM* change detection-based fast FCM, *RF* random forest, *ResNet* residual networks, *DLWC* deep learning with constrains, *ANN* artificial neural networks, *SID* spectral information divergence, *SAM* spectral angle match

with the manually delineated polygons. Further, we made efforts to enhance the accuracy of the detected landslide polygons by using different training data within a simple CNN architecture.

Study area

Kodagu, also known as Coorg, is a rural district in the state of Karnataka, India, covering an area of 4102 km^2. The study area is located on the eastern side of the Western Ghats at an elevation ranging from 45 to 1726 m above sea level (Fig. 1). Kaveri, which is the main river in Karnataka, originates in Talakaveri in Kodagu district. Kodagu is predominantly an agricultural region, producing rice and coffee, and various spices like pepper and cardamom and other agroforestry crops are cultivated within the region. The prevalent plantation crop in Kodagu district is coffee. It is the second-largest coffee production region in India after Chikmagalur district, and it accounts for about one-third of India's coffee production. Kodagu is rich in wildlife and has one national park and three wildlife sanctuaries despite being a small district. During the monsoon season (July–August), precipitation is intense and more or less continuous until the end of November. The average annual rainfall is about 4000 mm in hilly region. Heavy rainfall of about 1200 mm occurred during August 2018 (Fig. 2) and caused severe flooding and landslides in the region. The total rainfall that occurred in August exceeded the amount of the previous 4 years. The total damage caused by the August 2018 landslides was widespread and severe, and the total landslide area was two to three times larger than the landslide areas of the previous 4 years combined. The landslides occurring in our case study area are mainly debris flows. The study area is predominantly hill ranges covering dense forests, plantations, and cultivated valleys (Ramachandra et al. 2019). The study area is characterized by the highly dissected, undulating, and sloping structural hill ranges. Geologically, the study area comprises garnetiferous sericite schists and garnetiferous amphibolite, peninsular gneisses, and biotic gneisses with quartz (Vinutha 2015).

Data used and methodology

Datasets

Inventory dataset

In this study, a training dataset of polygons of the landslides for the Kodagu district was prepared from a manual delineation of landslides based on high-resolution PlanetScope imagery. The satellite images were taken from the Planet Labs Inc. PlanetScope, which includes more than 130 Dove satellites that provide 3 m spatial resolution multispectral images in four bands

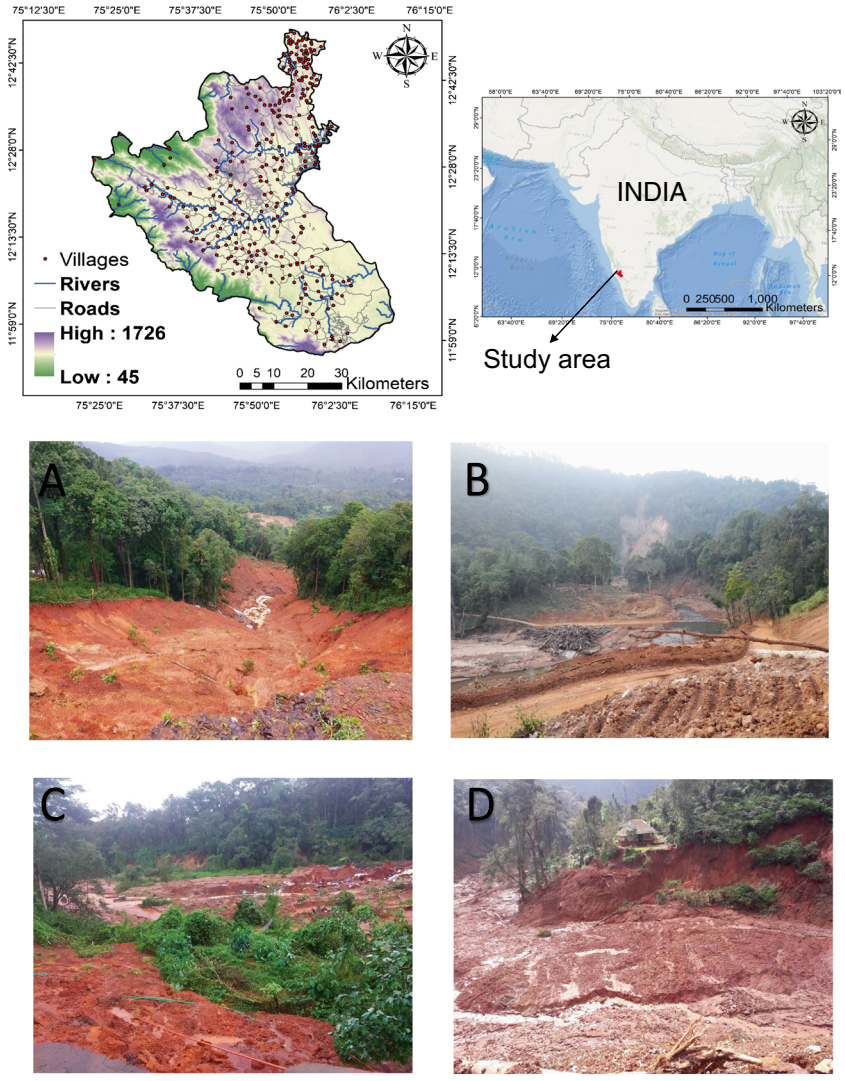
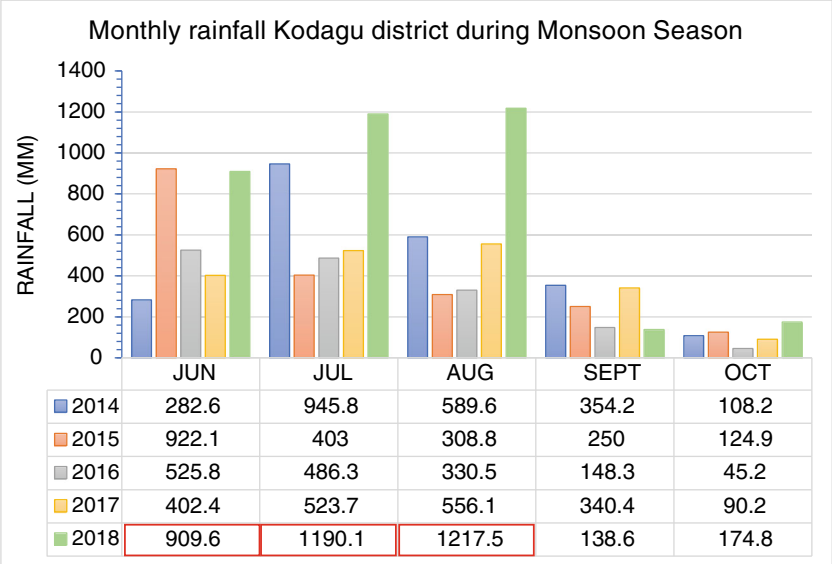


Fig. 1 The location of our study area, with villages, road, and river network. Photographs A–D show landslide features in the area

(RGB, NIR) red, 590–670 nm; green, 500–590 nm; blue, 455–515 nm; and NIR, 780–860 nm (Team 2018). We used cloud-free pre- and post-event PlanetScope images to map the landslide locations manually (Fig. 3). A total of 343 landslides were mapped as polygons, covering an area of 4140 km^2. The number of detected landslides differs from the 771 landslides reported by another study using object-based image analysis (OBIA) (Martha et al. 2019). This difference likely results from the automated landslide detection algorithm that was used. In our case, the individual parts of the same landslide were often counted as separate polygons, mainly when they were not connected due to the presence of shadows or vegetation in the images. The landslide inventory has different landslide types, namely, mud-flows, rock falls, and debris slides. The study area has hilly terrain and the landslide lengths vary, reaching up to 1828 m in length. The smallest manually mapped landslide is 276.23 m^2 and the largest is 81,342.87 m^2. Of the total 343 landslides, 93 are mudslides, 23 are rock falls, and 227 are debris-type landslides.

Optical data

In hilly terrain, dissected landscapes with rocks and barren areas show similar spectral characteristics as landslides (Moine et al. 2009). Fayne et al. (2019) observed that the red wavelength band provides spectral characteristics of landslides and barren areas in hilly terrain and forest-covered areas. The optical band of single RGB (red, green, blue) is useful for the identification of landslides, but it is not sufficient to differentiate landslides from vegetation growth in a shaded region. In such a case, an additional infrared band is useful to counteract the drawbacks of the mixed spectral response of landslides to only RGB spectral data. For the manual detection of landslides, we used the NDVI layer along with four bands of 3-m spatial resolution PlanetScope Dove imagery. The four PlanetScope spectral bands were used to calculate the normalized difference vegetation index (NDVI), which served as the basis for the landslide detection. The NDVI represents the surface reflectance, which provides an estimate of the vegetation growth or loss, which may affect landslide occurrence. The PlanetScope



	JUN	JUL	AUG	SEPT	OCT
2014	282.6	945.8	589.6	354.2	108.2
2015	922.1	403	308.8	250	124.9
2016	525.8	486.3	330.5	148.3	45.2
2017	402.4	523.7	556.1	340.4	90.2
2018	909.6	1190.1	1217.5	138.6	174.8

Fig. 2 Monthly rainfall (June–October) for the Kodagu districts in the Western Ghats of India (source: India Meteorological Department, Customized Rainfall Information System (CRIS)). The three red polygons show excess rainfall in the months of June, July, and August 2018

spectral bands were used to calculate the NDVI as the basis for the landslide modelling.

Slope

The selection of landslide-affecting factors depends on the local terrain conditions. We extracted the slope data from the 12.5-m resolution digital elevation model (DEM) that was created from the ALOS PALSAR data. The slope angle is crucial because the movement of mass is directly linked to the steepness of the slope, whereby steeper slopes are more prone to landslides. On the other hand, low angle slopes are more prone to the effects of channelized deposits, which results in rock fall and debris slides (Fan et al. 2018).

Convolution neural network

CNNs represent the state-of-the-art method in computer vision and image processing. Recently, CNNs have been applied in the domain of object detection and semantic segmentation due to the availability of labelled targeted images (Zhang et al. 2018). The use of CNNs is favourable for object detection and semantic segmentation because they have access to a large number of labelled images for training purposes, state-of-the-art algorithms, optimized CNN architectures, and GPUs (Guirado et al. 2017). Useful feature representations can be obtained by a CNN's multi-layer feed-forward neural networks, which allows the neural networks to recognize the feature differences in the image without using expert knowledge and defining rules (Ding et al. 2016). CNNs have a specific architecture, in which layers contain the pooling and convolutional layers, whereby the convolutional layers are considered to be the central part of a CNN architecture. The input image should be divided to the fixed window size patches for training the CNNs. The location of the centroid pixel of the window is selected based on the landslide bodies. Therefore, the fixed window size should be the minimum bounding box to cover the landslide in the image patches. These image patches are convolved by several trainable kernels and produce feature maps. Pooling layers are frequently used after the convolutional layers to subsample the resulting feature maps. Although there are various types of pooling strategies, the max pooling is the most widely used pooling method. Using max pooling, the CNN model can keep the maximum values from the results of each convolution layer.

The primary operations performed in any CNN can be summarized by the following equation (Zhang et al. 2018):

$$O^l = P\left(\sigma\left(O^{l-1} * W^l + b^l\right)\right)$$

where P refers to the pooling layer and the O^{l-1} is the result of the convolution layers of the lth layer, W^l and the b^l represent the weights and biases of the layer, respectively, and $\sigma()$ indicates the non-linearity function outside the convolutional layer.

In this study, an input window size of 32×32 pixels was used for landslide detection. According to our landslide inventory, we had several small landslides with different shapes. Some are elongated and thin and can almost look like an unpaved road rather than a landslide. Most landslides exhibit a mixture of topographic features, which makes them difficult to recognize. This input window size was selected as the optimum size based on a cross-validation for our case study area. To account for variability in the topographic factors and optical data, we structured a CNN model and a kernel size that varies from 5×5 to 3×3 for convolutional layers using max pooling layers with a 2×2 kernel size (Fig. 4). Our structured CNN model was prepared and trained in Trimble's eCognition software (eCognition Developer 2020). The statistical gradient descent function was used to optimize weightings through the network. Experimental results showed that using a batch size of 50 along with a learning rate of 0.0001, 3000 epochs resulted in the best detection results.

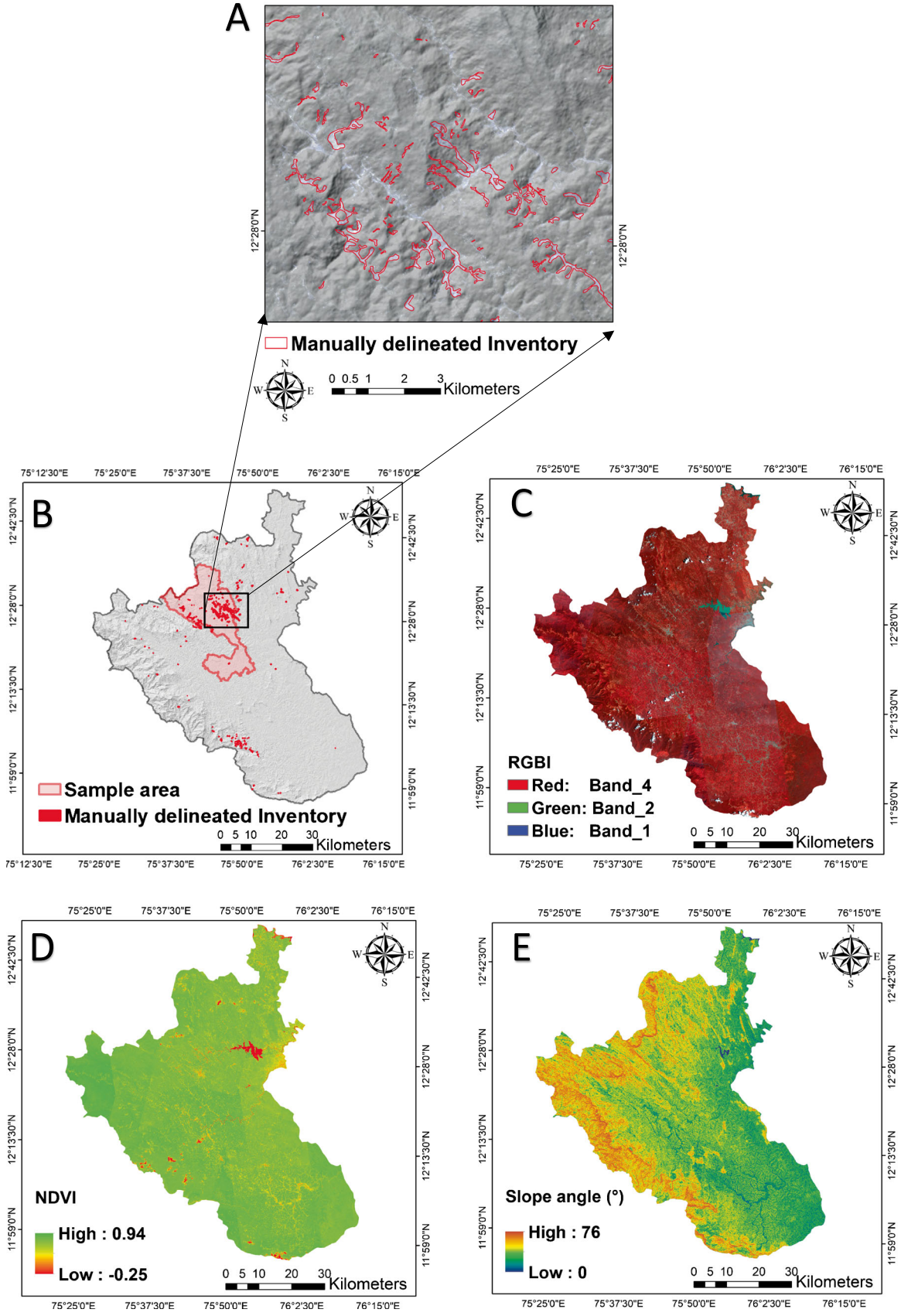


Fig. 3 (A) and (B) manually delineated landslide inventory prepared using visual image interpretation (C) false colour composite PlanetScope image, (D) normalized difference vegetation index (NDVI), and (E) slope angle

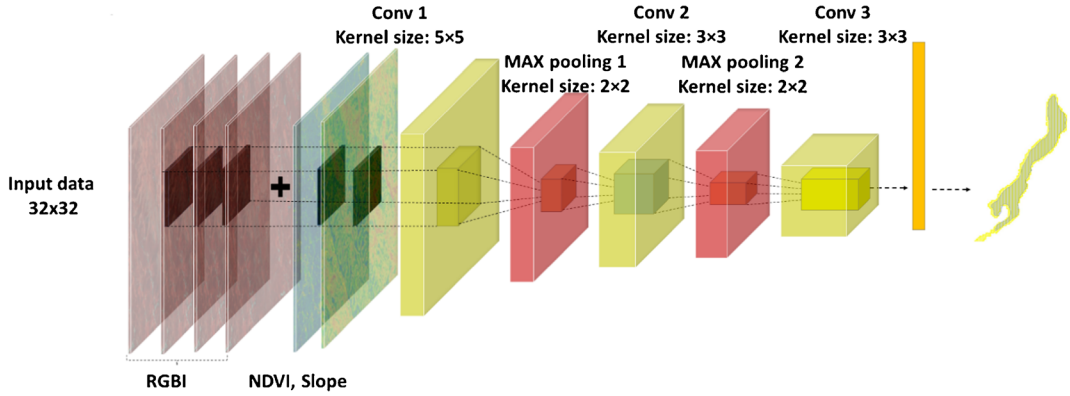


Fig. 4 The architecture of the CNN model, which is trained separately with two different training datasets

K-fold cross-validation

In this study, cross-validation was applied to determine the best model for landslide mapping and to decrease the negative effects of random sampling on the performances of the models. A fourfold cross-validation (CV) was applied based on various parameters such as the size of the database, different conditioning factors, and the number of computations within membership functions. The landslide-affected area was randomly divided into four equal folds of F1, F2, F3, and F4 where for any n and $m \in t$, size Fn = size Fm. The model runs k times and for any time of t, $t \leq k$. When the model runs at time t, 75% of the data without a subset of Ft was used for training the model, and 25% of the data was prepared for testing the model (Ghorbanzadeh et al. 2018). This method has been used by many researchers with various folds for different study goals. For example, Wiens et al. (2008) used a fivefold CV, and Ghorbanzadeh et al. (2019c) selected a fourfold CV for spatial prediction of wildfire susceptibility mapping. The distribution of our landslide inventory data within different four folds is shown in Fig. 5.

Landslide detection using CNNs

The architecture of the CNN model (Fig. 4) was trained with training datasets from outside the study site. Afterwards, the trained model was tested in the study site. We used the first CNN layer with a kernel size of five and continued with two CNN layers with a kernel size of three, adopted from Ghorbanzadeh et al. (2019a). The pooling layer was used to down sample the output of the CNN layer to produce a set of feature maps (Ghorbanzadeh et al. 2019a). The pooling layer reduces the spatial size of feature maps, thus reducing the computation volume for the remaining layers. In the CNN model, two max pooling layers of 2×2 were used. We fed our CNN model initially with a five-layer training dataset, including the optical data of the spectral bands RGBI and NDVI ($CNN_{RGBI,\ N}$), and then we added the topographic factor layer (slope steepness) to the previous dataset to train our $CNN_{RGBI,\ N,\ S}$ model. In the $CNN_{RGBI,\ N,\ S}$ model, we considered spectral bands along with the topographic factor (Fig. 6).

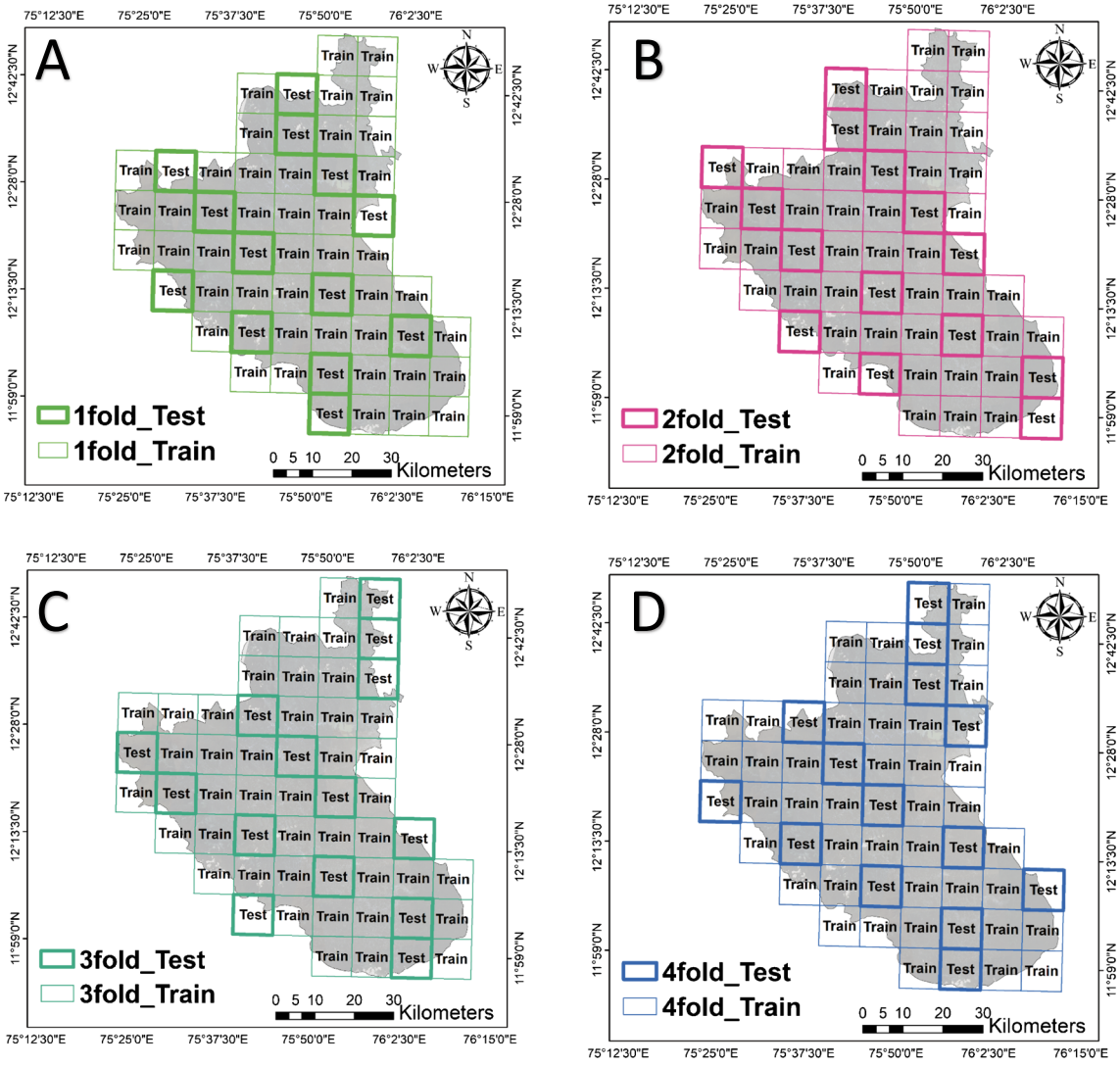
Comparison of landslide mapping using CNNs and manual detection

The trained CNN model was evaluated by employing to a sample area in Kodagu district. We used manually delineated landslide boundaries as ground truth, which were prepared using visual image interpretation of pre- and post-event PlanetScope imagery and landslide point data provided by the Geological Survey of India. We compared the manually delineated landslide boundaries with the landslide inventory data generated by training the CNN model separately, once with five layers with optical data and then with six layers with optical data and the slope layer.

The visually interpreted landslide dataset was separated into training and validation datasets because using training data enables the model to provide better predictions, and validation of the landslides improves the accuracy of the model. Choosing the right data spilt is important for the best results. Therefore, in the landslide dataset, a random 75/25 ratio was chosen for training/validation data. Increasing the proportion of validation data would mean a decrease in the model's prediction accuracy, therefore, a 4-fold cross-validation process is considered optimal. It consists of a random split of the dataset into four folds. Three out of the four folds are chosen to perform model training, while the last quarter is used for validation. The process is repeated by choosing another set of quarters for validation and the three others for training. This process is repeated three times until all four groups have been used for validation (Fig. 5). The four accuracy assessments obtained are averaged into one overall accuracy assessment. Validation is performed with the whole dataset, but a given sample is never used for training and validation at the same time. At each stage of the 4-fold process, 75% of the dataset is randomly selected as training data, while the rest is left for validation.

A number of accuracy assessment approaches were used to assess the performance of the applied CNN model by evaluating the consistency between the CNN and manually mapped landslide inventory (Ghorbanzadeh et al. 2019b). In this study, the performance of the CNN model was ascertained using four different metrics (precision, recall, F1 score, and the Matthews correlation coefficient (MCC)), which are based on confusion matrices with true positives (TP), false positives (FP), and false negatives (FN) (Figs. 7 and 8).

The precision is the proportion of CNN-derived landslide pixels correctly identified as landslides (Lormand et al. 2018). The recall is the proportion of visually mapped landslide pixels



	Fold 1	Fold 2	Fold 3	Fold 4
K = 1	train	train	train	test
K = 2	train	train	test	train
K = 3	train	test	train	train
K = 4	test	train	train	train

Fig. 5 (A, B, C, D) The applied fourfold cross-validation (CV) for the inventory dataset. Each colour represents a specific fold of the inventory dataset of the image patches both in the maps and in the table. The patches with bold text were used for testing, and the others were used for training the model

that were correctly detected by the CNNs (Liu et al. 2020). The F1 score is defined as the weighted harmonic mean of the precision and recall, used to evaluate the performance of the model (Liu et al. 2020). The higher the value of the F1 score, the better the performance of the model (Sameen and Pradhan 2019). The Matthews correlation coefficient (MCC) is useful to compare the binary classification of imbalanced datasets, and its values range from −1 to 1, where 1 represents a perfect classifier and 0 represents a classifier with random detection (Prakash et al. 2020) (Tables 2 and 3).

Analysis of landslide mapping using frequency area distribution

Landslide inventories are statistically analysed using frequency area distribution (FAD) curves, in which landslide areas are plotted against the cumulative landslide frequencies. In a study by Malamud et al. (2004a), observations show that the power law

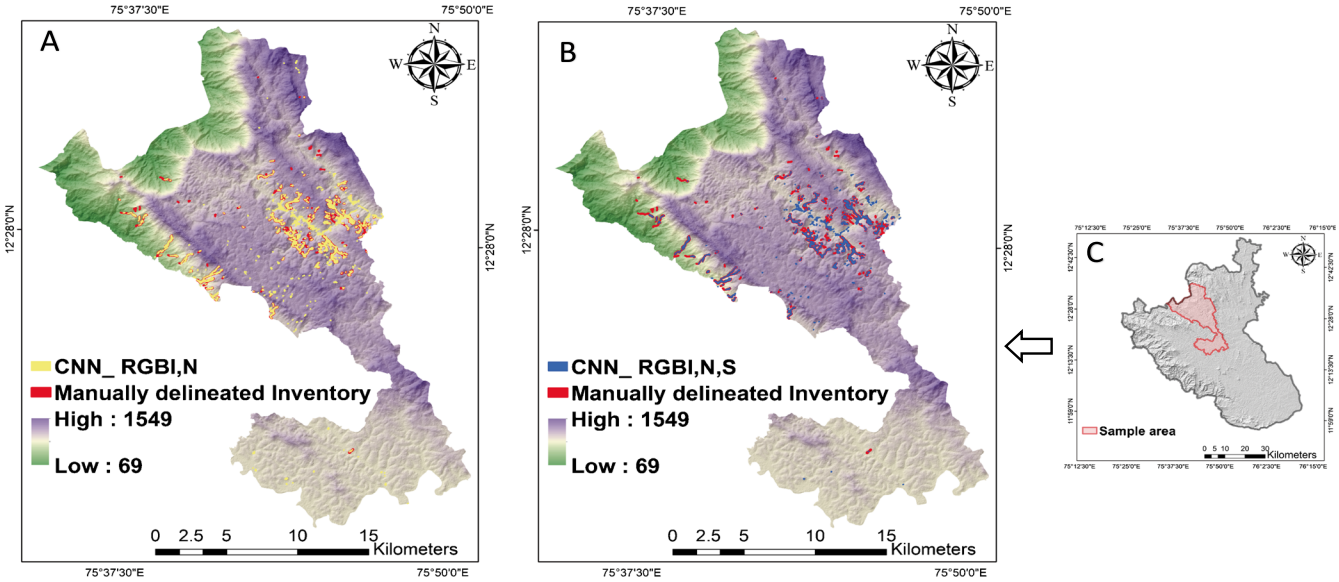


Fig. 6 Landslide detection results (A) CNN_RBGI, N and (B) CNN_RBGI, N, S; CNN_RBGI, N: convolutional neural network with RGB, infrared bands, and NDVI layer; CNN_RBGI, N, S: convolutional neural network with RGB, infrared bands, NDVI, and slope layer

applies for medium and large landslides. The probability of occurrence of a landslide of a particular size can be given by the power law equation:

$$p(x) = cX^{-\beta}$$

where X are observed values, c is a normalization constant, and β is the power law exponent.

Figure 9 shows the power law distribution for medium to large landslides and the divergence from the power law towards lower frequencies with a rollover point where the frequency decreases for smaller landslides. The trend of the FAD of most landslide inventories diverges from a power law for small landslides (Guzzetti et al. 2002; Malamud et al. 2004a; Stark and Guzzetti 2009; Tanyaş et al. 2019). The point where this divergence begins is defined as the cut-off point (Stark and Hovius 2001; Tanyaş et al. 2019). According to Van Den Eeckhaut et al. (2007), in a power law distribution, the slope of the distribution is defined by a power law exponent. The part that is represented by large events is referred to as the power law tail, as shown in Fig. 9 (with a scaling parameter, β). Malamud et al. (2004a) investigated four well-documented landslide events and concluded that rollover is a real phenomenon in landslide inventories that depends upon the bias and under-sampling of the smaller landslides.

They modelled the FAD for these four inventories and established theoretical curves to estimate the total landslide area triggered by an earthquake or rainfall event.

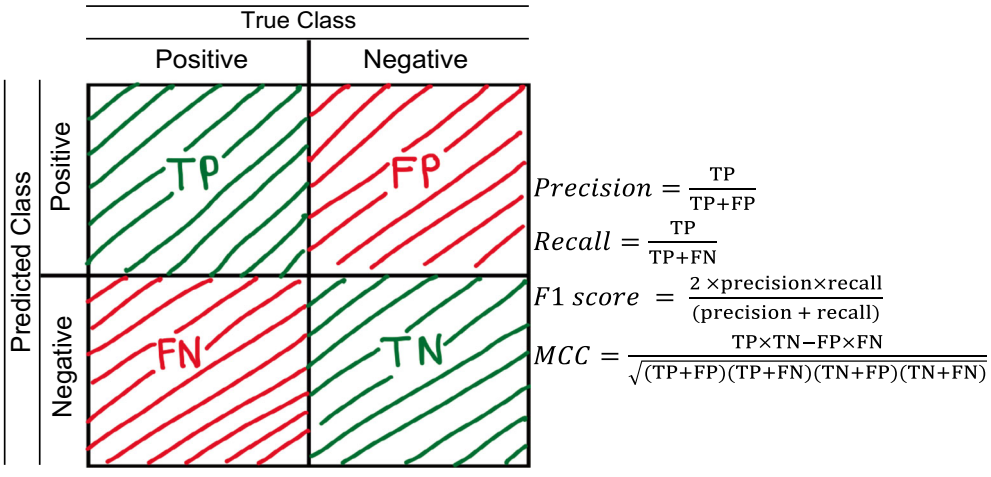


Fig. 7 Confusion matrix showing true class and predicted classes of landslides and other features and the four different metrics

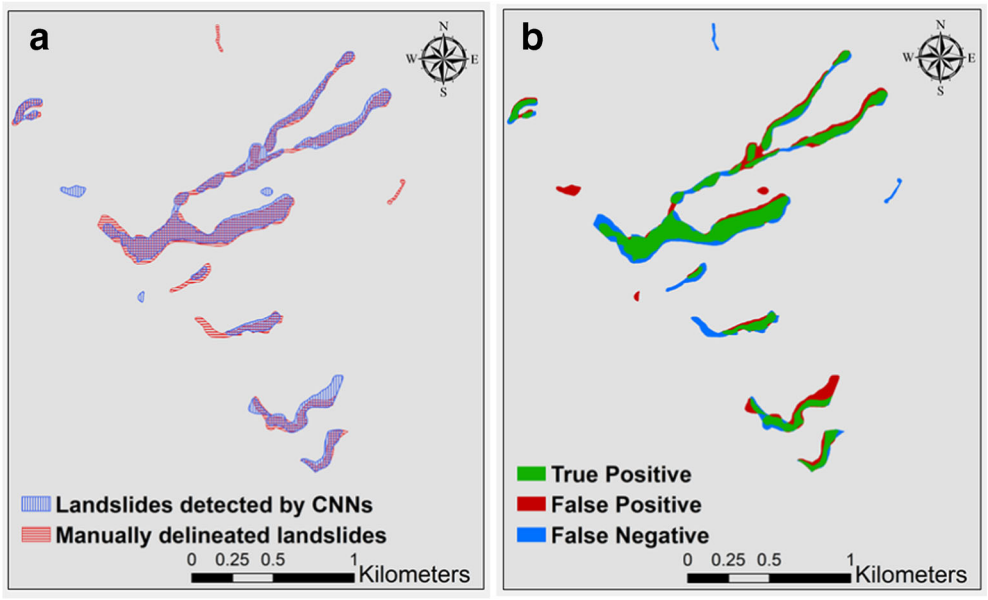


Fig. 8 (**a**) Inventory of landslide areas obtained from CNNs and manual delineation (**b**) true positive (TP), false positive (FP), and false-negative (FN) areas identified by comparing spatial overlaps between polygons of (**a**)

Malamud et al. (2004b) showed that the entire FAD of landslides could be explained by a three-parameter inverse gamma distribution (equation). This approach also described a way to estimate the landslide event magnitude (*mLS*). The *mLS* is the indication of the size of the landslide triggering event and gives an indication of the severity of the event in terms of landslide occurrence in a particular area for an event:

$$p(A_L;\rho,a,s) = \frac{1}{a\Gamma(\rho)}\left[\frac{a}{A_L - s}\right]^{\rho+1} exp\left[-\frac{a}{A_L - s}\right]$$

where ρ is the parameter primarily controlling power law decay for medium and large values, $\Gamma(\rho)$ is the gamma function of ρ, A_L is landslide area, a is the location of rollover point, s is the exponential decay for small landslide areas, and $-(\rho+1)$ is the power law exponent. Malamud et al. (2004b) provided a best fit for the power law exponent and showed that $-(\rho+1) = 2.4$.

Table 4 shows that the power law exponent of the analysed folds for CNN-derived inventories ranges from 1.37 to 2.22, which is lower than 'the given power law function exponent of 2.4' reported by Malamud et al. (2004b). Lower power law exponent values are lower as a result of using smaller dataset for analysis, as Malamud et al. (2004b) used three large landslide inventories from around the world. The smallest landslide areas mapped ranged from 2491 to 9407 m^2, and the largest landslides mapped range from 47,695.36 to 528,042.68 m^2 for the CNN-derived landslide inventories (see Table 4).

There is a scattered pattern of the plotted landslide probability density to the inverse gamma fit (see Fig. 10). Differences in the probability distribution and inverse gamma fit could result from gaps in the data of mapped landslides for given inventories, which means that some smaller landslides are missing or not mapped by the CNN model. The rollover points differ between inventories. For manual and CNN-derived inventories, the rollover points for smaller landslides vary between 454.84 and 10,125.55 m^2. In the CNN model-derived inventories, the rollover point ranges between 10,125.55 and 17,345.80 m^2, which is larger than manually delineated landslides because our model was not able to detect smaller landslides efficiently because of constraints in training

Table 2 The results of landslide detection in the study area based on the CNN model trained CNN_RBGI, N: convolutional neural network with RGB, infrared bands, and NDVI layer

Fold	TP (ha)	FP (ha)	FN (ha)	Precision (%)	Recall (%)	F1measure (%)	MCC (%)
1-fold	15.70	9.18	7.79	63.1	66.8	64.9	64.9
2-fold	135.10	70.87	88.00	65.6	60.6	63.0	62.9
3-fold	139.30	64.02	59.47	68.5	70.1	69.3	69.2
4-fold	72.24	41.15	42.16	63.7	63.1	63.4	63.4
Mean				65.2	65.1	65.1	65.1

Accuracies are stated as precision, recall, F1-measure, and MCC

Table 3 The results of landslide detection in the study area based on the CNN model trained CNN_RBGI, N, S: convolutional neural network with RGB, infrared bands, NDVI, and slope layers

Fold	TP (ha)	FP (ha)	FN (ha)	Precision (%)	Recall (%)	F1measure (%)	MCC (%)
1-fold	14.65	16.83	3.83	46.5	79.3	58.6	60.7
2-fold	150.26	95.18	62.39	63.8	73.8	68.4	68.5
3-fold	150.11	73.97	41.50	70.6	79.6	74.9	75.0
4-fold	86.88	54.15	24.81	63.8	78.6	70.4	70.8
Mean				61	78	68	68.8

Accuracies are stated as precision, recall, F1-score, and MCC

samples due to the smaller size of the study area. For smaller landslides, fold 2 shows more effectiveness, and for larger landslides, fold 3 shows more effectiveness as can be seen in power law tail in Fig. 10.

Discussion

This paper presents an approach to mapping landslides using a CNN in the hilly terrain regions of the Western Ghats in India. We used a simple CNN architecture with five and six input layers to train the model. We designed the CNN architecture using minimal input data for landslide detection in the study area. In recent studies, CNNs outperformed traditional machine learning algorithms in the detection of landslides (Liu et al. 2020; Ye et al. 2019). However, designing a CNN architecture and optimizing its parameters using sample strategies remain challenging tasks (Ghorbanzadeh et al. 2019a). We only used slope data as auxiliary topographical input information to remove the errors caused by spectral similarities in riverbeds and the built-up area in moderate slopes. The CNN model was trained with data, including the slope layer, which performed better than when using optical data alone by about 2.9 F1 score and 3.7 MCC mean values (Mondini et al. 2013). This higher accuracy is due to fewer FPs (almost half), which is attributed to the fact that the CNN model was trained with the slope data. The resulting accuracies of our designed CNN architecture are comparable with published studies that were based on much more complex CNN architectures, such as the U-Net and residential networks (Prakash et al. 2020).

The accuracy assessment metrics used for the validation of the landslide detection in this study demonstrate that using the CNN

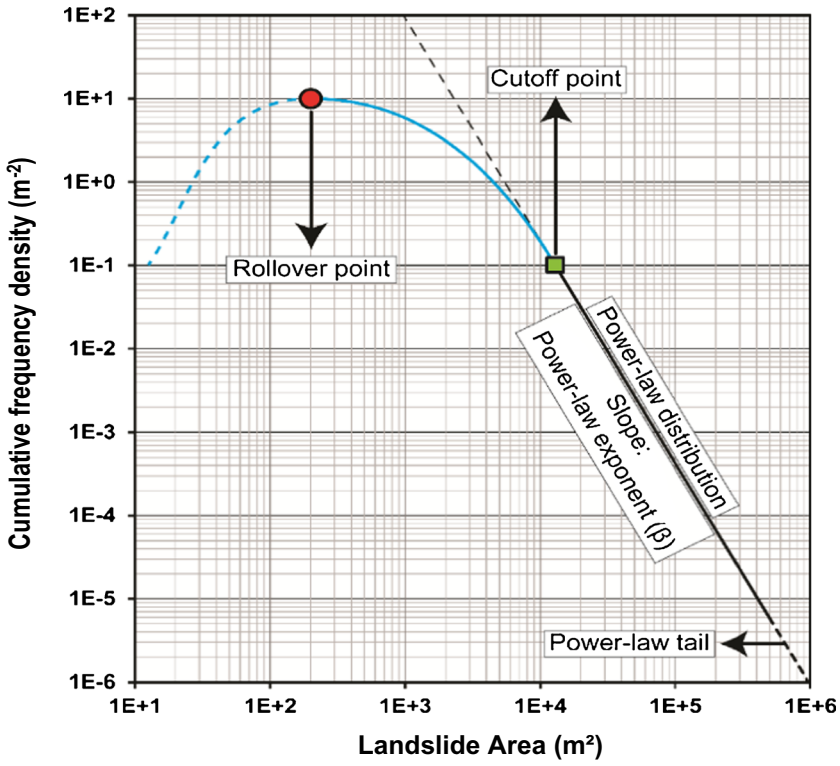


Fig. 9 Schematic representation of the main components of a cumulative frequency area distribution for (FAD) for a landslide inventory (Malamud et al. 2004a)

Table 4 Comparison of the frequency area statistics of landslide area

Inventories	Total number of landslides N_{LT}	Total area of landslides A_L km^2	Minimum area of landslides minA_L m^2	Maximum area of landslides maxA_L m^2	Power law exponent (β)	Rollover point (m^2)
Manual 1-fold	18	0.23	1524	39,370.03	2.02	5744.82
Manual 2-fold	167	2.03	176	446,831.61	1.37	454.84
Manual 3-fold	86	4.02	621	203,949.08	1.38	1287.01
Manual 4-fold	53	1.13	1610	146,899.24	1.66	5107.21
CNN 1-fold	9	0.24	2491	47,695.36	2.22	10,125.55
CNN 2-fold	46	2.05	5469	528,042.68	1.66	17,345.80
CNN 3-fold	40	1.90	5629	234,587.59	1.75	14,021.63
CNN 4-fold	35	1.04	9407	141,198.62	2.21	27,447

model provides automatic rainfall-induced landslide detection and inventory mapping using remote sensing data. However, it is not appropriate comparing our results with those from other studies in the Western Ghats, which used other object-based models for landslide detection and modelling (Martha et al. 2019). In this study, our model was trained with optical data from PlanetScope imagery with 3-m spatial resolution, whereas Martha et al. (2019) used Resourcesat-2,2A LISS-IV multispectral data with 5.8-m spatial resolution. The use of different datasets seems to produce differences in the detected landslides, especially because only the unvegetated areas were mapped and the connections were often not detected because of shadows and vegetation (Fiorucci et al. 2019). The model training strategy differs between CNN and other machine learning and object-based models. Adding additional heterogeneous training data usually reduces the convergence capabilities of a CNN model and, consequently, causes generalization and reduces the overall accuracy. However, by adding the slope data for training, the applied CNN model was able to decrease the false positives by distinguishing the landslide areas from non-vegetated areas such as the riverbeds, bare land, and built-up areas (Fig. 11). In our study, a total of 343 landslides were mapped as polygons, covering an area of 4140 km^2. The landslides along the riverbeds were difficult to detect using our CNN architecture. Also, water body and settlements areas were considered false positives in the model. There are some errors in the CNN slide detection results in built-up areas, forests, along the road network, and in riverbeds. A total of 14 polygons were false positives, whereby 6 were in forests, 6 in riverbeds, and one false positive each in built-up areas and along the road network. The false positives in forests cover an area of 11,937.6 m^2, which is about 2.04% of the CNN result. Similarly, the false positives in the built-up areas and road networks make up about 0.19% and 0.22% of the area of the CNN results, respectively. Most false positives were in riverbeds, which make up about 4.5% (2,656,085 m^2) of the CNN result. False positives in our results could be due to having fewer training samples in those classes because the size of our study area is smaller.

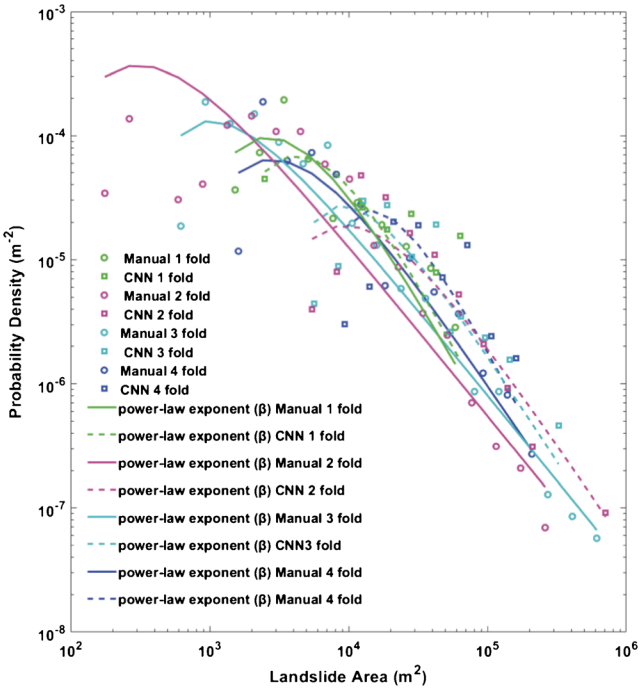


Fig. 10 Landslide frequency size distribution, representing the dependence of landslide probability density on the landslide area

The vegetation plays a huge role in detection of landslides as the model can distinguish the boundaries of landslides apart from vegetated areas using NDVI layer. Another limitation of our model was that it merged several individual landslides into one landslide. The fact that a number of landslides were mapped is attributed to the amalgamation of landslides due to merging of debris flows and slides after the event. The amalgamation of landslides is an issue that can be overcome in future studies with novel and optimized CNN architectures.

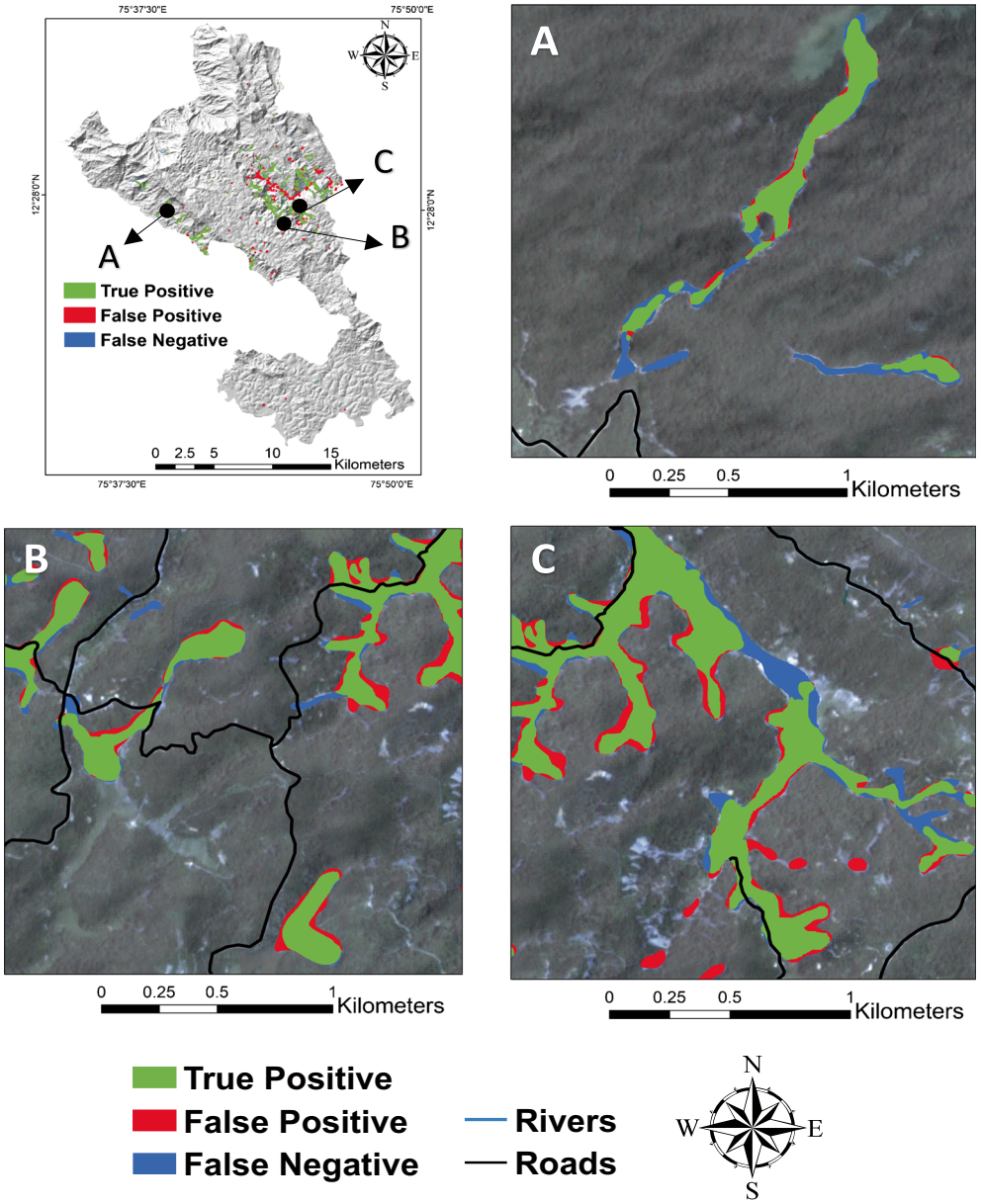


Fig. 11 Enlarged maps of the results of the CNN model, showing the detection of true and false positives and false negatives in the testing area: (A) forest area, (B) near-road network and hilly terrain area, and (C) near built-up areas

Conclusions

The main contribution of the present study is the automatic detection of landslides after major triggering events such as monsoon rainfalls. However, the main limitation for such study areas is having access to cloud-free images during the rainfall season. Developing semi- or fully automated approaches for landslide mapping is needed due to the substantially increasing frequency of natural disasters in recent years, causing significant concerns about the loss of human lives. In the selected study area, the loss of properties was mainly due to rainfall-induced landslides.

The occurrences of landslides worldwide are expected to increase due to urbanization, deforestation, and continued anthropogenic activities. Climate change has also contributed to variations or fluctuations in precipitation in the landslide-prone areas. Our study represents the semi-automatic rainfall-induced landslide detection and inventory mapping using remote sensing data for Kodagu district, which lies in the Western Ghats of India. We developed a CNN model to detect landslides based on various input data, including spectral information and the topographical slope factor. Our applied CNN model structure is simple and may not be superior to those used in previous studies listed in Table 1. However, our approach requires less human participation and can thus be considered a semi-automatic approach. The applied methodology is easily transferable to similar regions, like the Himalayas, and our trained model can also be used for a new landslide inventory dataset. However, in regions with less vegetation cover or steeper terrain, the model might require retraining based on the landslide inventory from these areas, which will

enhance the performance of the model. Moreover, the methodology can be used for detecting landslides caused by other triggering processes, e.g., earthquake-induced landslides. Therefore, this study and the applied methodology are useful for landslide inventory mapping and, consequently, for disaster mitigation management.

Acknowledgements

The authors are also grateful to the anonymous reviewers and the editor for their valuable comments/suggestions which have helped us to improve an earlier version of the manuscript.

Funding

Open Access funding provided by Paris Lodron University of Salzburg. This research was partly funded by the Austrian Science Fund (FWF) through the Doctoral College GIScience (DK W1237-N23) at the University of Salzburg—Open Access Funding by the Austrian Science Fund (FWF).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Chen Z, Zhang Y, Ouyang C, Zhang F, Ma J (2018) Automated landslides detection for mountain cities using multi-temporal remote sensing imagery Sensors 18:821

Cruden DM, Varnes DJ (1996) Landslides: investigation and mitigation

Das I, Stein A, Kerle N, Dadhwal VK (2011) Probabilistic landslide hazard assessment using homogeneous susceptible units (HSU) along a national highway corridor in the northern Himalayas. India Landslides 8:293–308. https://doi.org/10.1007/s10346-011-0257-9

Ding A, Zhang Q, Zhou X, Dai B (2016) Automatic recognition of landslide based on CNN and texture change detection. In: 2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC), IEEE, pp 444–448

Du Z, Yang J, Ou C, Zhang T (2019) Smallholder crop area mapped with a semantic segmentation deep learning method Remote Sensing 11:888

Duro DC, Franklin SE, Dubé MG (2012) A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery. Remote Sens Environ 118:259–272. https://doi.org/10.1016/j.rse.2011.11.020

eCognition Developer T (2020) "eCognition Developer User Guide"

Eskandari S, Reza Jaafari M, Oliva P, Ghorbanzadeh O, Blaschke T (2020) Mapping land cover and tree canopy cover in Zagros forests of Iran: application of Sentinel-2, Google Earth, and Field Data Remote Sensing 12:1912

Fan X, Domènech G, Scaringi G, Huang R, Xu Q, Hales TC, Dai L, Yang Q, Francis O (2018) Spatio-temporal evolution of mass wasting after the 2008 M w 7.9 Wenchuan earthquake revealed by a detailed multi-temporal inventory. Landslides 15:2325–2341

Fayne JV, Ahamed A, Roberts-Pierel J, Rumsey AC, Kirschbaum DJEI (2019) Automated satellite-based landslide identification product for Nepal 23:1–21

Fiorucci F, Ardizzone F, Mondini AC, Viero A, Guzzetti F (2019) Visual interpretation of stereoscopic NDVI satellite images to map rainfall-induced landslides. Landslides 16:165–174. https://doi.org/10.1007/s10346-018-1069-y

Ghorbanzadeh O, Blaschke T, Gholamnia K, Meena SR, Tiede D, Aryal J (2019a) Evaluation of different machine learning methods and deep-learning convolutional neural networks for landslide detection Remote Sensing 11:196

Ghorbanzadeh O, Meena SR, Blaschke T, Aryal J (2019b) UAV-based slope failure detection using deep-learning convolutional neural networks Remote Sensing 11:2046

Ghorbanzadeh O, Rostamzadeh H, Blaschke T, Gholaminia K, Aryal J (2018) A new GIS-based data mining technique using an adaptive neuro-fuzzy inference system (ANFIS) and k-fold cross-validation approach for land subsidence susceptibility mapping. Nat Hazards 94:497–517. https://doi.org/10.1007/s11069-018-3449-y

Ghorbanzadeh O, Valizadeh Kamran K, Blaschke T, Aryal J, Naboureh A, Einali J, Bian J (2019c) Spatial prediction of wildfire susceptibility using field survey GPS data and machine learning approaches 2:43

Guirado E, Tabik S, Alcaraz-Segura D, Cabello J, Herrera F (2017) Deep-learning convolutional neural networks for scattered shrub detection with Google Earth imagery arXiv preprint arXiv:170600917

Guzzetti F, Carrara A, Cardinali M, Reichenbach P (1999) Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study. Central Italy Geomorphology 31:181–216. https://doi.org/10.1016/S0169-555X(99)00078-1

Guzzetti F, Malamud BD, Turcotte DL, Reichenbach P (2002) Power-law correlations of landslide areas in central Italy. Earth Planet Sci Lett 195:169–183. https://doi.org/10.1016/S0012-821X(01)00589-1

Guzzetti F, Mondini AC, Cardinali M, Fiorucci F, Santangelo M, Chang K-T (2012) Landslide inventory maps: new tools for an old problem. Earth Sci Rev 112:42–66. https://doi.org/10.1016/j.earscirev.2012.02.001

Jin B, Ye P, Zhang X, Song W, Li S (2019) Object-oriented method combined with deep convolutional neural networks for land-use-type classification of remote sensing images. J Indian Soc Remote Sensing 47:951–965. https://doi.org/10.1007/s12524-019-00945-3

Juang CS, Stanley TA, Kirschbaum DB (2019) Using citizen science to expand the global map of landslides: introducing the cooperative open online landslide repository (COOLR). PLoS One 14:e0218657

Lei T, Zhang Y, Lv Z, Li S, Liu S, Nandi AK (2019) Landslide inventory mapping from bitemporal images using deep convolutional neural networks IEEE Geoscience and Remote Sensing Letters 16:982–986. doi:https://doi.org/10.1109/LGRS.2018.2889307

Liu P, Wei Y, Wang Q, Chen Y, Xie J (2020) Research on post-earthquake landslide extraction algorithm based on improved U-Net model. Remote Sens 12:894

Liu S, Qi Z, Li X, Yeh AG-O (2019) Integration of convolutional neural networks and object-based post-classification refinement for land use and land cover mapping with optical and SAR data. Remote Sens 11:690

Lormand C, Zellmer GF, Németh K, Kilgour G, Mead S, Palmer AS, Sakamoto N, Yurimoto H, Moebis A (2018) Weka trainable segmentation plugin in ImageJ: a semi-automatic tool applied to crystal size distributions of microlites in volcanic rocks. Microsc Microanal 24:667–675. https://doi.org/10.1017/S1431927618015428

Lu P, Stumpf A, Kerle N, Casagli N (2011) Object-oriented change detection for landslide rapid mapping. IEEE Geosci Remote Sens Lett 8:701–705. https://doi.org/10.1109/LGRS.2010.2101045

Malamud BD, Turcotte DL, Guzzetti F, Reichenbach P (2004a) Landslide inventories and their statistical properties. Earth Surf Process Landf 29:687–711

Malamud BD, Turcotte DL, Guzzetti F, Reichenbach P (2004b) Landslides, earthquakes, and erosion. Earth Planet Sci Lett 229:45–59

Martha TR, Kerle N, Jetten V, van Westen CJ, Kumar KV (2010) Characterising spectral, spatial and morphometric properties of landslides for semi-automatic detection using object-oriented methods. Geomorphology 116:24–36. https://doi.org/10.1016/j.geomorph.2009.10.004

Martha TR, Roy P, Khanna K, Mrinalni K, Kumar KV (2019) Landslides mapped using satellite data in the Western Ghats of India after excess rainfall during August 2018. Curr Sci 117:804–812

Meena SR, Ghorbanzadeh O, Hölbling D (2019) Comparison of event-based landslide inventories: a case study from Gorkha earthquake 2015, Nepal. Paper presented at the European Space Agency's 2019 Living Planet Symposium, Milan,Italy,

Moine M, Puissant A, Malet J-P Detection of landslides from aerial and satellite images with a semi-automatic method. Application to the Barcelonnette basin (Alpes-de-Hautes-Provence, France). In, 2009

Mondini AC, Marchesini I, Rossi M, Chang K-T, Pasquariello G, Guzzetti F (2013) Bayesian framework for mapping and classifying shallow landslides exploiting remote sensing

and topographic data. Geomorphology 201:135–147. https://doi.org/10.1016/j.geomorph.2013.06.015
Pradhan B, Singh R, Buchroithner M (2006) Estimation of stress and its use in evaluation of landslide prone regions using remote sensing data Advances in Space Research 37:698–709
Prakash N, Manconi A, Loew S (2020) Mapping landslides on EO data: performance of deep learning models vs. traditional machine learning models Remote Sensing 12:346
Qayyum A, Malik A, Saad MN, Mazher M (2019) Designing deep CNN models based on sparse coding for aerial imagery: a deep-features reduction approach. Eur J Remote Sen 52:221–239
Ramachandra T, Bharath S, Vinay SJPiDS (2019) Visualisation of impacts due to the proposed developmental projects in the ecologically fragile regions-Kodagu district, Karnataka 3:100038
Sameen MI, Pradhan B (2019) Landslide detection using residual networks and the fusion of spectral and topographic information. IEEE Access 7:114363–114373
Shahabi H, Jarihani B, Tavakkoli Piralilou S, Chittleborough D, Avand M, Ghorbanzadeh O (2019) A semi-automated object-based gully networks detection using different machine learning models: a case study of Bowen catchment, Queensland, Australia Sensors 19:4893
Stark CP, Guzzetti F (2009) Landslide rupture and the probability distribution of mobilized debris volumes. J Geophys Res Earth Surf 114:1–16. https://doi.org/10.1029/2008JF001008
Stark CP, Hovius N (2001) The characterization of landslide size distributions. Geophys Res Lett 28:1091–1094. https://doi.org/10.1029/2000GL008527
Tanyaş H, van Westen CJ, Allstadt KE, Jibson RW (2019) Factors controlling landslide frequency–area distributions. Earth Surf Process Landf 44:900–917. https://doi.org/10.1002/esp.4543
Tavakkoli Piralilou S et al. (2019) Landslide detection using multi-scale image segmentation and different machine learning models in the higher Himalayas Remote Sensing 11:2575
Van Den Eeckhaut M, Poesen J, Govers G, Verstraeten G, Demoulin A (2007) Characteristics of the size distribution of recent and historical landslides in a populated hilly region. Earth Planet Sci Lett 256:588–603. https://doi.org/10.1016/j.epsl.2007.01.040
Vinutha D (2015) Geomorphology and natural hazards in parts of Coorg district Karnataka state
Wiens TS, Dale BC, Boyce MS, Kershaw GP (2008) Three way k-fold cross-validation of resource selection functions Ecological Modelling 212:244–255. https://doi.org/10.1016/j.ecolmodel.2007.10.005
Xu C, Tian Y, Zhou B, Ran H, Lyu G (2017) Landslide damage along Araniko highway and Pasang Lhamu highway and regional assessment of landslide hazard related to the Gorkha, Nepal earthquake of 25 April 2015. Geoenvironmental Disasters 4:14–14. https://doi.org/10.1186/s40677-017-0078-9
Ye C et al. (2019) Landslide detection of hyperspectral remote sensing data based on deep learning with constrains. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing:1–14. doi:https://doi.org/10.1109/JSTARS.2019.2951725
Zhang C, Sargent I, Pan X, Li H, Gardiner A, Hare J, Atkinson PM (2018) An object-based convolutional neural network (OCNN) for urban land use classification. Remote Sens Environ 216:57–70. https://doi.org/10.1016/j.rse.2018.06.034

S. R. Meena (✉) · O. Ghorbanzadeh · T. G. Nachappa · T. Blaschke
Department of Geoinformatics—Z_GIS,
University of Salzburg,
5020, Salzburg, Austria
Email: sansarraj.meena@sbg.ac.at

S. R. Meena · C. J. van Westen
Faculty of Geoinformation Science and Earth Observation (ITC),
University of Twente,
Enschede, The Netherlands

R. P. Singh
School of Life and Environmental Sciences, Schmid College of Science and Technology,
Chapman University One University Drive,
Orange, CA, USA

R. Sarkar
Department of Civil Engineering,
Delhi Technological University,
Bawana Road, Delhi, India

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/348446732

Recent Development in Green Fuel Vehicles and their Future Advancements

Article · October 2020

CITATIONS
0

READS
3

2 authors:



Amit Sachdeva
Bharati Vidyapeeth Deemed University
1 PUBLICATION **0** CITATIONS
SEE PROFILE



Dadaso D Mohite
Bharati Vidyapeeth Deemed University
9 PUBLICATIONS **4** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project
Wheel track interaction, Lateral stability, Curving, Car body parameters, Indian railways, Suspension system. View project

Project
CNC cylindrical grinding, Dressing, Taguchi Method, ANOVA, Modeling, Optimization View project

All content following this page was uploaded by Dadaso D Mohite on 14 January 2021.

The user has requested enhancement of the downloaded file.

International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 10 | Oct 2020 www.irjet.net p-ISSN: 2395-0072

Recent Development in Green Fuel Vehicles and their Future Advancements

Amit Sachdeva[1], Dadaso D Mohite[2]

[1]*Research Scholar, Department of Mechanical Engineering, BV(DU), COE, Pune (INDIA)*
[2]*Assistant Professor, Department of Mechanical Engineering, BV(DU), COE, Pune (INDIA)*

Abstract - *In recent years there has been a major shift in the Automobile sector from the conventional internal combustion engine to more refined and environmentally more sustainable technologies to run our vehicles. Through this study, a comprehensive summary of past, recent and future development in the department of an environmentally clean transportation system is concluded. The study mainly focuses on green technologies that will be used to power our vehicles in the near future. A technological roadmap of development of different configuration of EV's including HEV, BEV and FCEV in the past year is concluded. The study of HEV is presented with details of generally used powertrain structure (series, parallel, combination and plug-in-HEV) along with details on BEV and major subsystems used in running pure electric vehicles such as types of power sources used, motors and power converters. Through this paper working of electric vehicle is discussed along with presentation on the different energy flow diagrams and finally, an overview of hydrogen-powered vehicles is presented with details of Fuel cell and future technologies. Finally, to conclude our study a discussion on global trends and various trades-offs that have to be made with electric vehicles for future trends and potentials areas of research for future use in EV is discussed.*

Key Words: Electric Vehicles [EV]; Hybrid Electric vehicles [HEV]; Fuel Cell Vehicles [FCV]; internal Combustion Engines [ICE]; ultra-capacitors; Battery Management System; Lithium-ion Battery; Battery Module; Supercapacitors; Energy Storage System [ESS]; Induction Motors [IM]; Permanent Magnet Motor [PMS]; Converters

1. INTRODUCTION

Increase in pollution and severe harm to the mother Earth has led us to look for alternative sources of energies that will help us to control and reduce harmful emissions from our Automobiles and other industries. In 2016, over 194 countries participated in Paris climate agreement enforcement and pledged to limit the average global temperature increase to less than 2-degree Celsius in this century [1]. According to the reports, more than 50% of the world's Green House Gases (GHGs) emissions comes from USA, China and India.

The transport sector astonishingly accounts for more than 80% of pollutants released from major cities in countries like USA, CHINA and INDIA etc. In our country usually the quantity of air pollutants, particulate matter, sulfur dioxide, nitrogen oxides (NO), ozone (O_3) and carbon monoxide (CO) are often above the National Ambient Air Quality Standards (NAAQS) [2]. During the 2014 assessment, World Health Organization (WHO) listed New Delhi among the top ten cities around the world having the worst PM!) pollution. Vehicular emission from ICE has been the major contributors to the rise in pollution and can constitute the following emission sources as shown in figure 1 [2].

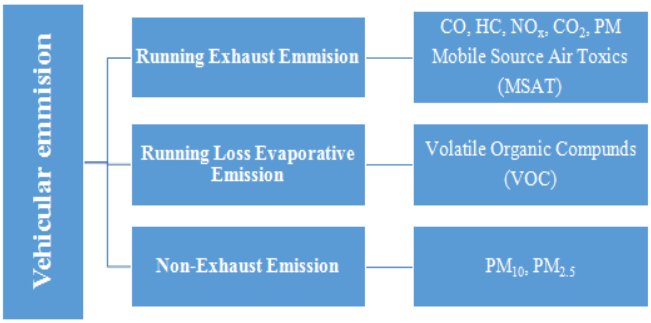


Figure 1: Types of vehicular emissions [2]

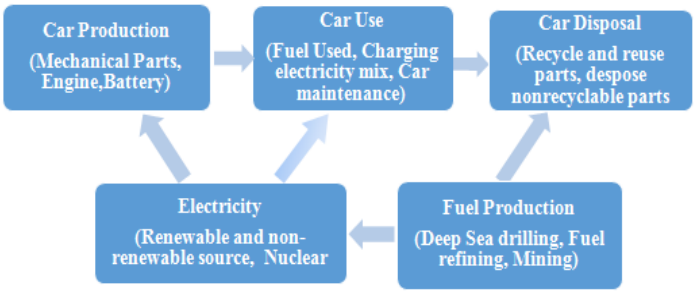


Figure 2: Sources of emission in life cycle of a car

The USA and China has already started wide spread adaptation of e-vehicles over fossil fuel powered vehicles to tackle the rising Green House Emissions in their country and as of 2019 they have highest stock of EV in the world however, India is still lagging behind its counterparts [1].

Electric vehicle transportation is potential resource to meet wide range of transportation policy goals and has boosted energy efficiency. EVs does not require any direct fuel for combustion and can fully depend on electricity. EVs will lead to great amount of energy to be secured, much better air quality, reduced noise or very less amount of noise, and have reduced the greenhouse gas emissions [3].

As of now INDIA has been lagging in the Electric vehicle department, but in coming years a lot of development will be there in terms of infrastructure and technologies to support and increase the penetration of EV's in Indian automobile industry. The Indian government has been trying to make this happen as soon as possible and plans to promote EV's. The Indian government aims at 100% e-mobility by year 2030. The foundation of India's journey to 100% e-mobility was laid in year 2017 when the first 100 e-vehicles unit were manufactured, and four charging stations were set-up [1].

But the usage of electric vehicles alone does not mean the reduction of air pollution. Electric vehicles will be effective in the reduction of pollution only if the major portion of electricity is obtained from renewable sources and the battery manufacturing plants are set-up far from the vehicle use region [2]. So not only the use of electric vehicles will reduce pollution, we need to develop new technologies to produce electricity such as nuclear power plants to deal with the increased requirement of power to run EV's. As of now, most of the electricity in India is generated by thermal power plants that use coal as their fuel. So, not only the development of new EV technologies, India must work on renewable sources of electricity generation too. So, a proper strategy has to be set up for the green future which is shown in figure 3.

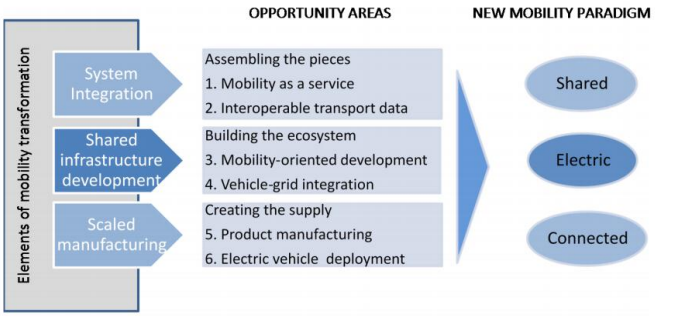


Figure 3: India's strategies to reduce vehicular emission and transform the whole mobility sector [2].

There has been a lot of research in the department of Green fuel vehicles and several review papers have focused on the general electric vehicles, electronic components, and energy management strategies. But there is a lack of published papers that present a comprehensive overview of EV's including powertrain configurations, power electronic components, energy management strategies, and new future technologies. This study fills this gap quite well by evaluating the practical vehicle applications and patents as well as published papers. The objective of this paper is to give a comprehensive overview of all-electric vehicles, evaluating the components used and the future of vehicles equipped with better BMS, IoT connectivity, and India's policies in favor of the development of EV's.

2. OVERVIEW ON ELECTRIC VEHICLES

Electric vehicles will be shaping our future for a better tomorrow. With less pollution, better air quality index, reduced noise pollution and efficient use of energy. Electric vehicles will be able to provide the growth to the stagnant automobile sector and increase the competitiveness in the industry. EV industry growth has attracted many investors and automobile companies towards electrified future. In the following section, the working principle and some light on the past present scenario of the electric vehicle are given.

2.1 Working Principle:

In simple words an electric vehicle works when electrical energy is used to power the motors from the controller which in turn receives the power from the rechargeable batteries. The battery pack stores energy in the form of electricity and supply power to the motors. The motor then uses the power (voltage) received from the battery pack to rotate and provide motion to the vehicle. The energy to be delivered to the motor is determined by the signal sent to the controller by potentiometer [4].

The article begins with a study on different EV configuration such as BEV, HEV, and FCBEV, then discussion on the engineering philosophy of EV development. Further, the advancement of Embedded systems used an overview of the EV market in INDIA is discussed.

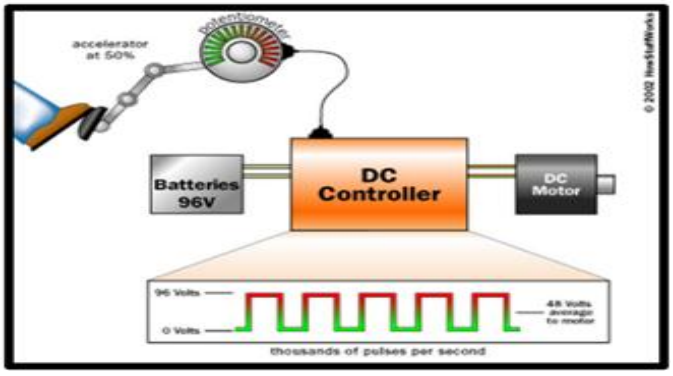


Figure 4: Working Principle of an Electric Vehicle [1]

2.2 Past Years Development

The first electric vehicle was developed in year 1834 and by the end of 19th century many auto-manufacturers in USA, Britain and France started making EVs. Figure 5 shows the World's first electric car. Due to the restrictions related to the batteries and therefore the rapid advancement in combustion engine (ICE) vehicles, EVs have almost vanished from the scene since 1930 [4]. The high cost, low top speed, and limited range of battery electric vehicles, compared to the combustion engine vehicles, led to a worldwide decline in their use; although electric vehicles have continued to be used in the form of electric trains and another niche uses [5].



Figure 5: World's first electric car. [5]

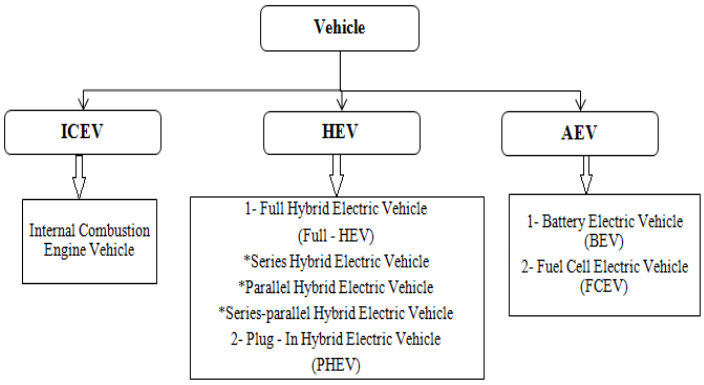


Figure 6: Classification of Vehicles [7]

During the start of the 21st century, with growing concern over the issues related to the hydrocarbon-fueled vehicles has led to an increased interest in electric and other alternative fuel vehicles. Increasing fuel prices, government regulation, and a general desire to scale back environmental damage have resulted in increased sales for fuel-efficient vehicles. Since 2010, there has been an increase in the combined sales of all-electric vehicles reaching over 5 million units delivered globally in September 2018, and the sales of light-duty all-electric vehicles and plug-in hybrids passed 8 million in December 2019 [5].

In the early times, the electric vehicles were powered by galvanic cells (batteries) and direct-drive reluctance motors. Then with further development first rechargeable batteries were used (Lead-acid batteries) with better induction motors but all EV's failed the race to newly developed IC engine vehicles in terms of efficiency, cost, and range.

2.3 Electric Vehicles at Present

The fist electric vehicle was introducing in the 18th century however the wide spread adaptation of EVs has been seen in the present decade only. Global EV sales have escalated from less than 10,000 in 2010 to 774,000 in 2016, surpassing two million cumulative sales. Electrification of vehicles is considered the main decarburization pathway for nearly all road-based transportation. Several countries on the account of worsening urban air quality have announced the intention to ban sales of internal combustion engine vehicles (ICEs) in the near future and replace the existing one with EVs [6].

3. ELECTRIC VEHICLE DRIVE-TRAIN

With recent development, vehicles can be classified into three main categories: internal combustion engine vehicles (ICEs), hybrid electric vehicles (HEVs) and all-electric vehicles (AEVs) as shown in figure 6 also. IC engine has been developed at the peak of their efficiency in the last several decades and now a sustainable source of energy requirement is been felt by all in the automobile industry. Now, further development is been made in the department of hybrid and electric vehicles.

From the above figure 6, we see that new technologies in vehicle propulsion are been developed other than internal combustion engines. Hybrid electric vehicles seem to be much practical in today's time due to the lack of infrastructure for the growth of AEV's. In HEV the ICE work in tune with the motors to provide better efficiency and less fuel consumption, the configuration of HEV's will be discussed in a further section. However, the HEV vehicles are also lagging behind in the Indian market due to their high cost and increased maintenance cost that normal ICE vehicles. So, the best alternative is to switch to AEV with the proper development of other sectors as well such as infrastructure for charging and production of electricity. The comparison of different characteristics of each drivetrain is given in table 1.

4. TYPES OF ELECTRIC POWER TRAIN

Through the past year's development, we have seen many applications of batteries and motor for automobile use. The structure of the drive-train is also different. So, in this section a brief discussion about the different types of power-train layout used for electric vehicles. Different types of power-train layout are hybrid-electric, All-electric and hydrogen fuel cell power-train.

4.1 HEV Power Train Configurations

In 1997, Toyota Prius was the first HEV to be introduced in the automobile market and since then more than 150 different models of HEV have been developed around the world. In hybrid electric vehicles (HEV) the conventional internal combustion engine (IC) is combined with an electric propulsion system that forms the combined hybrid vehicle drive train. The electric motors are provided to assist the conventional IC engines to provide either better fuel economy or better performance. So, the working of hybrid electric vehicle is mainly focused on the following key points:

- Maximize the fuel economy.
- Minimize vehicular emissions
- Minimize the system cost
- Better driving performance

Table -1: Characteristics of ICEVs, BEVs, HEVs, PHEVs and FCEVs [7]

Characteristics	ICEVs	BEVs	HEVs	PHEVs	FCEV
Propulsion System	ICE based	Electric motor based	Electric motor & ICE	Electric motor & ICE	Electric motor based
Energy storage	Fuel tank	Battery Ultra-capacitor Flywheel	Fuel tank Battery Ultra-capacitor Flywheel	Fuel tank Battery Ultra-capacitor Flywheel	Fuel cell Battery Ultra-capacitor
Energy source infrastructure	Refueling station	Electric charging facility	Electric power Refueling station	Electric charging station Refueling station	Hydrogen cylinder Hydrogen refiner & refueling station
Advantages	Fully commercialized Matured technology Better performance Simple operation Reliable	Zero emission Quite Smooth operation Energy efficient Independency from petroleum product Commercialized	Low emission Higher fuel economy Long driving range Reliable Commercialized Durability	High fuel efficient Lower emission Extended electric driving range V2G or G2V capability Quiet and smooth Operation	Ultra-low emission Competent driving range Highly efficient Independency from petroleum products Reliable Durable High cost
Drawbacks	Less efficient Harmful emission Poor fuel economy Comparatively bulky	Limited driving range Poor dynamic response High recharging time	Complex system Costly Bulky Increased component	High complexity Higher initial cost Battery technology Impact on grid	Slow dynamic response Not commercialized Sophisticated electronic controller
Major issues	Harmful emissions Fuel economy Dependency on petroleum products	Size & weight of battery pack Infrastructure for charging station	Size & weight of battery pack & ICE Integration of components	Charging station infrastructure Size & weight of battery pack & ICE Impact on grid	Cost of fuel cell Infrastructure for hydrogen conditioning, storage and refilling system

4.1.1. HEV's power-train configuration can be of four types [9]

a) Series Hybrid Electric Vehicles (SHEV):

In series hybrid electric vehicles there is no mechanical connection between ICE and wheels. SHEV involves internal combustion engine (ICE), generator, battery packs, rectifier, capacitors, converters and electric motors as shown in Figure 7. These configurations do not need supercapacitors (also known as ultracapacitors).

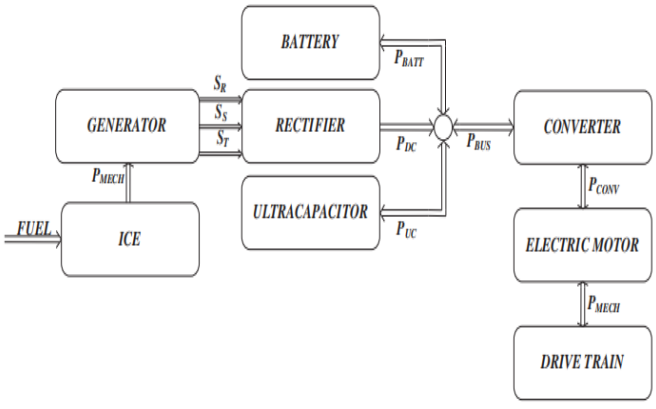


Figure 7: The Power Flow of SHEV Power Train [9]

In this type of configuration, the ICE is switched off when the battery is feeding power to the wheels and ICE only turns on when the batteries are low on charge. General working of a SHEV can be understood on the following way that when the power demand on the electric motor is less than the output of generator, the remaining power is used to charge ultracapacitor banks and the battery pack and if the power demand of the motor is higher than the output power of the generator than the additional power is supplied from the ultracapacitors [9]. So, the engine always works at it BEP.

b) Parallel Hybrid Electric Vehicles

In parallel hybrid electric vehicles, both IC engine power output and electrical power output are connected in parallel to drive the wheels as shown in figure 8. In this type of configuration, the IC engine is in always-on mode at almost constant power output at maximum efficiency point

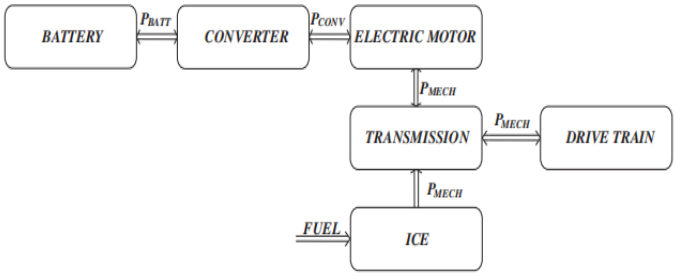


Figure 8: The Power Flow of PHEV Power Train [9]

The power flow can be understood as when the power required by the transmission is higher than the output power of ICE, the electric motor is turned on. In this arrangement the ICE and the electric motors both supply power to the transmission and in case the output power of the engine is more than that required by the transmission, the additional power is used for charging the batteries and the ultracapacitors [9]. This configuration leads to high maintenance and complex power management system algorithm. Yet this is the most commonly used hybrid vehicle powertrain.

c) Combination of Parallel and Series HEVs

This type of configuration is a combination of series and parallel hybrid system. As the name implies this configuration is neither fully parallel nor series hybrid system. Figure 9 shows the architecture of this drive. This configuration consists of a feature of both series and parallel hybrid systems. This type of drive is equipped in passenger vehicles or small automobiles.

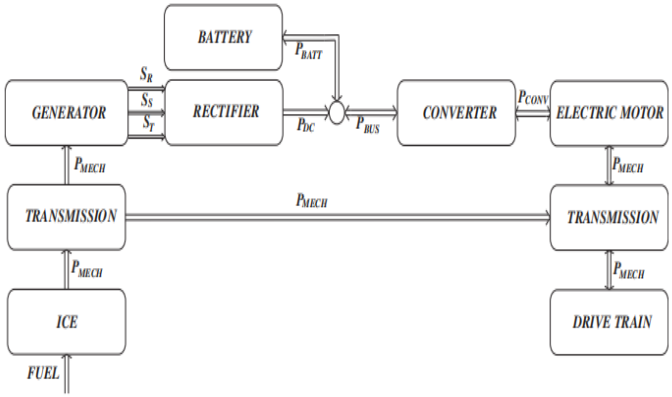


Figure 9: The Combination of Parallel and Series Hybrid Configurations [9]

This configuration of a combination of series and the parallel hybrid system works in conjunction with a power split device, a planetary gear that makes the power to be sent to wheels from the ICE. A planetary gearbox connecting the electric motor and the ICE is shown in Figure 10. With this configuration the speed of the engine can simply be adjusted by varying the speed of the generator.

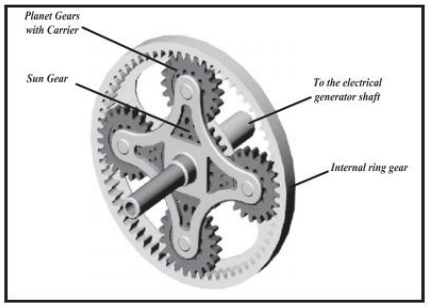


Figure 10: View of Planetary Gear System [9]

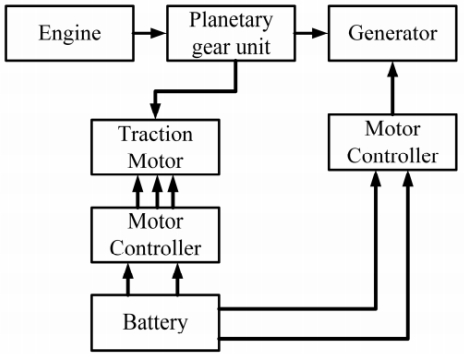


Figure 11: Drive train of SPHEV Using Planetary Gear

In figure 11. A planetary combination (PC) hybrid in Toyota Prius is shown. The figure shows the planetary gear unit coupled with ICE, alternator coupled to the sun gear and output torque is transmitted to the differential. The working of this power train is similar to that shown in figure 14.

d) Plug-in Hybrid Electric Vehicles (PHEV)

In plug-in hybrid electric vehicles, the capacity of the IC engine is much smaller than that of the electric propulsion system. In these vehicles, we have to charge the batteries every time as the IC engine is not capable of charging the batteries on the go and the main driving force is achieved by electric motors. These vehicles are a step towards all-electric vehicle (AEV), but there are some advantages of PHEV over AEV that are:

- Charging time is less.
- The car can be driven on ICE alone when the batteries die out.
- The vehicle range is more.
- Higher efficiency and less pollution
  PHEVs' has the ability to run solely on electric power only, which further reduces the carbon footprint smaller than the HEVs.
- The fuel cost associated with PHEV is also less.

A roadmap to the technological advancement in electric vehicles is shown in figure 12. At the beginning of the electric vehicle stage, a large engine is supported by a smaller battery and motor but as there were more advancement and need for clean vehicles the larger engines are being replaced by big batteries and motors.

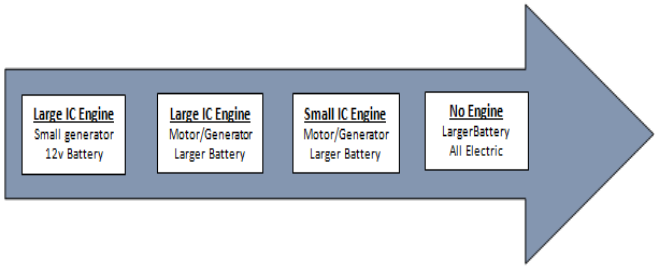


Figure 12: The Evolution Steps of the HEV Technologies [9]

4.2: Battery Electric Vehicles (BEV)

Battery electric vehicles are also known as All-electric vehicles (AEV) in which the vehicle propels by motors working on electricity. There is no IC engine assistance as in Hybrid Electric vehicles. BEV's have one or several high-energy batteries onboard which drive the traction motor with the help of embedded systems, which will be discussed in further sections. All BEV's have to rely solely on energy store in the batteries for propulsion, even with today's development in batteries technologies BEV's can only have an estimated range of 200-250km for smaller vehicles and 400-450km for heavy vehicles on a full charge. BEV's are simple in construction, operation, and easy to maintain. These vehicles are eco-friendly as they do not produce greenhouse gases (GHG) and produce less noise while running. All-electric drive train are very efficient and easier to maintain as compared to its ICE counterparts. The mentioned advantages coupled with limited driving range make them perfect vehicle for use in urban areas. [10]

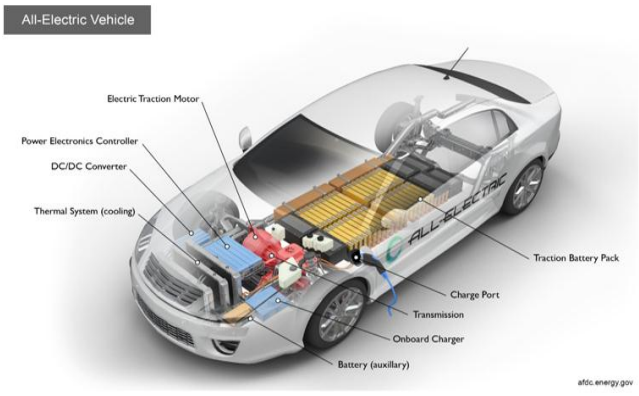


Figure 13: Layout of an All Electric Vehicle (AEV) [11]

Figure 13 shows the layout of a BEV and the major hardware components of the vehicle. The figure shows a front-wheel-drive BEV in which the motor drives the front wheels and the battery pack is placed at the center of the vehicle floor for better weight distribution. The weight of the BEV is higher than normal IC engine vehicles due to battery pack and a lot of research is going on to reducing the weight of the batteries and improve the handling of the vehicles. The weight of the vehicle severely affects the driving range that can be obtained from battery electric vehicles.

EVs as whole can be considered as a combination of different subsystems. The interaction between the systems is what required to make an EV work and for proper functioning multiple technologies can be employed [10]. From figure 14 an observation on the major components of EV's can be made but figure 14 show how the components interact with each other. Observing figure 14 a conclusion can be made that some parts interact significantly with other parts, whereas some parts interact less with other parts. But it is the combined works of all these parts/systems which make an EV operate.

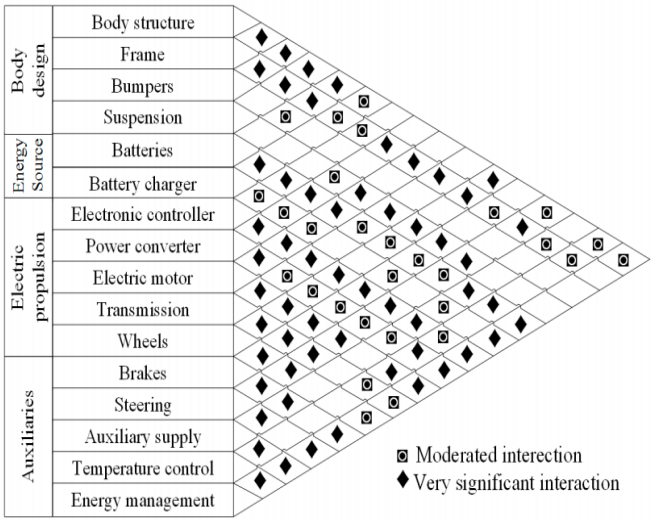


Figure 14: Major EV Subsystems and their Interactions [10]

4.2.1 EV Configurations:

It is concluded from figure 17 that how different subsystems interact. It is clear from figure 17 that some components of EVs are closely related while some have fewer interaction with other parts. In this section, a detailed explanation on how these systems interact with each other and make an EV work at its full potential is explained.

An electric vehicle is quite flexible in power transmission, unlike its ICE counterparts. In a BEV the power is transmitted through flexible electrical wires rather than rigid mechanical links in ICE or HEV. BEVs have only one moving part i.e. motor and can be controlled by different control algorithm and techniques.

All EV's are equipped with 3 major subsystems that are an energy source, electric propulsion, and the auxiliary.

The Energy source subsystem consists of:

a) The energy storage devices (battery, fuel cell, ultra-capacitor)
b) EMU (Energy management unit)
c) Energy charge unit

The Electric propulsion subsystem comprises of:

a) Electric motors
b) Power converter
c) Controller
d) Transmission
e) Driving wheels

The auxiliary subsystem consists of:

a) Power steering unit
b) Temperature control unit
c) Auxiliary power supply

In figure 15, the mechanical links are represented by the black lines, electric link between components are represented with Green lines and the control information communications are represented with blue lines. The motor speed is based on input from the accelerator pedal which interacts with the electronic controller a part of an embedded system; hence named as throttle-by-wire. The energy management unit along with electronic controller works together, to regulate regenerative braking and its energy recovery. It also works with the energy-refueling unit to control refueling and to monitor the usability of the energy source. The auxiliary power supply provides the required power with different voltage levels for all EV auxiliaries, especially the temperature control and power steering units.[12]

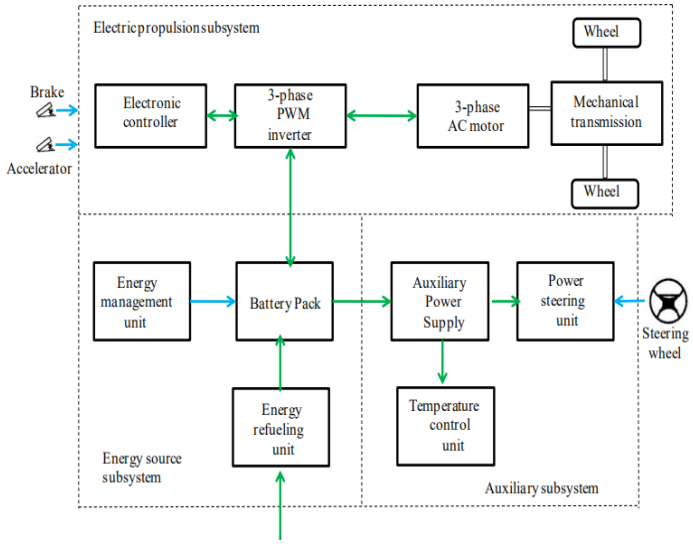


Figure 15: Modern configuration of Electric Vehicles [12]

The components mentioned in figure 18 are further discussed in the following section along with their specification and why they are best fit for modern EVs is also discussed.

4.2.2 Electric Vehicle design:

Modern EVs can be configured in many possible ways with variation in placement of different components of the propulsion system and the power sources. From figure 16, the different drive train configuration of EV's can be understood. The placement of electric motors can on the front wheels/axel or on the rear wheels/axle. Each configuration has its own characteristics and benefits. Motors are placed on front wheels in small or passenger vehicles so that they do not have a problem climbing steep hills and rear placement of motor is usually done on sports cars which require high accelerations of the line.

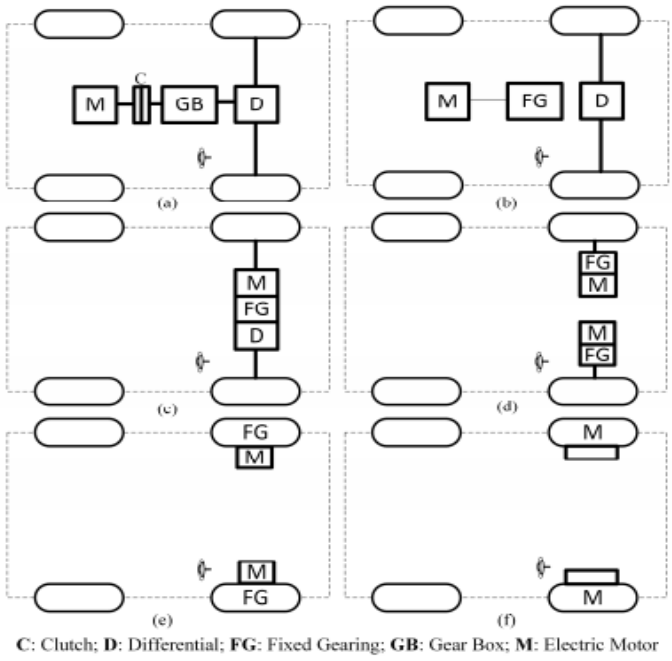


Figure 16: Different configuration of front-wheel drive (FWD) EV. (a) FWD vehicle which has motor powering the front-wheels (b) No clutch configuration (c) A single unit comprising of motor, gears and differential powering the front wheels (d) 2 motors are placed at separate front-wheels to obtain differential action (e) A modification of fig 19(d) configuration with fixed gearing placed within the wheels (f) A low-speed motor placed on the wheel rim to remove the mechanical gear system [10]

In the followings figure the bottom layout of different Vehicle is given. Through the figure, it can be observed that Motors can be placed on front-axel, rear or on both. The layout is designed with keeping many factors in mind such as weight of the car, cost and mileage range. Figure 17 Shows a rear-wheel drive configuration in which the battery and motor are placed at the rear-axel.

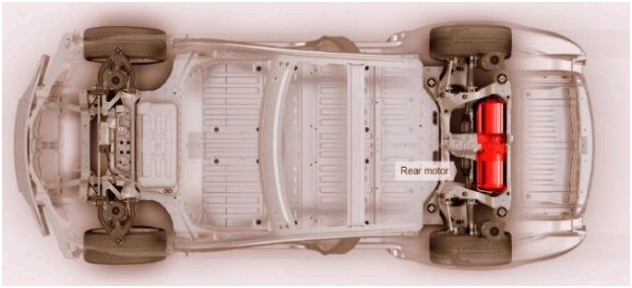


Figure 17: Tesla Model S, Rear Wheel Drive Configuration [10]

When there is a need for more traction and power, all-wheel drive (AWD) configuration is also used in electric vehicles by placing motors on the front and rear wheels/axels. All-Wheel

drive systems lead to increases in cost, weight, and complex power management strategies. These vehicles are equipped with torque vectoring systems to provide better performance and handling. Figure 18 shows a layout of AWD vehicle with electric motors placed on each front and rear axle.

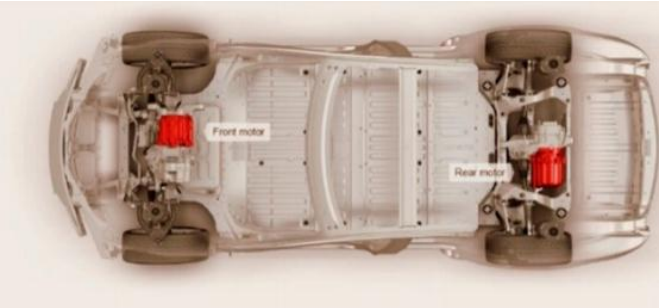


Figure 18: All-Wheel Drive Configuration [10]

5. MAJOR SUB-SYSTEM USED IN ELECTRIC VEHICLES

The working principle of the electric vehicles was discussed in earlier sections. Through this section, an explanation of the details of major sub-system used in electric vehicles is discussed with their working and interaction with other components. These sub-systems make an EV work and can be categorized in Energy Sources, Motors, Inverter and Converters, Embedded Systems and Charging Modules. Below the details of each such sub-system of EV mentioned above is given.

5.1. Energy Sources for EV's

EV's require electrical energy for propulsion. This energy requirement is fulfilled by different energy storage devices onboard of the vehicle. Each energy storage device is selected on the basis of some predefined criteria's and among them high energy density and high-power density being the most important ones. High specific power is required to give the motors great push thus increase acceleration and high energy density is required for long driving range. There are several other characteristics that have to be sort after to make a reliable power source such as fast charging, large cycle count, and long service. A lot of research goes into finding the right energy storage system (ESS) with desirable characteristics. Modern EV's are a combination of 2 or more. Energy-storage system are the devices which stores the electric-charge required to make the motor run and can be categorized as discussed below:

5.1.1Battery

Batteries are fundamentally an energy storage device which provides energy for the population of electric vehicles. Batteries are composed of two electrodes dipped in an electrolyte, the electrolyte acts as a medium for exchange of ions between electrodes to produce electric charge. Batteries have been used in vehicles since the early days, earlier the batteries were known as auxiliary batteries which provided energy to start the engine or for vehicle internal electronics.

But with the development of high energy capacity traction batteries, we are able to power the entire vehicle movement in conjunction with electric motors. The first battery to be developed was **Lead-Acid battery** by Gaston Plante in 1859. After 40 years of research, Nickle-Cadmium battery was developed in 1899 by Waldemar Jungner. Ni-Cd batteries made significant improvement in terms of storage capacity but were backed down due to its disadvantages including a voltage suppression issue that occurred as the battery ages, known as the memory effect.

After the Lead-acid and Ni-Cd battery, the research continues throughout the 20th century to find the best suitable battery for EV. It was in 1985 when the first Lithium-ion battery was developed and powers most of the EV's today. Later ZEBRA batteries and Nickel-Metal Hydride batteries were developed. Researchers are still in search of new and better battery composition to attain the desired performance goals. Some of the prominent batteries types once used or are in current development state are lead-acid, Ni-Cd, Ni-Zn, Zn/air, Ni-MH, Na/S, Li-polymer and Li-ion batteries. In this section, only the LI-ion battery will be discussed and compared to other alternative batteries available. As most of the EV's today run on Li-ion battery it is very important to understand the chemistry of the battery, cost and lithium deposits in the world. Through the means of the figure 19 it can also be seen that most modern EV's uses Li-ion batteries of different capacity and power.

Hyundai Ioniq	Li-Ion Battery	Capacity: 28kWh	Power: 88kW/118CP	Price: 29500€
Nissan Leaf	Li-Ion Battery	Capacity : 30kWh	Power: 80kW/107CP	Price: 30680€
VW E- Golf	Li-Ion Battery	Capacity : 24.2 kWh	Power: 100kW/136CP	Price: 37590€
Tesla Model S	Li-Ion Battery	Capacity : 100 kWh	Power: 193kW/259CP	Price: 123000€
Renault Twizzy	Li-Ion Battery	Capacity : 6.1 kWh	Power: 4 kW/5 CP	Price: 6750€

Figure 19: Comparative evaluation of different electric vehicles in market [13]

5.1.2 Lithium-Ion Battery Composition and Construction

In Li-ion batteries, the charge carriers are the lithium ions. In lithium-ion battery family, there are many different cell chemistries available and more are to be developed in future. The negative electrode of Li-ion batteries is principally composed of carbon (e.g. Graphite) or lithium titanite (Li4Ti5O12) and some other novel materials are under development such as Li metal and Li (Si) [14]. The electrodes are immersed into an electrolyte medium which are typically

composed of a mixture of lithium salts (e.g. LiPF6) and an organic solvent (e.g. diethyl carbonate) to allow for charge transfer [14].

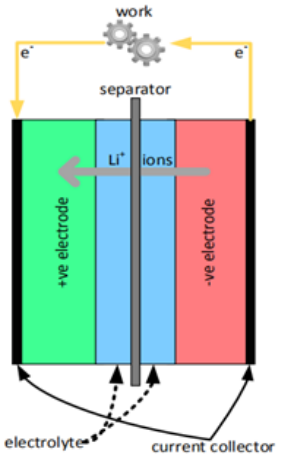


Figure 20: Construction of Li-ion Battery Cell [15]

The arrangement of Li-ion battery is shown pictorially in figure 20 which illustrates the ion flow in the battery when operating as an energy source (i.e. aa a galvanic device)—when the batteries are used to supply electric current than the electrons travels from negative electrode to the positive one and simultaneously the Li+ ions travel to the positive electrode through the electrolyte so that electro neutrality is maintained. During the charge mode i.e. when batteries act as electrolytic device or accept charge the flow of electrons and Li+ ions are reversed [14]. Through figure 21 a comparison of Li-ion batteries with other types of batteries is given on a plot of energy density VS weight. It simply shows that why the Li-ion batteries are the top priority of EV manufacture as the primary energy-storage device in modern electric vehicles.

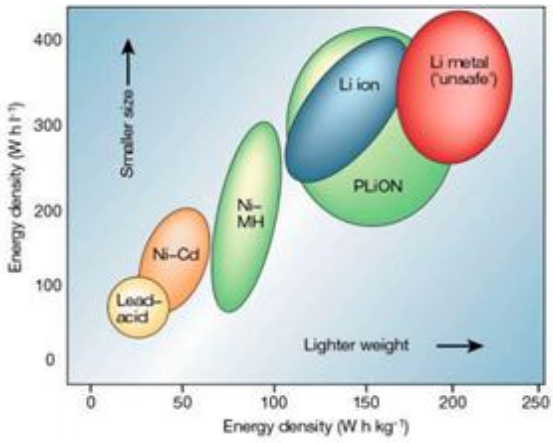


Figure 21: Range plot of EVs equipped with different composition of batteries [14]

The Li-ion battery family as of now consists of many practically applicable compositions which are Lithium Titanium Oxide (LTO), Lithium Nickel Manganese Cobalt Oxide (NMC), Lithium Nickel Cobalt Aluminum Oxide (NCA), Lithium Manganese Oxide (LMO), Lithium Cobalt oxide ($LiCoO_2$), and Lithium Iron Phosphate (LFP) [15].

Batteries for the EV are usually selected on the basis of specific parameters. Parameters on which batteries are compared are battery specific energy, specific power, energy density, weight and size. A comparison of different batteries on the basis of different parameters is shown in table 2.

5.1.3 Conclusion on Selection of Batteries:

Referring to table 2, it can be concluded that Na-$NiCl_2$ batteries have a better energy consumption rate (12.6 kWh/100km) as compared to other batteries. Na-$NiCl_2$ batteries are also important due to it's advantages such as low cost, high number of lifecycles and better functioning under harsh conditions. One major drawback of this technology is that these batteries tend to heat-up, which causes the battery electrolyte to solidify when the vehicle is not in use. So, in this case additional costing of cooling and heat management system is required which maintains the temperature and prevents failure.

According to studies, Li-S batteries are the one with highest energy consumption (17.2 kWh/100km) and are most suitable for high energy storage systems. Mi-MH batteries have intermittent properties and are considered outdated as well. In the recent development Li-Ion batteries have acquired the position of the most popularly used batteries for EV application due to its continuously declining cost, moderate energy consumption (14.7 kWh/100km), advance manufacturing technique, increased life cycle, light weight and high energy storage as well as driving range make them most suitable for EV application today.

Table 2: Comparing Lead-acid, Ni-MH, ZEBRA and Li-ion.

Battery type	Specific Energy (Wh/Kg)	Energy Density (W/Kg)	Mass of battery for EV to run a 100Km with 20 Wh/100km	Specific power (W/Kg)
Lead-acid	35-40	80-90	500-600 kg	285
Ni-MH	50-60	100-140	300-400 kg	200
ZEBRA (Na-$NiCl_2$)	100	160	200 kg	170
Li-ion	150-200	250-400	100-140 kg	260

5.1.4. Construction of Battery

The Battery packs used in electric vehicles consists of numerous battery cells connected in series. For example, in

Tesla model S has 7104 Li-ion cells which make a total of 85KWh battery pack. The cells used to make the battery module can be in different form and shape. The different form of battery cell used in EV is depicted in figure 22 and the cooling arrangement is shown in figure 23.

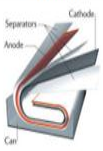
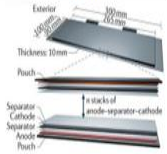
Shape	Cylindrical	Prismatic	Pouch
Diagram			
Electrode Arrangement	Wound	Wound	Stacked
Mechanical Strength	++	+	−
Heat Management	−	+	+
Specific Energy	+	+	++
Energy Density	+	++	+

Figure 22: Cell designs and Relative Strengths and Weaknesses [13]

All battery modules in EV consist of the battery module, battery management system and the cooling system. The Li-ion batteries used in EV today tend to overheat with use and are prone to failure which may lead to fires. The battery module provides protection to the batteries, manage batteries efficiently and safely. A cooling system is needed to make sure that the batteries operate at an optimum temperature range and provide consistent power output without the fear of fire breakouts due to high-temperature.

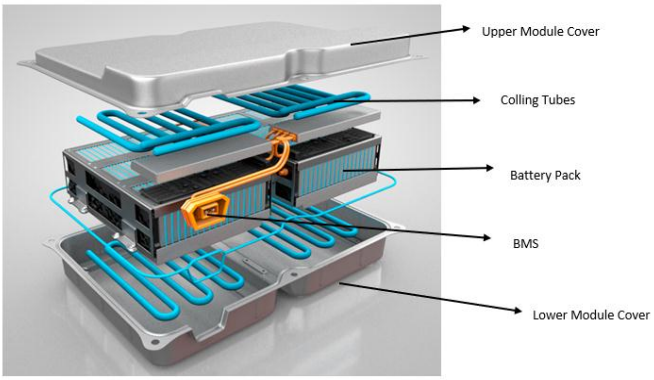


Figure 23: Arrangement of battery cells in the battery module. Special cooling tube are used to dissipate heat generated by the batteries [16]

Figure 24 shows the layout of the battery with the arrangement of the cooling tubes, cooling pads and upper-lower module cover. The design of battery modules placed in the vehicle has undergone significant development in the last 2-3 decades. Battery-module consists of many singular battery-cells to provide the required power output and takes a lot of space so, a proper layout is to be made to ensure better weight-distribution and thermal management in the electric vehicles. Through the figure 24 it can be observed that the how the battery module is designed and arranged on the vehicle chassis. The figure 24 shows the chassis design of the upcoming Jaguar I-Pace. [33]

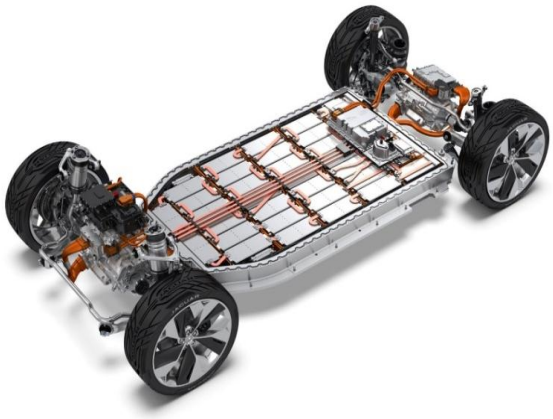


Figure 24: Chassis Design of the Electric-Vehicle [33]

5.1.5 Super Capacitors

The next category of the Energy Storage System (ESS) is Supercapacitors also referred to as Ultracapacitors (Electrochemical Capacitors). These electronic devices behave like a high-capacity capacitor with a capacitance value much higher than other capacitors. They have lower voltage limits that bridge the gap between electronic capacitors and rechargeable batteries. It can store about 10 to 100 times more energy per unit volume or mass than electrolytic capacitors. These electronic devices can accept and deliver charge at a much faster rate as compared to batteries and can tolerate many more charge and discharge cycles than rechargeable batteries.

Supercapacitors have gained a lot of application in the EV industry requiring many rapid charge/discharges cycle, rather than long term compact energy storage. Ultracapacitors comprises of two electrodes separated by an ion-enriched dielectric liquid. When a potential is applied across the UCS, the negative ions are attracted by the positive electrode and the negative electrode gathers the positive ions. In this way the charge gets stored on the electrodes and provide a considerably high-power density.

During the process there is no chemical reaction involved at the electrode, this way they tend to have long cycle life but also makes them low in energy density [9]. Super-capacitors have low internal resistance which makes it highly efficient, but also causes high output current when charged at extremely low SOC. The basic construction of an ultracapacitor is shown in figure 25.

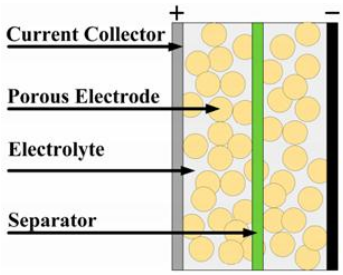


Figure 25: Construction of an Ultracapacitor [9]

While designing the power train of an EV the engineers have to take consideration of vehicle performance and energy storage requirements. From the above statement, we can apply that, a battery-based ESS has several challenges providing the impetus to look for additional solutions. While using battery-based ESS's high power density is required to meet the peak power demand. Although batteries with higher power densities are available, they are highly-priced as compared to the lower power density counterparts. So, to get high power density from the batteries we need to increase their size. However, this also causes an increase in cost. In addition to these issues, applications that require instantaneous power input and output typically find batteries suffering from frequent charge and discharge operations, which have an adverse effect on battery life.

In order to solve the problems listed above researchers have come up with an idea to combine a high-power density, but a lower energy density capacitor to the existing battery energy storage system to improve the efficiency. UC's are good options to compensate for the high peak of usage during short periods of time when the battery power is not sufficient. The combination of the battery-SC is known as a hybrid energy storage system (HESS). The basic idea of a HESS is to combine ultracapacitors (UCS) and batteries to achieve better overall performance [22]. Table no. 4 shows the typical characteristic of ultracapacitor cells compared to that of Li-ion battery.

Table 3: Comparison of Characteristics of UC and Li-ion Battery [22]

Chemistry	Nominal Cell Voltage (Volt)	Energy Density (Wh/kg)	Power Density (kWh/Kg)	Cycle Life (times)
UC	2.5/2.7	2-30	4-10	Over 1,000,000
Li-ion	3.6	250-400	150-180	Below 50,000

With the research and development in the field of energy storage systems (ESS), there are categories of ultracapacitor developed on the basis of chemical and physical composition of the energy storage mechanisms, active electrode material, cell characteristics, and their performance potentials (energy density, power density, etc.).

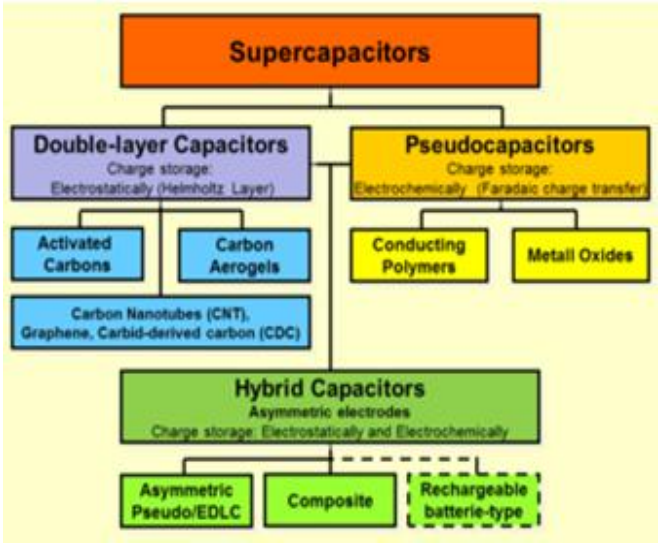


Figure 26: The Hierarchical Classification of Supercapacitors and Capacitors [19]

Through figure 26, it can be observed that modern age super-capacitors are divided into three major categories that will be discussed in detail below. These super-capacitors are divided into different categories on the basis of their cell structure and chemical formulation.

i. Double-Layer Capacitors (Carbon/Carbon) [21]

Most of the electrochemical capacitors currently available in the market are termed as electric double-layer capacitors (EDLC). The charge separation in a microscopically thin layer formed between a solid, conducting surface and liquid electrolyte results in energy storage in double-layer capacitors. The dominant electrode material is microporous, activated carbon [12,13]. The double-layer is formed in the microspores of the high surface area carbon material. Either an aqueous or organic electrolyte can be used.

ii. Pseudo-Capacitors [19]

These type of electrochemical capacitors uses metal oxides or conducting polymer electrodes with a high amount of electrochemical pseudo capacitance additional to the double-layer capacitance. Pseudo capacitance is a result of Faradaic electron charge-transfer with redox reaction, intercalation or electron sorption.

iii. Hybrid Capacitors [19]

Hybrid capacitors are usually a mixture of electrodes with differing characteristics, one exhibiting mostly electrostatic capacitance and other one electrochemical capacitance. Hybrid capacitors available are Lithium-ion capacitors

5.1.6 Regenerative Braking

After so many years of research and development, it has been recognized that combining ultra-capacitors with batteries would drastically reduce the stress on the batteries in electric vehicles application in which the batteries are subjected to high current pulses in both charge and discharge.

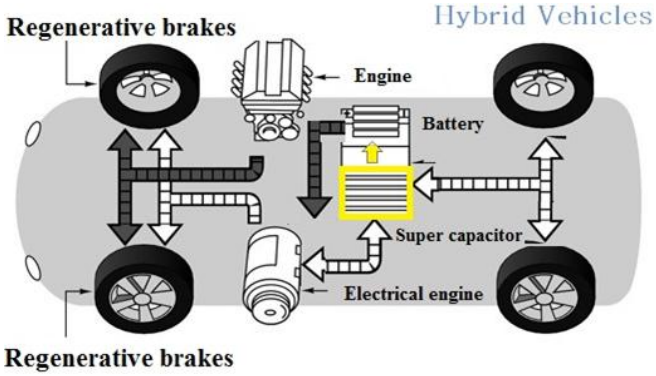


Figure 27: Arrangement of Supercapacitor in EVs [23]

From the arrangement shown in the figure 27, it is clear that the ultracapacitors have become an integral part of the EV powertrain. Many countries with less of lithium deposits have invested heavily in the development of UCs for their electric vehicles (E.g.: INDIA). The properties of UCs of getting charged and discharged quickly have significantly increased the life of the Li-ion batteries. The **figure 34** also shows the regenerative braking process in which the energy can be obtained by the backflow of current from the motor while braking and can be stored into the capacitors. The regeneration process would not have been possible without the supercapacitors property of getting recharged in 5-10 seconds.

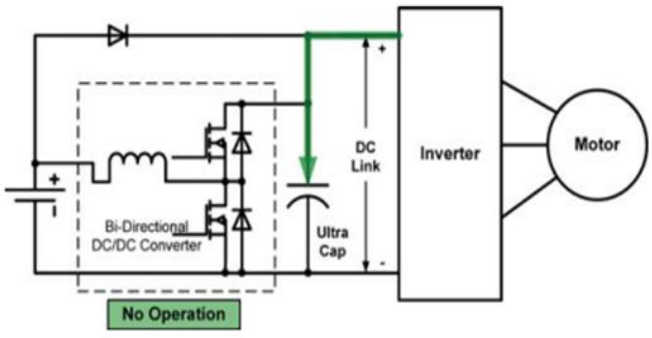


Figure 28: Flow of Current during Regenerative Process [21]

Due to recent research and development, we have been able to collect the energy produced while braking. This is called regenerative braking. Super or Ultra-capacitors play a very important role in regenerative-braking as they charge very quickly and also releases the charge a faster rate than typical batteries.

UCs can have a very large no. of charge and discharge cycles as compared to batteries so they are most suitable to charge through the small amount of energy produces while braking and supply this energy to the motors when battery power is not sufficient, like while heavy acceleration or in traffic.

5.2 Traction Motors

The next important parts in all-electric vehicles are the motors, because if the efficiency and the performance of motors are not satisfactory then the development in energy sources is of no use. The motor converts the electrical energy received from the power source into tractive motion which makes the vehicle move; same as that when fuel is burnt in the IC engine the engine converts thermal energy to mechanical motion. Motors also act as a generator during the regenerative braking process and send energy back to the energy source [10]. The selection of Tractive motor drive for an EV mainly considers the efficiency, reliability and cost of motors.

From a commercial point of view, the electric motors adopted or under serious consideration for electric vehicles includes the DC Motor, Induction Motor (IM), the Permanent Magnet (PM), Synchronous Motor, and the Switched Reluctance Motor(SRM). These five major types of tractive motors are considered best for EV application. Figure 29 and 30 show the different classification of DC and AC tractive motors for EV and HEV application.

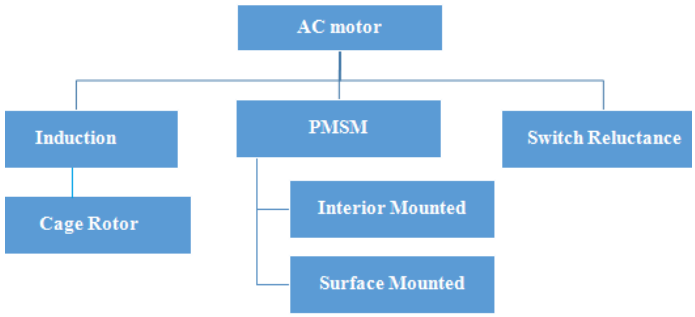


Figure 29: Classification of DC Motors [25]

In this section, the major five types of tractive motors will be discussed and compared on their characteristics and the most suitable for EV & HEV application is determined. All tractive motors can be classified into 2 types either AC or DC. Both the motors differ from each other in terms of working principle and construction.

However, AC motor drive has some clear advantages over DC motor drive such as higher efficiency, higher power density, effective regenerative braking, robustness, reliability and less need of maintenance

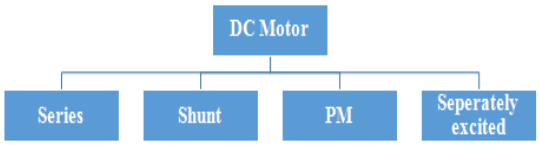


Figure 30: Classification of AC Motors [25]

In figure 31 the cross-sections of different electric motors for use in traction motors is shown.

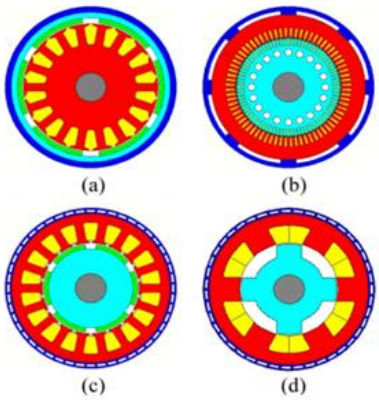


Figure 31: Cross section view of main traction motors used in EV a) DC motor. (b) IM (c) PM brushless motor. (d) SRM [25]

Through further discussion, explanation about different traction motors used in EVs today is concluded giving the advantages and disadvantages of each motor and deciding the most suitable for EVs today. Also, some highlights on their performance are given.

5.2.1 DC Motors

A DC motor is a class of rotary electrical motors that converts direct current electrical energy into the tractive effort. The most common type relies on the forces produced by the magnetic field. DC motors have been prominent in EV propulsion systems because of their torque-speed characteristics, simple design and their speed control is simple.

On the contrary, dc motor drives have a bulky construction, low efficiency, low reliability, and higher need of maintenance, mainly due to the presence of the mechanical commutator (brush and rings) [25]. After the growth of vector control for AC motors (synchronous and induction), the DC motors' attraction in traction applications diminished [24]. Nowadays DC motors are used for low power applications.

5.2.2 Induction Motors

An induction motor or asynchronous motor is an AC motor which produces electromagnetic induction from the magnetic flux of the stator winding to get the required current within the rotor to produce the tractive torque [27].

An Induction motor is of two types: wound type and squirrel-cage type. These are famously known as "Tesla motor". For EV traction motor application Squirrel-cage AC motors are quite desirable due to their reliability, robustness, less maintenance and the ability to work in a hostile environment. IM has become much mature in terms of technology than other AC motors in recent years. In induction motors, torque and field control can be decoupled using vector control methods and speed range can be extended using flux weakening in the constant power region.

The typical characteristics of an induction motor are shown in Figure 32.

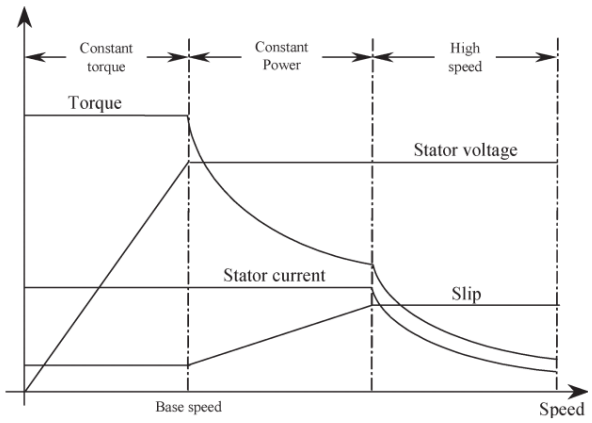


Figure 32: Characteristics of induction motor [24]

Induction motors also have some shortcomings which continue our search for more reliable and efficient traction motors for EV application. These drawbacks are mainly high loss, low efficiency, low power factor, and low inverter-usage factor, which become more significant for the high speed, large power motor. [25]

5.2.3 Permanent Magnet Synchronous (PMS) Motor

The Permanent magnet synchronous (PMS) motor is also known as brushless AC motor (BLAC). These motors are the most serious competitor to the induction motor in traction application in EVs. Many car manufacturers have already used these motors in their EVs as shown in the table below. These motors have several advantages: higher power density, higher efficiency and the more effective distribution of heat into the environment [24]. However, these motors have a narrow constant power region and to widen the speed range without over sizing the motor, conduction angle of power convertor is used in traction application of the motor. Figure 39 shows the characteristics of the BLAC motor with & without conduction angle control. With the use of conduction angle control speed range of BLAC can be extended to three or four times the base speed.

The conduction angle control is done by multiphase pole-changing IM drive. However, the shortcoming of the motor is that they tend to demagnetize due to the heart or armature reaction so a lot of attention has to be paid while designing the cooling and motor control system

5.2.4. Switched Reluctance Motor (SRM)

SRM motors have been developed as a good traction motor for EV application in last years. The development of SR motors has some advantages over other type of traction motors that are: simple & rigid construction, fault tolerance, simple control and excellent torque-speed characteristics.

However, there are some disadvantages too such as high noise, high torque ripple, special convertor topology and electromagnetic interference.

5.2.5 Brushless DC motor (BLDC)

This type of motor is conceptually the result of reversing the stator and the rotor of permanent magnet DC motor. BLDC are fed by rectangular waves in contrast to BLAC motors which are fed by sinusoidal waves. In these motors the stator are made up of permanent magnets (PM) and the rotors have brushes Their main advantages are the deletion of the brushes, their compactness, high efficiency and high energy density. However, the disadvantage overshadows the advantages due to which this motor has lost its application in the EV industry. The major drawbacks to this motor are its bulky structure, low efficiency, high heat generation because of the brushes. The heat produced is difficult to remove as it is generated in the center of the rotor and there is associated loss in efficiency with increasing temperature.

Table 4: Comparison of all Motors [10]

Motor Type	Advantage	Disadvantage	Vehicles Used In
Brushed DC Motor	• Maximum Torque at low speed	• Bulky structure • Low efficiency • Heat generation at brushes	Fiat Panda Elettra (Series DC Motor), Conceptor G-Van (Separately excited DC Motor)
Permanent Magnet Brushless DC Motor (BLDC)	• No rotor copper loss • More efficiency than induction motors • Lighter and smaller • Better heat dissipation • More specific power • More torque density	• Short Constant Power • Decreased torque with increase in speed • High cost because of PM	Toyota Prius (2005)
Permanent Magnet Synchronous Motor (PMSM)	• Operable in different speed ranges without using gear systems • Efficient and compact • Suitable for in-wheel applications • High torque even at very low speed	• Huge iron loss at high speeds during in-wheel operation	Toyota Prius, Nissan Leaf, Soul EV
Induction Motor (IM)	• The most mature commutator less motor drive system • Can be operated like a separately excited DC motor by employing field orientation control		Tesla Model S, Tesla Model X, Toyota RA V4, GM EV1
Switched Reluctance Motor (SRM) **Synchronous Reluctance Motor**	• Simple and robust construction • Low cost and High speed • Less chance to hazard • Long constant power range • High power density • Robust and Fault tolerant • Small and Efficient	• Very noisy • Low efficiency • Larger and heavier than PM machines • Complex design and control • Problems in controllability and manufacturing • Low power factor	Chloride Lucas

5.3 Power Convertors in EV's

In electric vehicles the energy storage system/devices (ESS) as discussed above, stores and supply electricity for the functioning of the electric vehicles. The system usually consists of battery and ultracapacitors (UC) only and both these electric devices stores energy as a DC charge. All-electric vehicles have to be charged before use and this charge is obtained by EV chargers which supply AC current to the vehicle's ESS. So, to store the charge in DC device AC charge has to be converted to DC charge and this is done by the converter (AC/DC).

When the ESS is fully or partially charge and power has to supplied to motors for propulsion from the batteries. Modern motors are AC charge type as discussed in the last section, so the charge from batteries i.e. DC has to be converted to AC again which is done by **Invertors (DC/AC)**. Figure 33 shows the typical placement of different power convertors in EV with the energy flow direction.

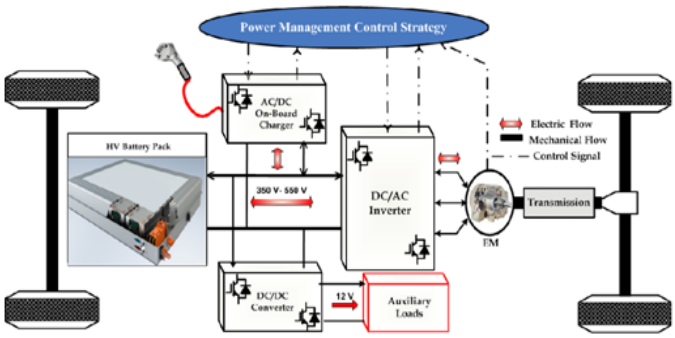


Figure 33: Power Conversion in Electric Vehicles [28]

As seen in the above figure all the arrangement of converters DC/AC or AC/DC is required to recharge the energy storage of EVs and then use it to propel the vehicles. In the above figure, it shows an AC/DC converter at the charging point which converts AC charge from power-grid to store DC charge in batteries. Then a DC/DC converter is shown for auxiliary loads which include the power required for electronics inside the vehicle like heater/AC or lightning inside the vehicle. Different auxiliary devices work at different DC voltages but battery supplies a rated DC voltage at all charge, so a DC/DC converter can be used to interface the elements in the electric powertrain by boosting or chopping the voltage levels [28]. Finally, a DC/AC inverter is shown in the figure which converts the DC charge from the battery to AC charge for the working of the motors. These converters are linked to a DC link to work as bi-directional converter while recharging the batteries or UC during the regeneration process. A simplified arrangement of convertor and invertors in EVs is shown in figure 33.

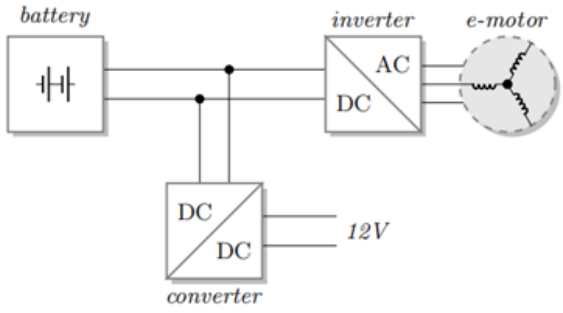


Figure 33a: Illustration of a Basic Electric Power Train of a Full Electric Vehicle [29]

6. HYDROGEN POWERED VEHICLES

The use of hydrogen as a mean to power our vehicles has attained the attention of many researchers and auto manufacturers as clean & more efficient energy source for our vehicles. The development of Hydrogen-powered vehicles is still in its infancy and no major applications of the Fuel cell technology have been made but, this fact only justifies the limitation of current times for the use of hydrogen in our vehicles. However, Hydrogen fuel cell technology has a very upscale potential to be used as fuel in our vehicles in the very near future [33]. The major interest of the use of hydrogen in vehicles is due to higher efficiency than IC engine, high part-load capacity, no tailpipe emission (only water vapors), quite running, higher driving range (over 500km) and faster refueling. In the year 2002, the Bush Administration along with industry-government cooperation announced the "Freedom CAR" initiative, to develop a fuel cell vehicle. [32]

Through this initiative, the use of Hydrogen to power our vehicles was started either as direct injectable fuel to IC engines or as a fuel cell to power motors in EVs replacing batteries. A comparison between all category of green vehicles is done in table. 6.

Table 5: Comparison between HEV, EV and Fuel Cell Vehicles [31]

Hybrid	**Substantial (100-200%)**	**Some zero-Emission range possible**	**Near term (2-7 y)**	**Substantial**	**Grade climbing ability or towing capacity may be reduced**
Fuel Cell	Very High (150-300%)	Low to zero tailpipe and total	Mid term (7-12 y)	Very high (>20%)	Potential petroleum independence
Battery-Electric	Very High (300%)	Zero tailpipe	Near term (2-7 y)	Very high (>20%)	Energy storage, range concerns, low petroleum use
Vehicle system	Fuel Economy Improvement Potential	Criteria Emissions	Years to mass Market introduction	Current Incremental Cost	Other Issues

As part of electric vehicle research, this section will be dedicated to the hydrogen fuel cell technology to replace the battery in BEV. The Fuel cells the most crucial part of hydrogen-electric vehicles. Hydrogen used in fuel-cells act as a energy carrier that can be produced from low-carbon source and has high specific energy as compared to most batteries. The first mass-produced fuel-cell electric vehicles (FCEVs), which use polymer electrolyte membrane (PEM) fuel cells, were introduced in 2013–2014 by Hyundai, Toyota and Daimler [31].

6.1. Fuel Cells

Fuel cells are the powerhouse for future electric vehicles. Fuel cells generate electricity by electrochemical reaction. A fuel cell consists of a negative electrode (ANODE), a positive electrode (Cathode) and the electrolyte between them.

The Hydrogen is stored in an onboard tank and supplied to the anode where it gets oxidized. The ions created to travel through the electrolyte to the cathode and combine with the other reactant introduced [34]. The creation of ions as the

product of oxidization of hydrogen at anode produces electricity.

Hydrogen is preferred as fuel for FCs due to its high energy content and does not produce pollution (produces only water on combustion) and is abundantly available in nature in the form of different compounds such as hydrocarbons.

The chemical reaction inside a Fuel-cell is stated below: [34]

$$2H_2 + O_2 = 2\ H_20$$

The figure 34 shows the working principle of a fuel cell in which fuel and oxygen are taken in, the exhaust of oxidized hydrogen (H2O) and electricity produced as the product of the chemical reaction is shown. Figure 35 shows the construction of fuel cell.

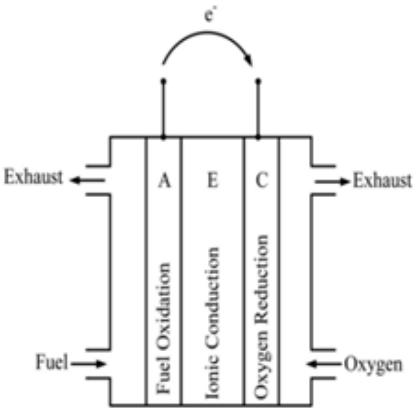


Figure 34: Working Principle of Fuel Cell [10]

Fuel cell application in EV has many advantages from producing electricity efficiently from fuel, noiseless operation, faster refueling, minimal emissions, durability and the ability to provide high density current output. A major drawback of this technology is its high price and as hydrogen has lower energy density compared to petroleum derived fuel. Therefore, larger fuel tanks are required for FCEVs to store large quantities of hydrogen fuel. So, additional cost goes into making larger tanks and also to make them capable enough to contain large amount of hydrogen fuel without the risk of any explosion in case of an accident.

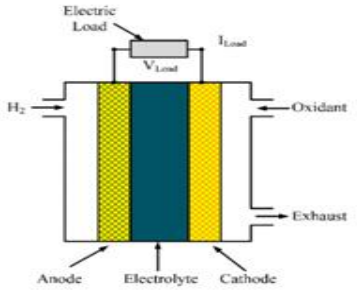


Figure 35: Fuel Cell Configuration [10]

Today hydrogen production is very expensive and less efficiency is transferred to from well-to-wheel pathway. The below give figures shows the entire pathway for energy transfer for battery and hydrogen powered vehicles. For battery energy the entire pathway is 77% efficient. Whereas for FCV's, the energy from the electric plant is used for the electrolysis process to produce hydrogen gas from water.

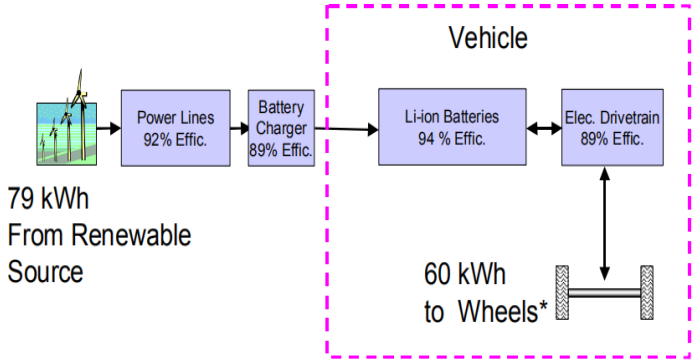


Figure 36: Well-to-Wheel Energy Pathway for Battery Electric Vehicle. [32]

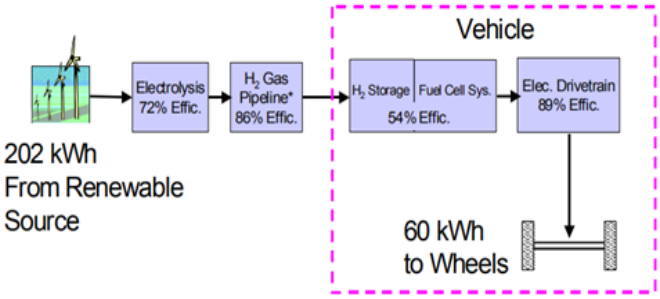


Figure 37: Well-to-Wheel Energy Pathway for Fuel Cell Vehicle [32]

As concluded from the above figure 36 and figure 37, a lot of energy goes into production and use of hydrogen as a fuel for fuel cell vehicle and a major amount of this energy is produced with fossil fuels which put severe stress on Power Grid.

- FCEVs have higher purchase prices than conventional vehicles, and similarly to BEVs, this is attributed to their electrochemical power supply. The hydrogen storage tank and fuel-cell system are the most expensive components because of the inclusion of expensive materials and equipment such as platinum, carbon fibre, humidifiers and heat exchangers. The cost of nearly all these components will decline considerably with increased manufacturing volumes, with the main exception being platinum-group metal (PGM) catalysts owing to their scarcity. [30].

7. CONCLUSION AND FUTURE DEVELOPMENT

Even though a lot of development and revolution has been made in terms of Green transportation and it has been presented through this paper, but there is a lot to do in future to make these Vehicles feasible and economically viable. The paper presents an overview of the Technological roadmap of the technologies from the earliest to the recently developed technology for EV's. Detailed analysis of different Electric Vehicles is performed to know the advantages and disadvantages of each type- HEV, BEV, FCEV. A study on different components that makes EV run is laid out with the earliest and recent development in them—Batteries, Motors, Power convertors and Super capacitors. Through the study, it was concluded that Li-ion battery and Permanent brushless AC motor are the most attractive ones for future application In EV's considering their price and energy density. Different Power-train configuration for Hybrid as well as for pure electric vehicle was studied with proper emphasis on energy flow and energy management systems. Lastly, a discussion on Fuel-cell Electric is presented to make the basics for future development, as this technology will most probably run our vehicles in future with zero emissions and very less cost. Furthermore, this paper provides with basic analysis and information for future research in the Electric Vehicle domain.

REFERENCES

1. Ankit Kumar, S K Choudhary and Chethan K N (2018) 'COMMERCIAL VIABILITY OF ELECTRIC VEHICLES IN INDIA', International Journal of Mechanical Engineering and Technology (IJMET), Volume 9 (Issue6, June 2018), pp. 730–745 [Online]. Available at: http://iaeme.com/MasterAdmin/uploadfolder/IJMET_09_06_082/IJMET_09_06_082.pdf (Accessed: 5th December 2019)
2. R. Vidhi and P. Shrivastava, "A Review of Electric Vehicle Lifecycle Emissions and Policy Recommendations to Increase EV Penetration in India," Energies, vol. 11, no. 3, p. 483, Feb. 2018.
3. Rishabh Jain (2018) 'The Progress of Electric Vehicles', International Research Journal of Engineering and Technology (IRJET), Volume 05(Issue 10, Oct 2018), pp. 2395-0072 [Online]. Available at: https://www.researchgate.net/publication/333311907_THE_PROGRESS_IN_ELECTRICAL_VEHICLE (Accessed: 5th December 2019).
4. C C Chan (2015) ELECTRICAL ENGINEERING – Vol. III - Electric Vehicles - C CChan. Encyclopedia of Life Support Systems (EOLSS) [Online]. Available at: https://www.eolss.net/Sample-Chapters/C05/E6-39A-05-08.pdf (Accessed: 5th December 2019).
5. Wikipedia contributors (23 May 2020 04:19 UTC) History of the electric vehicle, Available at: https://en.wikipedia.org/w/index.php?title=History_of_the_electric_vehicle&oldid=958322433 (Accessed: 5th November 2019).
6. Cano, Z.P., Banham, D., Ye, S. et al. (2018) 'Batteries and fuel cells for emerging electric vehicle markets.', Nat Energy, Volume 03(), pp. 279–289 (2018) [Online]. Available at: https://doi.org/10.1038/s41560-018-0108-1 (Accessed: 5th December 2019).
7. Saqib Jamshed Rind, Yaxing Ren, Yihua Hu, Jibing Wang, and Lin Jiang (2017) 'Configurations and Control of Traction Motors for Electric Vehicles: A Review', Chinese Journal of Electrical Engineering, Volume 03(No. 3), pp. 279–289 (2018) [Online]. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8250419 (Accessed: 5th December 2019).
8. Monica Ralston and Nick Nigro (2011) 'PLUG-IN ELECTRIC VEHICLES: LITERATURE REVIEW', CENTRE FOR CLIMATE AND ENERGY SOLUTION, (), pp. [Online]. Available at: https://www.ourenergypolicy.org/wp-content/uploads/2012/04/PEV-Literature-Review.pdf (Accessed: 5th December 2019).
9. Bayindir, Kamil & Gözüküçük, Mehmet &Teke, Ahmet. (2011). A comprehensive overview of hybrid electric vehicle: Powertrain configurations, powertrain control techniques and electronic control units. Energy Conversion and Management. 52. 1305-1313. 10.1016/j.enconman.2010.09.028.
10. F. Un-Noor, S. Padmanaban, L. Mihet-Popa, M. Mollah, and E. Hossain, "A Comprehensive Study of Key Electric Vehicle (EV) Components, Technologies, Challenges, Impacts, and Future Direction of Development," Energies, vol. 10, no. 8, p. 1217, Aug. 2017 [Online]. Available at: http://dx.doi.org/10.3390/en10081217
11. Somayaji Y., Mutthu N.K., Rajan H., Ampolu S., Manickam N. (2017). Challenges of Electric Vehicles from Lab to Road. 2017 IEEE Transportation Electrification Conference (ITEC-India), https://ieeexplore.ieee.org/document/8333880/
12. IIT Guwahati (2013-02-08) Introduction to Hybrid and Electric Vehicles, Available at: https://nptel.ac.in/courses/108/103/108103009/ (Accessed: 5th November 2019).
13. Iclodean, Calin &Varga, Bogdan &Burnete, N &Cimerdean, Denisa&Jurchiş, B. (2017). Comparison of Different Battery Types for Electric Vehicles. IOP Conference Series: Materials Science and Engineering. 252. 012058. 10.1088/1757-899X/252/1/012058.
14. Miao, Yu &Hynan, Patrick & von Jouanne, Annette &Yokochi, Alexandre. (2019). Current Li-Ion Battery Technologies in Electric Vehicles and Opportunities for Advancements. Energies. 12. 1074-1094. 10.3390/en12061074.
15. Chian, T & Wei, W & Ze, E & Ren, L & Ping, Y & Bakar, N &Fauzan, Ts Dr Mohd Faizal & Sivakumar, S. (2019). A

Review on Recent Progress of Batteries for Electric Vehicles. International Journal of Applied Engineering Research. 14. 4441-4461.

16. Smith S.A (2013-02-08) Flord and scheugenpflug collaborate to scale up EV thermal designs, Available at: 16.https://www.google.com/url?sa=i&url=https%3A%2F%2Fchargedevs.com%2Ffeatures%2Flord-and-scheugenpflug-collaborate-to-scale-up-ev-thermal-designs%2F&psig=AOvVaw0r4t3cVI7HE-ihRkhSw-hg&ust=1581436584041000&source=images&cd=vfe&ved=0CAMQjB1qFwoTCOD9jaSux-cCFQAAAAAdAAAAABAD (Accessed: 5th November 2019).
17. Paul W. Gruber Pablo A. Medina Gregory A. Keoleian Stephen E. Kesler Mark P. Everson Timothy J. Wallington (2011) 'Global Lithium Availability', Journal of Industrial Ecology, Volume 15(Issue 5), pp. [Online]. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1530-9290.2011.00359.x (Accessed: 5th December 2019).
18. Cano, Z.P., Banham, D., Ye, S. et al. Batteries and fuel cells for emerging electric vehicle markets. Nat Energy 3, 279–289 (2018). https://doi.org/10.1038/s41560-018-0108-1
19. Wikipedia contributors (14 May 2020 02:16 UTC) Supercapacitor, Available at: https://en.wikipedia.org/w/index.php?title=Supercapacitor&oldid=956564666 (Accessed: 5th November 2019).
20. Andrew Burke (2009) 'Ultracapacitor technologies and application in hybrid and electric vehicles', International Journal of Energy Research, Volume 34(Issue 2), pp. 298-395 [Online]. Available at: https://doi.org/10.1002/er.1654 (Accessed: 5th December 2019).
21. J. Cao and A. Emadi, "A New Battery/UltraCapacitor Hybrid Energy Storage System for Electric, Hybrid, and Plug-In Hybrid Electric Vehicles," in IEEE Transactions on Power Electronics, vol. 27, no. 1, pp. 122-132, Jan. 2012.doi: 10.1109/TPEL.2011.2151206
22. Kouchachvili, Lia &Yaïci, Wahiba&Entchev, E. (2018). Hybrid battery/supercapacitor energy storage system for the electric vehicles. Journal of Power. 374. 237–248. 10.1016/j.jpowsour.2017.11.040.
23. Hashemnia, Mohammad Naser &Asaei, Behzad. (2008). Comparative study of using different electric motors in the electric vehicles. 1 - 5. 10.1109/ICELMACH.2008.4800157.
24. S. J. Rind, Y. Ren, Y. Hu, J. Wang and L. Jiang, "Configurations and control of traction motors for electric vehicles: A review," in Chinese Journal of Electrical Engineering, vol. 3, no. 3, pp. 1-17, December 2017.
doi: 10.23919/CJEE.2017.8250419
25. M. Zeraoulia, M. E. H. Benbouzid and D. Diallo, "Electric Motor Drive Selection Issues for HEV Propulsion Systems: A Comparative Study," in IEEE Transactions on Vehicular Technology, vol. 55, no. 6, pp. 1756-1764, Nov. 2006.doi: 10.1109/TVT.2006.878719
26. Wikipedia contributors (17 May 2020 20:16 UTC) Induction motor, Available at: https://en.wikipedia.org/w/index.php?title=Induction_motor&oldid=957240578 (Accessed: 5th November 2019).
27. Monzer Al Sakka, Joeri Van Mierlo and Hamid Gualous (2010) 'DC/DC Converters for Electric Vehicles', in SerefSoylu (ed.) Electric Vehicle.:intechopen, pp. DOI: 10.5772/17048.
28. Lukasiewycz, Martin & Steinhorst, Sebastian & Andalam, Sidharta & Sagstetter, Florian &Waszecki, Peter & Chang, Wanli & Kauer, Matthias & Mundhenk, Philipp & Shanker, Shreejith& Fahmy, Suhaib & Chakraborty, Samarjit. (2013). System architecture and software design for Electric Vehicles. Proceedings - Design Automation Conference. 1-6. 10.1145/2463209.2488852.
29. Cano, Zachary & Banham, Dustin & Ye, Siyu & Hintennach, Andreas & Lu, Jun & Fowler, Michael & Chen, Zhongwei. (2018). Batteries and fuel cells for emerging electric vehicle markets. Nature Energy. 3. 279-289. 10.1038/s41560-018-0108-1.
30. Craig Davis, Bill Edelstein, BillEvenson, AvivaBrecher and Dan Cox (2003) Hydrogen Fuel Cell Vehicle Study, American Physical Society.
31. Eaves, Stephen & Eaves, James. (2004). A Cost Comparison of Fuel Cell and Battery Electric Vehicles. Journal of Power Sources. 130. 208-212. 10.1016/j.jpowsour.2003.12.016.

BIBLOGRAPHY



First Author: Amit Sachdeva is a Mechanical Engineer, studying MBA Business Analytics from Delhi Technological University. He was awarded the B.Tech Degree from Bharati Vidyapeeth (DU), College of Engineering, PUNE in 2020.



Co-Author: Dadaso D Mohite is Mechanical Design Engineer, Working as Assistant Professor at Bharati Vidyapeeth (DU), COE, Pune. He was awarded the B.Tech (Mechanical) Degree from BV (DU), COE, Pune in 2013, and M.E. Mechanical (Design) from SPPU, Pune in 2015

View publication stats

2020 IEEE Intelligent Vehicles Symposium (IV)
October 20-23, 2020. Las Vegas, USA

Robust and Adaptive Traffic Signal Control for Unstructured Driving Scenarios in the Developing World

Ishan Maheshwari[1,*] Ujwal Padam Tewari[2,*] Chetan Arora[3]

***Abstract*— Adaptive traffic signal control (ATSC) systems are key to efficient traffic management in smart cities. While RADAR or inductive loop based systems are regularly used for the task in the developed world, financial and infrastructure constraints precludes use of sophisticated and expensive sensors in the developing countries. RGB cameras are cheap and easy to deploy, making their usage attractive for ATSC. However, unstructured driving scenarios, and the requirement to work 24×7 pose a significant challenge to the camera based systems. In this paper we propose to use traffic tailback length (TTL) for training a Deep Reinforcement Learning (RL) Agent in an ATSC system. We show that TTL can be calculated by simply performing a 'background subtraction' on the input video captured from an RGB camera. The background subtraction operation is well studied in computer vision and can be robustly performed in low illumination, and rainy, snowy, or hazy weather. The operation does not require detecting vehicles explicitly, thus making the proposed system applicable in unstructured driving conditions, involving unseen vehicles. We show, in a simulated environment using SUMO, that our technique outperform the state of the art, which uses explicit vehicle detection information as well as speed of the vehicle, in terms of vehicle throughput, queue length, and average delay.**

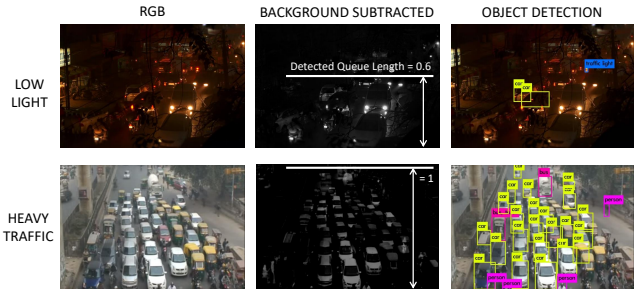


Fig. 1: *The figure shows that the traffic tailback length (TTL), which is defined as the distance of the last vehicle standing in a lane at an intersection, can be robustly computed (compared to object detection) even in low illumination and congested self occluding traffic scenarios, by performing 'background subtraction' on the input video. Our Reinforcement Learning (RL) agent uses TTL as a cue to learn traffic patterns and regulate phase switching for optimal throughput.*

I. Introduction

Vehicle density in the cities, especially in the developing world, is increasing exponentially. This is both due to the population growth accompanied by increasing urbanization, as well as rise in the car ownership with improving economic prosperity. However, the road infrastructure is often not able to keep pace with this increase due to financial as well as space constraints. In such a situation adaptive traffic signal control becomes the key to achieving maximum utilization of the existing road infrastructure.

A traffic controller is a system designed to regulate traffic at a single or a set of traffic intersections using static or adaptive rules. While the former creates a fixed phase switching routine based upon the historical traffic data, the adaptive controllers measure the traffic density and vehicle delays at each approach, and then generate an optimal green time for various phases so as to maximize the vehicle throughput at the junction. A typical adaptive traffic signal controller (ATSC) may include sensors, such as RADAR, inductive loops, or cameras, and uses conventional or modern artificial intelligence techniques to learn the incoming traffic patterns. While RADAR and inductive loops are costly and difficult to deploy, the cameras are cheap, and non-intrusive. However, the camera based systems are difficult to operate in low illumination or adverse weather conditions, such as rain, snow or haze, when the visibility is low. Besides, these systems often assume lane discipline, and vehicle types to be known a-priori (to detect and track the object and measure the wait time). The assumptions usually do not hold in unstructured traffic in the developing world. The line of sight requirement in camera and RADAR based systems poses additional challenges in the highly congested traffic scenarios due to partial or complete occlusion by other vehicles. In Fig. 1, we show some of the indicative challenging examples for an ATSC system relying on explicit object detection.

Our focus in this paper is on designing a practical and robust adaptive traffic signal control system specifically tailored to the conditions and requirements of the developing countries. Our system uses RGB cameras due to their cost advantage over RADAR, and non-intrusive nature compared to loops which require digging the road. We acknowledge the challenges posed to the camera by low illumination scenario in the night making it difficult to detect vehicles on the road. Here, we make an important observation and contribution that even in the night, the illumination at the road[1] is usually sufficient for a camera to identify the parts of the road occupied by the vehicles. This allows us to observe the traffic tailback at a carriageway. We show in this paper that merely knowing traffic tailback length (called 'TTL'

[1] Delhi Technological University, New Delhi
[2] Indian Institute of Information Technology Vadodara
[3] Indian Institute of Technology Delhi, New Delhi
* Research intern with Chetan Arora during the project.

[1]due to the illumination from vehicle head-lamps as well ambient light from the street-lamps

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 18,2021 at 09:52:55 UTC from IEEE Xplore. Restrictions apply.

hereon) is sufficient to learn the incoming traffic pattern, and adaptively adjust the phase switching in an ATSC system to achieve better throughput rates compared to the existing state of the art which makes use of explicit vehicle detection and individual vehicle speeds to decide the phase green timings.

We exploit the advances made in two separate technological domains to design our system. *Background Subtraction* is well studied area in computer vision. In our problem context we learn the background which includes static part of the road, and then compute the foreground representing portions of the road occupied by vehicles. The farthest part of the occupied region is returned as TTL. We use Gaussian Mixture Models [13], [14] to represent the background.

We use advances in artificial intelligence algorithmic techniques, and especially model free reinforcement learning (RL) to train our ATSC system for predicting phase switching time intervals. In the context of our problem, we propose to represent the state of our RL system as a sequence of TTL on all the carriageways, and the corresponding signal phase for last h time instances. Although our RL controller agent can be trained for non-cyclic traffic light switching, we have enforced the traffic lights to be cyclic in nature to ensure the traffic flow does not become chaotic. Therefore, unlike previous works [8], our controller has a much smaller action space - {Switch, Not Switch}. A Switch action sends the traffic light into the next phase of its cycle, whereas a Not Switch action retains the previous phase. As we show in the experiments, this simplifies the training as well as reduces the number of phase cycles an agent needs to observe before it is able to learn the patterns and starts performing optimally.

Contributions: The specific contributions of this work are as follows: (1) We propose to use Traffic Tailback Length (TTL) to learn traffic pattern at an intersection. We show that TTL can be computed by off the shelf, low cost RGB cameras, even in low illumination and adverse weather scenarios. (2) We suggest a reinforcement learning based smart traffic signal controller agent that uses TTL and predicts the phase switching sequence. We make several innovations in design of RL agent, especially for compact and effective state and action representation. (3) We conduct exhaustive experiments using our RL agent in a simulated environment, SUMO [4]. We have tested our agent in (a) changing traffic patterns (b) obstacles in a carriageway, and (c) various rewards such as sign, linear or square function of TTL. We show that in all the scenarios, our agent improves the state of the art in terms of average delay faced by a vehicle as well as overall throughput at the intersection.

II. Related Work

We categorize previous works for adaptive traffic signal control (ATSC) on the basis of information used for designing the RL state space.

One of the major challenges in designing an ATSC system via RL is the choice of state space, and reward function, due to their significant effect on the overall learning process. Touhbi et al. [9] have suggested normalized queue length, obtained by dividing the queue length with lane length, for generating the state representation. Abdulhai et al. [2] have used the more informative state space by including the queue lengths on the four approaches as well as the elapsed phase time. Liang et al. [5] use a different state space representation based on position and speed of vehicles at an intersection. The action space is defined by selecting every phase's duration in the next cycle. The authors have proposed Double Dueling Deep Q Network (3DQN) algorithm for training their model. Genders and Razavi [3] have proposed to divide the road surface in equal sized cells, and then generate a state representation composed of three vectors. the first vector indicates the presence of a vehicle in each cell, the second records the speed of the vehicle in each cell (0 if vehicle not present) and the third represents the current phase. The representation requires gathering information from expensive sensors and division into cells implicitly assumes lane driving.

Richter et al. [7] have used policy gradient methods for the training. However, they have assumed that all vehicles move at a uniform speed at all times. Further, their controller requires much more elaborate information to create state space, viz cycle duration, phase duration, as well as the information from the neighbouring intersection. Mohamad and Abdulhai [8] have used positions and speeds of the vehicles in the at the intersection as inputs in the form of an image like structure wherein if there is a vehicle on the street, the specific cell corresponding to that part of the street is filled with 1, otherwise 0. Zheng et al. [12] proposed a novel designed called FRAP that models the problem as a competition between pairs of traffic light phases. Through such competition modeling, they claim to achieve invariance to symmetrical cases such as rotation in traffic flow. Zang et al. [11] introduced a novel framework called MetaLight that improved FRAP in traffic signal control by optimizing its model structure and updating paradigm. Xiong et al. [10] use demos of human-operated traffic light signals to train their agent. This avoids the exploration strategy that is required by reinforcement learning agents otherwise.

III. Proposed Method

A. Simulator

Testing a ATSC system directly in the real-world can be impractical and poses severe risks. To do a proof of concept, we have adopted a simulator based approach. This is inline with the current state of the art, which has also been tested on the simulators. For our simulations, we have used an open-source traffic simulation package 'Simulation of Urban MObility' (SUMO) [4]. However, the frames captured from the simulator do not match the perspective that is expected from the real-world CCTV cameras mounted on the intersections. Therefore, we apply warping and transformation on the simulator frames using OpenCV library [1] to match them with the frames captured from the real camera. This allowed us to test our algorithm on a realistic settings.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 18,2021 at 09:52:55 UTC from IEEE Xplore. Restrictions apply.

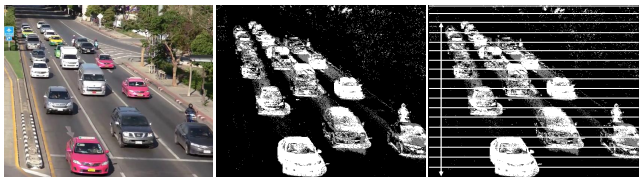


Fig. 2: Various stages of our TTL estimation technique.

B. Traffic Tailback Length (TTL)

For separating the foreground from the background, we use a Gaussian Mixture Model (GMM) based Background/Foreground Segmentation Algorithm [13], [14]. Figure 2 shows a result from our background subtraction module on an image taken from a traffic intersection. Note that the background subtraction step helps us to find the foreground pixels, irrespective of the vehicle type, and do not assume any lane driving assumption. The module is also robust against varying illumination and poor visibility due to adverse weather conditions.

To estimate TTL from an incoming frame, we scan the corresponding foreground mask, row by row. For each row we count the number of pixels marked foreground. The row is declared foreground if the count is beyond a threshold. We start counting from the bottom row and move upwards one at a time. The process is stopped when we receive a certain number of consecutive rows which are not foreground. The number of foreground rows is then returned as the estimated TTL. Figure 2 shows an example result.

It may be noted that even when we use a simulated environment, we warp and transform the frames generated by the simulator to visually match the frames captured from a real camera on an intersection. We use our background subtraction module on these frames to compute the TTL. These computed TTL values are then used in the RL agent for a fair comparison with other state of the art.

C. Reinforcement Learning

The goal of a reinforcement learning agent is to observe the environment *state* (in our case TTL in each of the 4 carriageways at the intersection), and then learn to predict an *action* (to switch a given phase or not), which maximize the long term expected *reward* (we explore various rewards). We use *Q-learning* algorithm [6] to train our agent.

The *Q-learning* starts by defining an *action-value* for each action in a particular state, denoted as $Q(s, a)$. The action value, $Q(s, a)$, represents the expected reward in the long term if an action a is chosen given an environment state s. The function $Q(s, a)$ is also known as *Q-function*. Formally, the *policy* of an RL agent, $\pi(s, a)$, is defined as the probability of taking an action, a, given a state, s,. Naturally, $Q(s, a)$ is dependent upon the policy, which governs what actions will be subsequently taken. Mathematically:

$$Q^{\pi}(s_t, a_t) = \mathbb{E}\big[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ...|s_t, a_t\big],$$

where R_{t+1} is the reward received at time $t + 1$, and γ is known as *discount factor* for the future rewards.

For scenarios closer to real-world, the state space and action space can be extremely large. So function approximation techniques are used to estimate the Q-function. When neural networks are used to estimate the Q-function, they are known as Deep Q-Networks (DQN) [6].

Initially, Q-value for each state-action pair is randomly initialized. As the agent gains experience by exploring various environment states and receiving rewards by choosing a particular action, it updates the Q-values by using well known *Bellman Equation* as follows:

$$Q(s, a) \leftarrow Q(s, a) + \alpha\big[R(s, a) + \gamma \max_{a'} Q(s', a') - Q(s, a)\big].$$

Here s' is the new state reached after taking action a in state s, and α is the step size analogous to learning rate in the machine learning systems. *TD-target* is defined as the sum of the current reward and the discounted maximum possible Q-value in the the next state for any action. The DQN training algorithm works by minimizing the *TD error*, defined as follows:

$$\text{TD Error} = R(s, a) + \gamma \max_{a'} Q(s', a', w) - Q(s, a, w),$$

where w are the parameters of a neural network to estimate the Q-function.

Target Network: We use the idea of fixed Q-targets [6], where a separate network, called the target network, is used for estimating the target values and after every n steps, the parameters of the DQN are copied to the target network. The updated TD error can be written as follows:

$$\text{TD Error} = R(s, a) + \gamma \max_{a'} Q(s', a', \bar{w}) - Q(s, a, w),$$

where $\bar{w}$ are the weights of target network. Use of target network significantly improves the DQN training.

Experience Replay: While training a neural network for supervised learning, the samples in a mini batch are uncorrelated. In RL, rewards are sequentially correlated that tend to bias a DQN towards focusing on the latest and forget the previous experiences. To address this one creates a *replay buffer* that stores experience tuples while the agent interacts with the environment. For training, a small batch of tuples is randomly sampled from the replay buffer for training the DQN. This method of creating a memory buffer and randomly sampling from it is called *experience replay*.

In detail, each tuple in replay buffer is of the form (s_t, a_t, r_t, s_{t+1}). Initially, the buffer is empty. The agent takes random actions till the experience buffer is of the size *init_replay_size*. A maximum size limit, *max_replay_size* is also specified for the buffer. Once the maximum size is reached, each new experience is added by removing the oldest entry from the buffer. For our experiments, we use initial size as 350 tuples and maximum size as 8000 tuples.

ϵ-greedy Policy: While training, we use ϵ-greedy policy [6] to ensure our agent explores the state space before converging to optimal policy. We specify an exploration rate, $\epsilon \in [0, 1]$, and at every time step, with probability $1 - \epsilon$, we *exploit* the current knowledge in the form of learnt Q-function, and take action based on the best Q-value. However,

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 18,2021 at 09:52:55 UTC from IEEE Xplore. Restrictions apply.

(a)

	W-E	S-N	E-W	N-S
t-1	0	1	0.7	0.4
t-2	0	1	0.7	0.3
t-3	0.1	0.9	0.4	0.3
t-4	0.4	0.4	0.40	0.2
t-5	0.7	0.8	0.5	0.2
t-6	0.9	0.8	0.4	0.1
t-7	1	0.8	0.4	0
t-8	1	0.4	0.3	0

(b)

W-E Phase				S-N Phase				E-W Phase				N-S Phase			
1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0

(c)

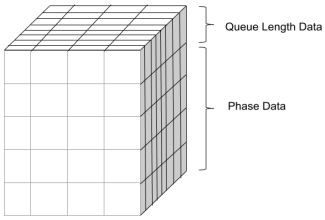


Fig. 3: Figure depicting the steps involved in state construction.

with probability ϵ, we *explore* by choosing a random action. In our experiments, we start with a value of 1 for the ϵ, and uniformly decrease it to 0.4 in 3000 steps. This ensures more exploration when the Q-values are uncertain, and more exploitation when we are confident on the learnt Q-values.

D. Proposed DQN Architecture

1) State Space: We record the state information every second, whereas, the granularity of an action in our formulation is of $k = 8$ seconds. In between two action instants, we maintain a history of k time instants for the state representation. The state representation consists of information on queue length ($\in [0, 1]$) in all the carriageways as well as the traffic light phase at the instant. To make it easy to process using standard convolutional neural network (CNN) architecture, we represent the state as a 3D volume of size $5 \times 4 \times k$, where each slice of 5×4 maintains the information at an instant. The first row in a slice contains the queue length in each carriageway. Rows 2 to 5 contains the phase information. Row 2 is filled with all 1 if the signal for WE carriageway is green, or with all 0, if this signal is red. Similarly Row 3, 4, and 5 represents signal phase for SN, EW, and NS carriageways respectively. Though, the information could have been encoded in a single scalar instead of a row of size 4, we use the particular structure to make it a uniform grid which helps in training a CNN architecture. Figure 3 shows the state representation pictorially.

2) Action Space: To avoid any kind of chaotic vehicle flow, we follow a cyclic order of traffic light phases. We propose using two actions {Switch, Not Switch} where the Switch action changes the phase to the next phase of the cycle, and Not Switch keeps the phase unchanged. A maximum limit of consecutive Not Switch actions is kept to ensure that the controller does not neglect any lane completely. This limit is set high enough so that it is practically unreachable for a trained controller.

3) Reward Function: We observe the queue lengths in various carriageways to find the two scores:

$$\mathcal{S}_a = q_{north} + q_{south} + q_{east} + q_{west}, \text{ and}$$
$$\mathcal{S}_s = q^2_{north} + q^2_{south} + q^2_{east} + q^2_{west},$$

where q_* denotes the TTL in each of the four carriageways (north, east, south, west). While both the scores serves to bias our agent towards maintaining smaller TTL, $\mathcal{S}_s$ additionally bias the agent towards maintaining relatively shorter TTL in each of the carriageways. This is because at larger values, the square function increases more rapidly than the linear. This biases the agent towards clearing carriageways with larger TTL first. Based upon the scores we propose following reward functions:

$$R_1 = \text{sign}(\mathcal{S}_a^t - \mathcal{S}_a^{t+1})$$
$$R_2 = \mathcal{S}_a^t - \mathcal{S}_a^{t+1} \quad (1)$$
$$R_3 = \mathcal{S}_s^t - \mathcal{S}_s^{t+1}.$$

Here S^t and S^{t+1} are the scores at time instants t and $t+1$ respectively, and sign$(\cdot)$ is the function which outputs 1 if difference of scores is positive and -1 otherwise. We analyze the performance of our RL agent with all the above reward functions in the experiment section.

4) Network Architecture: Our proposed architecture for the DQN is shown in Figure 4. The first layer in our network is a convolutional layer of size 5×1. The kernel essentially convolves all the information from a carriageway across all the time instants. We learn 10 such filters in the first layer. The output of the layer is thus 10 feature maps of size 1×4. We flatten the output here to make a single vector of size 40. This vector is then inputted to a fully connected layer of size 20. The third layer is the output layer with 2 neurons, giving the probability of the two actions. Algorithm 1 describes precise steps of the overall training algorithm.

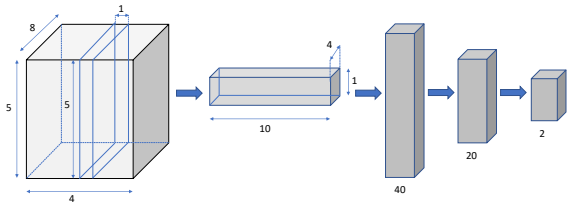


Fig. 4: Architecture of our Deep Q-network



Fig. 5: Comparison of our proposed rewards on a particular traffic pattern. Please see the text for details.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 18,2021 at 09:52:55 UTC from IEEE Xplore. Restrictions apply.

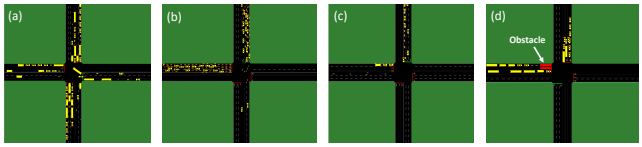


Fig. 6: Scenarios tested on a single intersection. In (a), equal traffic is supplied to all the carriageways. In (b), equal traffic is supplied to two of the carriageways, the other two carriageways are kept empty. In (c), the traffic in the two non-empty carriageways is made unequal. In (d), an obstacle (labeled in red) is introduced in one of the carriageways.

IV. Experiments

For comparison with round-robin algorithm, we use a controller that switches traffic lights after every 40 seconds, called Fixed Switching Time 40, or FST-40. We use [8] as representative of state of the art for comparing with our model. Since no implementation is publicly available, we have made our implementation of their algorithm as well. We use state space and reward function as suggested in [8], but kept our action space for easier comparison. We compare the methods on three metrics:

1) Average Delay: At every second, we compute the delay faced so far by the vehicles standing in all the carriageways. The average of all the delays is computed as average delay at that instant.
2) Queue Length: This is average of the queue lengths (TTL) at all the carriage-ways, computed every second.
3) Vehicles Passed: This is cumulative count of the vehicles which have passed the intersection till that time instant.

A. Comparing Proposed Reward Functions

In the proposed approach, we have explored three rewards to train our ATSC agent (Eq. 1). While R_1 and R_2 give linear reward with the decreasing queue lengths, R_3 gives as per the square of queue lengths. The purpose of R_3 is to penalize increasing longer queues more compared to equivalent increase in the shorter queues. Figure 5 compares the performance of our agent trained with the three rewards. The simulation is tested with traffic in NS carriageway 9 times that of EW carriageway. In all the three cases the agent takes some time to learn the traffic pattern, but after observing the traffic pattern for some time, the performance of the agent is similar on all the three rewards. As we show later in this section, the performance under various proposed rewards differs only when we create an obstacle in one of the carriageways. We have used R_2 as our reward function for subsequent testing with state of the art.

B. Comparison with State of the Art

The controller was compared in a variety of scenarios (Figure 6). In the initial scenario all the four inbound carriageways have equal traffic. This scenario is optimal for the round-robin algorithm. Our controller also learns the round-robin strategy with a shorter switching time and is able to match the performance of the round-robin algorithm in terms of vehicle throughput, while significantly reducing the cumulative average delay and queue length (Figure 7(a)).

Next, we removed traffic from two carriageways, leaving only two carriageways with equal and non-zero traffic. The controller adapts to this traffic configuration and allocates more time to occupied carriageways than empty carriageways, resulting in superior performance than FST 40. Our method also outperforms state of the art controller [8] which is only able to reduce the cumulative average queue lengths over FST 40 by a narrow margin (Fig. 7(b)).

To further increase complexity, traffic was modified to be unequal in the two carriageways with a ratio of 9:1. This ratio was made to exchange between the 2 carriageways (making it 1:9) at fixed time intervals. Three scenarios were created where this time interval for exchange was 10 minutes, 15 minutes and 30 minutes respectively. The performance was measured on cumulative average delay (Figure 7(c)) and vehicle throughput (Figure 7(d)). When the traffic is exchanged every 30 minutes, the state of the art is able to achieve a better vehicle throughput by a small margin at the cost of a significantly higher delay. This is because the state of the art preferred allocating the green signal almost completely to the denser carriageway. On the other hand, our controller makes sure to allocate green signal to both the carriageways resulting in better queue length management while maintaining a high enough throughput. For the other two scenarios where exchange in traffic happens more frequently at 10 and 15 minutes, our controller outperformed the state of the art and FST 40 in both the metrics. The state of the art performs better than FST 40 for some part of the episode and then worse than FST 40, because it is not able to adapt quickly enough to changing traffic resulting in sub-optimal performance in later stages of the episode.

Next, we stopped exchanging the traffic in the two carriageways and kept the traffic equal in them, while introducing an obstacle on one of the two carriageways, lowering its vehicle speeds. Our controller was compared with FST 40 to check whether it detects the inequality in carriageway vehicle speeds and behaves differently for obstacle carriageway. Figure 8(a) shows that it performs better than FST 40.

To observe the way our proposed controller handled this inequality, we did a carriageway-wise comparison in terms of green time allocated per cycle (Figure 8(c)) and average queue length (Figure 8(b)) observed in obstacle carriageway and non-obstacle (but non-empty) carriageway. We also analyzed the results for our agent trained with rewards R_1 and R_3 to see if the difference in reward function influences the behavior of our controller. Our controller with reward R_2 allocated a significant amount of green time to the non-obstacle carriageway to push the overall throughput higher and to minimize overall queue length by making a trade-off with the average delay time of vehicles on the obstacle carriageway. We also observed that R_1 focuses on queue lengths of both carriageways by allocating more green time to obstacle carriageway than R_2 but fails to find any significant improvements in the queue length of the carriageway with

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 18,2021 at 09:52:55 UTC from IEEE Xplore. Restrictions apply.

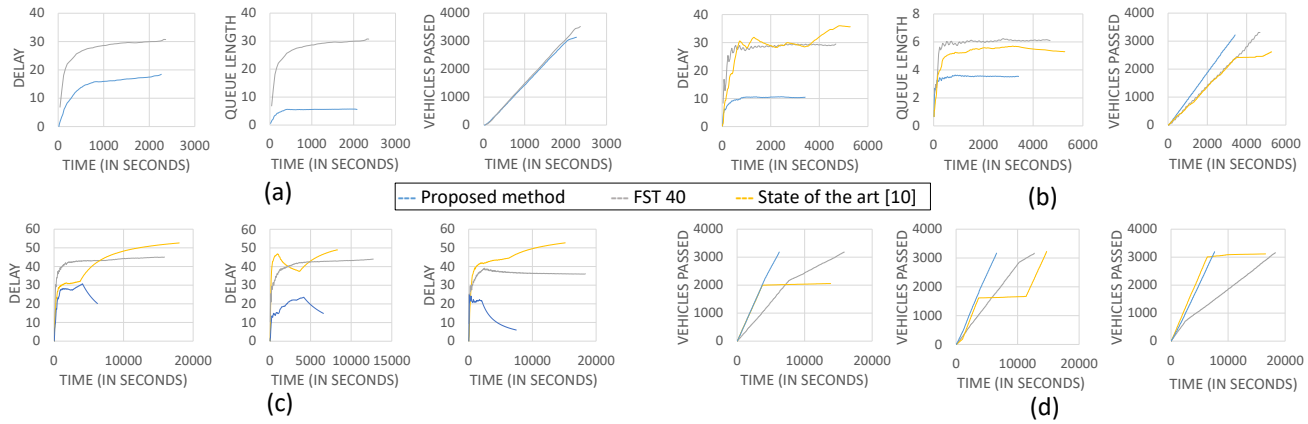


Fig. 7: The graphs depict the comparison of various controllers, including our proposed model, round robin controller having a switching time of 40 seconds (FST 40), and current state of the art model [8].

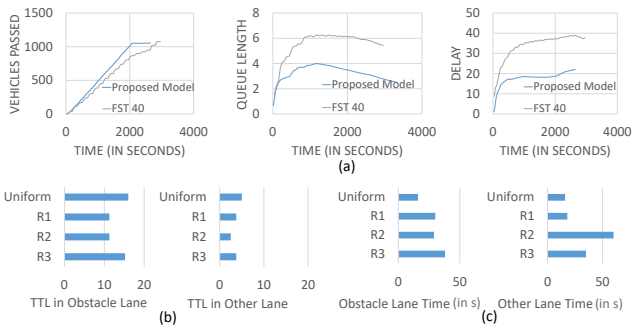


Fig. 8: In (a), the graphs depict the comparison of our proposed model against a round-robin controller (FST-40) in an obstacle scenario. In (b), the graphs depict the behavior of various reward functions.

obstacle because of its lower vehicle speeds. R_3 allocates equal green time to both the carriageways to prevent queue in carriageway with obstacle to grow too long.

V. Conclusion

In this paper we have presented a Reinforcement Learning based Adaptive Traffic Signal Controller which uses only the traffic tail back length at each carriageway to learn the incoming traffic patterns and adapt phase switching sequence for optimal throughput. We show that the traffic tail back length can be computed simply and robustly by background subtraction operation on the incoming video frames. The operation is robust to degraded visibility due to adverse weather or night time operation. Further the operation does not assume any lane driving or prior availability of vehicle types. We tested our agent on a variety of commonly occurring scenarios in developing world. In all our experiments the proposed system outperforms the state of the art. We believe our low cost system will make adaptive traffic control practical for use in the developing countries.

References

[1] Open Source Computer Vision Library (OpenCV) Version 3.4.2. https://github.com/opencv/opencv, Last Checked: May 2020.

[2] Baher Abdulhai, Rob Pringle, and Grigoris J. Karakoulas. Reinforcement learning for true adaptive traffic signal control. *Journal of Transportation Engineering*, 129(3):278–285, 2003.

[3] Wade Genders and Saiedeh Razavi. Using a deep reinforcement learning agent for traffic signal control. *CoRR*, abs/1611.01142, 2016.

[4] Daniel Krajzewicz, Jakob Erdmann, Michael Behrisch, and Laura Bieker. Recent development and applications of SUMO - Simulation of Urban MObility. *International Journal On Advances in Systems and Measurements*, 5(3&4):128–138, December 2012.

[5] Xiaoyuan Liang, Xunsheng Du, Guiling Wang, and Zhu Han. Deep reinforcement learning for traffic light control in vehicular networks. *CoRR*, abs/1803.11115, 2018.

[6] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, February 2015.

[7] Silvia Richter, Douglas Aberdeen, and Jin Yu. Natural actor-critic for road traffic optimisation. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1169–1176. MIT Press, 2007.

[8] S. M. A. Shabestary and B. Abdulhai. Deep learning vs. discrete reinforcement learning for adaptive traffic signal control. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 286–293, Nov 2018.

[9] Saad Touhbi, Momahed Babram, Tri Nguyen-Huu, Nicolas Marilleau, Hbid Moulay Lhassan, Christophe Cambier, and Serge Stinckwich. Adaptive traffic signal control : Exploring reward definition for reinforcement learning. *Procedia Computer Science*, 109:513–520, 06 2017.

[10] Yuanhao Xiong, Guanjie Zheng, Kai Xu, and Zhenhui Li. Learning traffic signal control from demonstrations. pages 2289–2292, 11 2019.

[11] Xinshi Zang, Nan Xu, Huaxiu Yao, Guanjie Zheng, Kai Xu, and Zhenhui Li. Metalight: Value-based meta-reinforcement learning for online universal traffic signal control, 2020.

[12] Guanjie Zheng, Yuanhao Xiong, Xinshi Zang, Jie Feng, Hua Wei, Huichu Zhang, Yong Li, Kai Xu, and Zhenhui Li. Learning phase competition for traffic signal control, 2019.

[13] Z. Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 2, pages 28–31 Vol.2, Aug 2004.

[14] Zoran Zivkovic and Ferdinand van der Heijden. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recogn. Lett.*, 27(7):773–780, May 2006.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 18,2021 at 09:52:55 UTC from IEEE Xplore. Restrictions apply.

2020 IEEE International Conference for Innovation in Technology (INOCON)
Bengaluru, India. Nov 6-8, 2020

Sensorless control of PMSM using Direct and Model References Adaptive system estimation techniques and observing the effect of using PI Fuzzy logic controller in MRAS Model

Mohamed Abdallah Anwar Said,
M.Tech Student, Delhi Technological University, Delhi, India
Maas.volt@gmail.com

Mini Sreejeth,
Professor, Delhi Technological University, Delhi, India
minisreejeth@dce.ac.in

A.B.K.Bhattacharya,
Professor, Delhi Technological University, Delhi, India
a.b.bhattacharya@dce.edu

***Abstract*-As per the extremely high competitiveness among modern PMSM motor and drive manufacturers for attracting more customers to buy their products, a huge load has been added upon engineers for developing better products with better techniques. One of the most competitive aspects for adding attraction to the PMSM is the existence of a sensorless estimation technique option in the Drive system replacing the ordinary encoder based one. This is not considered to be a complete replacement of encoder or and speed sensor, as having a direct and clear signal remains as the best method for obtaining the best motor performance. But as for many reasons and under many conditions any failure could happen to the encoder, which could put the motor completely out of service in case no other speed and position estimation method exists. Thanks to modern engineers and mathematicians, many algorithms have been developed to overcome this problem. Using pure mathematical based estimations, PMSM's both speed and position can be estimated and implementing those estimation techniques to the PMSM Drive soft and hardware add a competitive value to those motors as they become more reliable and durable.**

Keywords: Sensorless control, estimator,MRAS,observer

I. INTRODUCTION

Designing a sensor-less feedback approach became a wide field of research as many algorithms and approaches have been introduced. Some approaches depend on some direct calculations others depend on using a non-linear algorithms for estimation. The model used for direct estimation is called basic estimation technique, and the algorithm depends on the switching logic of the VSI. The model reference adaptive system (MRAS) is one of the promising means in the field of sensorless control as it is considered to be flexible in dealing with the nonlinearity. By implementing both direct and MRAS techniques over the same system a better understanding of systems behavior will be generated and it will be easier to compare between the effectiveness of each estimation technique from the aspects of performance and tuning flexibility.

In the design of both techniques PI controller is widely used for eliminating the error. The main PI controller used in the speed PI controller ensures that motor speed tracks the reference speed smoothly. PI is not the only speed controlling technique that can be used, many other techniques can be used such as Fuzzy logic controllers and Artificial intelligence based controllers. Fuzzy logic controller commonly is not used alone while designing the controller, but a mixed Fuzzy and PI technique is used for better and faster performance. This PI Fuzzy technique will be discussed while being implemented in MRAS estimator based model in the subsequent sections.

II. MODELING OF PMSM AND PMSM DRIVE

A dynamic model of PMSM motor can be designed mathematically after taking the following assumptions into consideration.

1) Saturation neglected

2) EMF is sinusoidal,

3) Losses are neglected

4) No field current dynamics.

The stator voltage equations are:

$$V_q^r = r_s\, i_q^r + \omega_r \lambda_d^r + \frac{d\lambda_q^r}{dt} \tag{1}$$

$$V_d^r = r_s\, i_d^r - \omega_r\, \lambda_q^r + \frac{d\lambda_d^r}{dt} \tag{2}$$

The flux linkage is:

$$\lambda_q^r = L_q i_q^r \tag{3}$$

$$\lambda_d^r = L_d i_d^r + \lambda_m \tag{4}$$

Thus the stator voltage equations can become

$$V_q^r = r_s\, i_q^r + L_q \frac{di_q^r}{dt} + \omega_r\, L_d i_d^r + \omega_r \lambda_m \tag{5}$$

$$V_d^r = r_s\, i_d^r + L_d \frac{di_d^r}{dt} - \omega_r\, L_q i_q^r \tag{6}$$

Where,
r_s = stator resistance

L_d = d-axis inductance

L_q = q-axis inductance

ω_r = rotor rotational speed

λ_m= permanent magnet flux.

Electromagnetic torque can be expressed as

$$T_{em} = \frac{3}{2}\frac{P}{2}\left(\lambda_d^r i_q^r - \lambda_q^r i_d^r\right) \tag{7}$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:01:46 UTC from IEEE Xplore. Restrictions apply.

And by appropriate substitution of equations (3) and (4) in (7) it becomes.

$$T_{em} = \frac{3}{2}\frac{P}{2}\left[\lambda_m i_q^r + (L_d - L_q) i_q^r i_d^r\right] \quad (8)$$

The torque equation can also be expressed as

$$T_e = T_m + J\frac{d}{dt}\omega_r + B\,\omega_r \quad (9)$$

Using the above equations the Simulink model for the PMSM is built in MATLAB as shown in Fig.1.

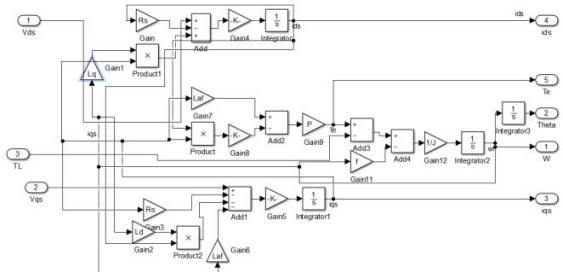


Fig. 1. Simulink Model of PMSM motor

The simulation is carried out on a 1 kW PMSM drive whose technical specification is given at Table 1

TABLE I. PMSM PARAMETERS

Rated Output (kW)	1.0
Rated Torque (Nm)	3.18
Maximum Torque (Nm:)	9.54
Rated Speed (RPM)	3000
Maximum Speed (RPM)	5000
Rated Current (Amps)	7.3
Maximum Current (Amps)	21.9
Rotor Moment of Inertia (Kg.m2 x 10-4)	2.65
Mechanical Time Constant (msec)	0.74
Torque Constant - KT (Nm/A)	0.44
Voltage Constant - KE (mV/rpm)	16.8
Armature Resistance (Ohm)	0.20
Armature Inductance (mH)	2.0
Electrical Time Constant (msec)	10.26
Maximum Radial Shaft Load (Newton)	490
Maximum Thrust Shaft Load (Newton)	98

Field oriented control is the common approach used while designing a drive model with a sensor-less control. Fig.2 represents the basic FOC drive model structure used in all the following studies of sensorless estimators.

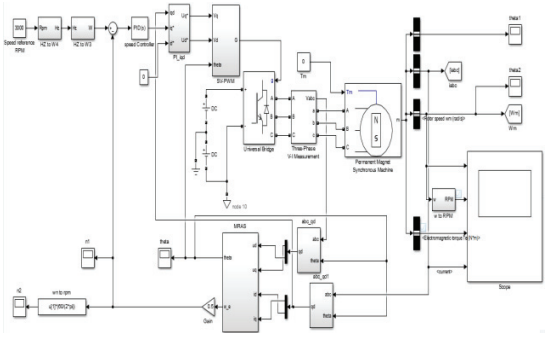


Fig. 2. Simulink model of PMSM Drive module

III. BASIC ESTIMATION TECHNIQUE

Equations (10) and (11) describe the algorithm used for building the direct estimator.

Calculating both v_α and v_β

$$v_\alpha = \frac{v_{dc}}{3} * (2\,s_{I1} - s_{I3} - s_{I5}) \quad (10)$$

$$v_\beta = \frac{v_{bat}}{3} * (\sqrt{3}(s_{I3} - s_{I5})) \quad (11)$$

Where,

s_{I1}, s_{I2} and s_{I3} = switching logic (set either of the upper three legs of VSI)

i_α , i_β Values are calculated by

$$i_\alpha = i_a \quad (12)$$

$$i_\beta = \frac{1}{\sqrt{3}}(i_b - i_c) \quad (13)$$

Calculating the stationary components of flux Ψ_α , Ψ_β

$$\Psi_\alpha = \int (v_\alpha - R_s * i_\alpha)dt + \psi_{\alpha 0} \quad (14)$$

$$\psi_\beta = \int (v_\beta - R_s * i_\beta)dt + \psi_{\beta 0} \quad (15)$$

The resultant flux linkage ψ_s can be obtained by

$$\psi_s = \sqrt{\psi_{alpha}^2 + \psi_{beta}^2} \quad (16)$$

the estimated rotor position θ_{est} can be expressed by

$$\theta_{est} = \tan^{-1}\left[\frac{\psi_{beta}}{\psi_{alpha}}\right] \quad (17)$$

The estimated rotor speed $\omega_{m\,est}$ speed can be expressed by

$$\omega_{m\,est} = \frac{1}{p}\frac{d}{dt}(\theta_{est}) \quad (18)$$

Each of these equations is implemented in Simulink as a single block for each and by assembling them, direct estimator is built.

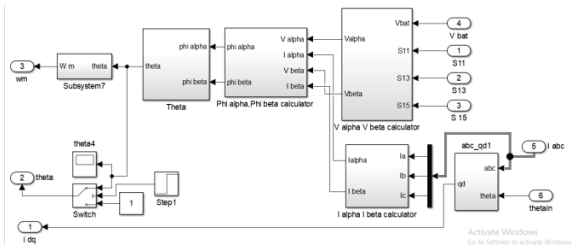


Fig. 3. Basic Estimator Design

IV. MODEL REFERENCE ADAPTIVE SYSTEM (MRAS)

The basics of MRAS technique is represented by two independent models:

1) Reference model which is independent on the variable to be estimated.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:01:46 UTC from IEEE Xplore. Restrictions apply.

2) Adjustable model which is dependent on the variable to be estimated.

The two models use different sets of inputs to calculate the same state variables which are in turn fed to a certain adaption mechanism. The adaption mechanism uses the difference between the two signals to tune the estimated variable and feed it back to the adjustable model. The estimated value will this way be driven to its true value. The MRAS Scheme is illustrated as shown in Fig.4.

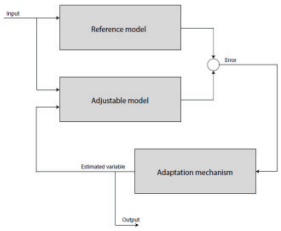


Fig. 4. Block diagram of Model Reference Adaptive System

When using MRAS to estimate the rotor speed and position of a PMSM, a motor variable, e.g. stator current or power, should be chosen as the state variable.

Two different models are then used: The reference model calculates the state variable without using the rotor speed or position and the adjustable model calculates the same state variable with the help of the rotor speed or position value given by the adaptive mechanism.

The following figure Fig.5 represents the structure of MRAS algorithm.

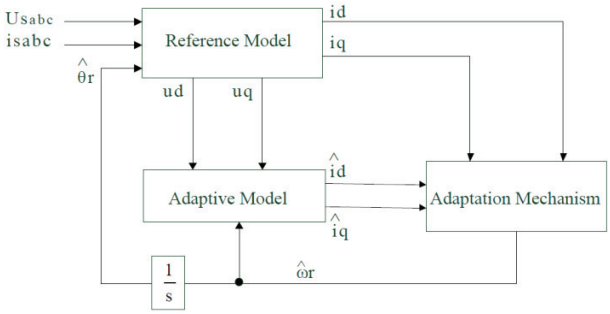


Fig. 5. Block diagram of sensorless of MRAS for PMSM Drive

The following equations represent the algorithms used for building the MRAS estimator.

$$V_d = R\, i_d + L_d \frac{di_d}{dt} - \omega_r L_q i_q \tag{19}$$

$$V_q = R\, i_q + L_q \frac{di_q}{dt} - \omega_r L_d i_d + \omega_r \psi_r \tag{20}$$

Where

V_d, V_q = stator voltage component in d-q frame of axes.

i_d, i_q = stator current component.

L_d, L_q = stator inductance.

R = stator resistance.

ψ_r = rotor flux.

ω_r = rotor speed.

$$\frac{di_d}{dt} = -\frac{R}{L_d} i_d + \frac{\omega_r L_q}{L_d} i_q + V_d \frac{1}{L_d} \tag{21}$$

$$\frac{di_q}{dt} = -\frac{\omega_r L_d}{L_q} i_d - \frac{R}{L_q} i_q + V_q \frac{1}{L_q} - \frac{\omega_r \psi_r}{L_q} \tag{22}$$

Both equations (21),(22) can be represented in matrix form as;

$$\frac{d}{dt}\begin{bmatrix} i_d \\ i_q \end{bmatrix} = \begin{bmatrix} -\frac{R}{L_d} & \omega_r \\ -\omega_r & -\frac{R}{L_q} \end{bmatrix} \begin{bmatrix} i_d \\ i_q \end{bmatrix} + \begin{bmatrix} \frac{1}{Ld} & 0 \\ 0 & \frac{1}{Lq} \end{bmatrix} \begin{bmatrix} V_d \\ V_q \end{bmatrix} \tag{23}$$

This equation represents the reference model

$$pi' = Ai + Bu' \tag{24}$$

The estimated values can be represented by the following model (the adjusted model)

$$\frac{d}{dt}\begin{bmatrix} \hat{i_d} \\ \hat{i_q} \end{bmatrix} = \begin{bmatrix} -\frac{R}{L_d} & \hat{\omega_r} \\ -\hat{\omega_r} & -\frac{R}{L_q} \end{bmatrix} \begin{bmatrix} \hat{i_d} \\ \hat{i_q} \end{bmatrix} + \begin{bmatrix} \frac{1}{Ld} & 0 \\ 0 & \frac{1}{Lq} \end{bmatrix} \begin{bmatrix} V_d \\ V_q \end{bmatrix} \tag{25}$$

Which, can be simplified into

$$p\hat{i}' = A\hat{i} + Bu' \tag{26}$$

Defining the generalized error as

$$e = i' - \hat{i}' \tag{27}$$

Where

$$e_d = i'_d - \hat{i_d}' \tag{28}$$

And

$$e_q = i'_q - \hat{i_q}' \tag{29}$$

From this we can get that

$$\begin{bmatrix} \frac{de_d}{dt} \\ \frac{de_q}{dt} \end{bmatrix} = \begin{bmatrix} -\frac{R}{L_d} & \omega_r \\ -\omega_r & -\frac{R}{L_q} \end{bmatrix} \begin{bmatrix} e_d \\ e_q \end{bmatrix} - J\,(\omega_r - \hat{\omega_r}) \begin{bmatrix} \hat{i_d}' \\ \hat{i_q}' \end{bmatrix} \tag{30}$$

Where

$$J = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \tag{31}$$

And this equation can be simplified as

$$\frac{de}{dt} = A_g - W \tag{32}$$

Where

$$A_g = \begin{bmatrix} -\frac{R}{L_d} & \omega_r \\ -\omega_r & -\frac{R}{L_q} \end{bmatrix} \tag{33}$$

And

$$W = (\hat{\omega_r} - \omega_r).J.\hat{i_s}' \tag{34}$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:01:46 UTC from IEEE Xplore. Restrictions apply.

Adaptive law can be obtained by inverse solution of Popov integral inequality.

$$\omega_r^{\wedge} = \left(K_p + \frac{K_i}{s}\right).i_s * i_s^{\wedge} \quad (35)$$

$$\omega_r^{\wedge} = \left(K_p + \frac{K_i}{s}\right)\left(i_d'\, i_q^{\wedge\prime} - \; i_q'\, i_d^{\wedge\prime}\right) i_q^{\wedge\prime} \quad (36)$$

$$\omega_r^{\wedge} = \left(K_p + \frac{K_i}{s}\right).\left(i_d\, i_q^{\wedge} - i_q i_d^{\wedge} - \frac{\psi_r}{l}.\left(i_q - i_q^{\wedge}\right)\right. \quad (37)$$

Where $i_d^{\wedge}, i_q^{\wedge}$ are obtained from the adjusted model, and i_d, i_q are obtained model.

The PI controller in the adaption mechanism allows the estimator to perform in satisfactory way, and it helps in increasing the operation range.

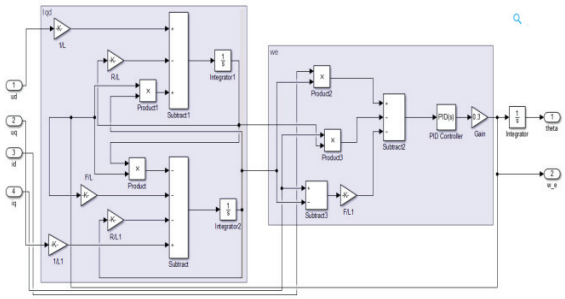


Fig. 6. Simulink model of MRAS module

Fig.6 represents the MRAS module by Simulink. This module is completely dependent on the pre mentioned MRAS equations. By integrating this Module with a sensorless model for the drive and by tuning the PI parameters perfectly, an acceptable performance can be achieved.

V. PI - TUNING

The PMSM Drive contains four PI controllers as follows:

- MRAS model PI controller
- Speed PI controller
- Vq PI controller
- Vd PI controller

The existing PI units are showed as following



Fig. 7. MRAS model PI controller

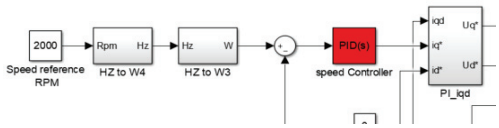


Fig. 8. Speed PI Controller

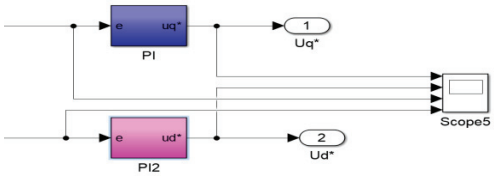


Fig. 9. Vq,Vd PI controllers

The PI tuning parameters are as per following.

TABLE II. PI MODEL VALUES

	P	I
MRAS PI controller	1	0.02
Speed PI controller	0.3	10
Iq PI controller	200	20000
Id PI controller	200	20000

The reason of not using the PID instead of PI is that the D component adds instability to the system.

VI. INTEGRATED PI FUZZY CONTROLLER

For the speed controller of PMSM a regular Fuzzy is not used, instead an integration between fuzzy and PI controller are used for achieving a good performance.

Fig.10 represents the interior mechanism of the fuzzy logic controller.

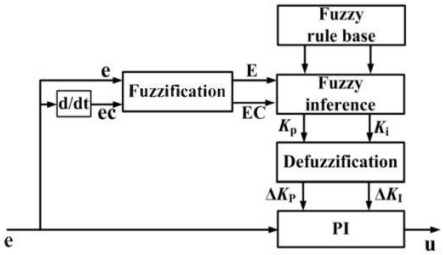


Fig. 10. Fuzzy logic Controller for PMSM

The structure of the PMSM Fuzzy speed controller is described as per following.

1) Error input which is the drift value between the reference speed and the estimated speed
2) Error and error change (e, ec) and they are the input to the fuzzification box
3) Fuzzification box and inside it E and EC are fuzzy variables of e and ec, Kp and Ki are fuzzy variables of $\Delta\, Kp\ and\ \Delta K$i,. The range of inputs and outputs are given. E and EC= {-3, -2.5, -2, -1.5,-1,-0.5, 0, 0.5, 1, 1.5, 2, 2.5, 3}, Δ Kp and ΔKi, = {-0.3, -0.25, -0.2,- 0.15,-0.1,-0.05, 0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3}. The quantum factors Ke and Kc are used for transforming the inputs into the fuzzy variables E and EC, respectively. They can be described as

$$E = e * K_e \quad (38)$$

$$EC = ec * K_C \quad (39)$$

4) Fuzzy rule base and the fuzzy logic rules of both Kp and Ki are represented by the following tables while Fig. 20

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:01:46 UTC from IEEE Xplore. Restrictions apply.

represents the membership functions governing both the input and the output

TABLE III. FUZZY LOGIC RULES OF KP

		EC						
		NB	**NM**	**NS**	**ZO**	**PS**	**PM**	**PB**
E	**NB**	PB	PB	PM	PM	PS	ZO	ZO
	NM	PB	PB	PM	PS	PS	ZO	NS
	NS	PM	PM	PM	PS	ZO	NS	NS
	ZO	PM	PM	PS	ZO	NS	NM	NM
	PS	PS	PS	ZO	NS	NS	NM	NM
	PM	PS	ZO	NS	NM	NM	NM	NB
	PB	ZO	ZO	NM	NM	NM	NB	NB

TABLE IV. FUZZY LOGIC RULES OF KI

		EC						
		NB	**NM**	**NS**	**ZO**	**PS**	**PM**	**PB**
E	**NB**	NB	NB	NM	NM	NS	ZO	ZO
	NM	NB	NB	NM	NS	NS	ZO	ZO
	NS	NB	NM	NS	NS	ZO	PS	PS
	ZO	NM	NM	NS	ZO	PS	PM	PM
	PS	NM	NS	ZO	PS	PS	PM	PM
	PM	ZO	ZO	PS	PS	PM	PB	PB
	PB	ZO	ZO	PS	PM	PM	PB	PB

5) Defuzzification and the max-min composition centroid method are employed in defuzzification. According to this rule, Kp and Ki are calculated, but they are fuzzy valves. The scaling factors y_1 and y_2 are used for transforming them into accurate values (Δ Kp and ΔKi), as given by

$$\Delta K_p = K_p * y_1 \quad (40)$$

$$\Delta K_i = K_i * y_2 \quad (41)$$

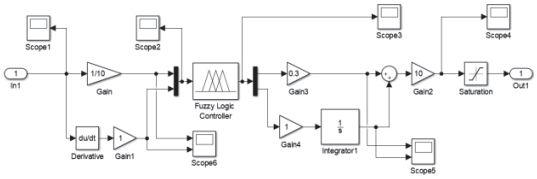


Fig. 11. Fuzzy speed controller Structure

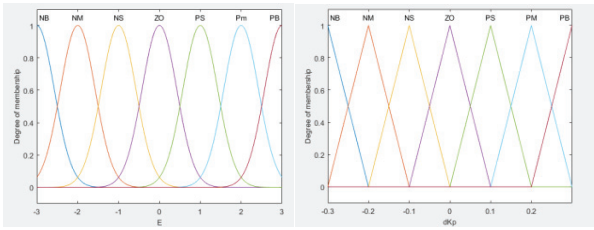


Fig. 12. Fuzzy controller membership functions

VII. PERFORMANCE EVALUATION

Using MATLAB Simulink basic estimator, MRAS estimator and fuzzy controller are tested and the following set of figures describe the performance of each of those models.

Fig.13 defines the speed performance of the basic estimator at variable torque and 2000 RPM reference speed. It is observed that even with sudden torque change, the observer manages to track the reference speed, maintain the output speed equal to the reference speed 2000 RPM.

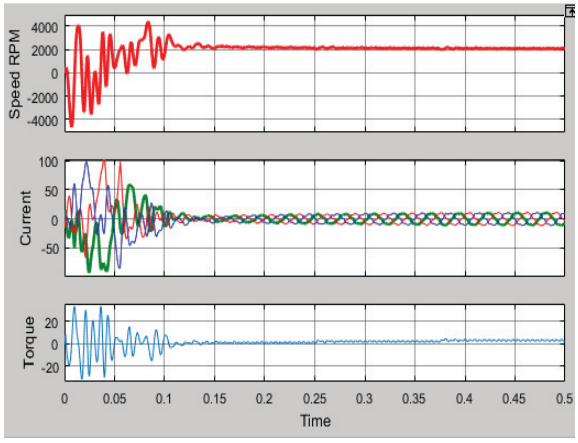


Fig. 13. Dynamic performance of PMSM at Variable torque (0, 1, 2, 3) Nm and 2000 reference speed (Using Basic estimator)

Fig.14 defines the speed performance of the MRAS estimator at variable torque and 1000 RPM reference, the MRAS performance is observed at sudden torque changing (0.125 sec between each new torque change). Current changing also gives a realistic behavior as it is logic to increase with torque increase. And the speed maintains reaching the set point (1000 RPM) after any torque change.

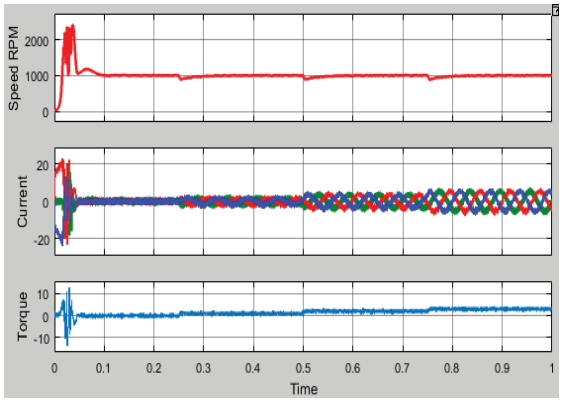


Fig. 14. Dynamic performance of PMSM at variable torque and 1000 RPM reference speed (Using MRAS Estimator)

In Fig.15 using a repeating sequence stair dynamic change performance over Time can be checked as per following.

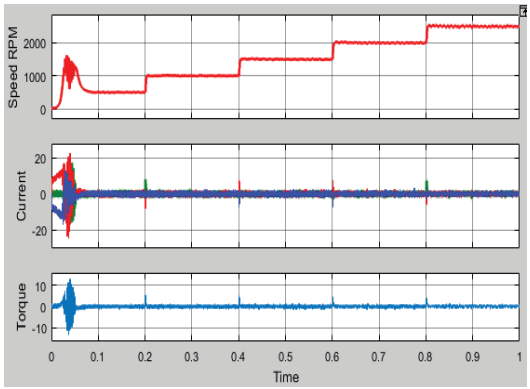


Fig. 15. Dynamic performance of PMSM for speed change (500, 1000, 1500, 2000, 2500) RPM respectively (Using MRAS Estimator)

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:01:46 UTC from IEEE Xplore. Restrictions apply.

Fig.16 represents the PMSM at reference speed (2000) and zero torque, while a fuzzy logic speed controller is used. It is clear how the controller is trying to keep up with the reference speed but the output still not perfect as output speed keeps oscillating slightly over the reference speed maintain a higher speed.

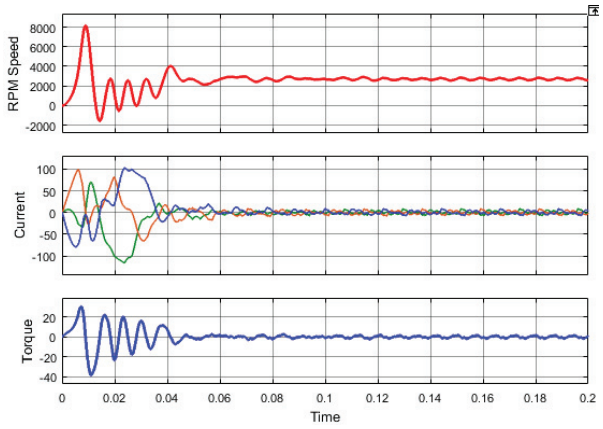


Fig. 16. Dynamic performance of PMSM for 2000 RPM reference (Using FUZZY Controller)

VIII. CONCLUSION

MRAS estimator proves to be a very powerful observer for performing a sensorless control over PMSM. Even if basic estimator is getting the desired output but the overshoot in the transient region makes it inappropriate for practical applications. This problem of course can be solved if soft started sis unused or if enough time will be given to the drive to get from zero speed till maximum speed. Fuzzy logic controller also is considered to be a promising technique, and regardless to its week performance but with more enhancement and further research this technique might be developed more. Researches also are conducting in integrating fuzzy with ANN technique for getting better performance but such technique still wide complicated to be applied.

REFERENCES

[1] Hui Deng, Guang-Zhong Cao, Su-Dan Huang, Lai-Juan Shi, Zhi-Ming He, A Sensorless Vector Strategy for the PMSM Using Improved Sliding Mode Observer and Fuzzy PI Speed Controller, IEEE Trans.

[2] PMSM motor is being used widely now in all industrial aspects regarding to the high.

[3] Hyun Lee and Jangmyung Lee," Design of Iterative Sliding Mode Observer for Sensorless PMSM Control, IEEE TRANSACTIONS ON CONTROL SYSTEMS TECHNOLOGY, VOL. 21, NO. 4, JULY 2013.

[4] Saeed Masoumi Kazraji, Ramin Bavili Soflayi, Mohammad Bagher Bannae Sharifian," Sliding-Mode Observer for Speed and Position Sensorless Control of Linear-PMSM" doi: 10.2478/ecce-2014-0003.

[5] Mustafa DURSUN1* Ali Fuat BOZ2 Murat KALE1 Murat KARABACAK2 "SENSORLESS SPEED CONTROL OF PERMANENT MAGNET SYNCHRONOUS MOTOR WITH HYBRID SPEED CONTROLLER USING MODEL REFERENCE ADAPTIVE SYSTEM" Vol 3, No 1, 24-37, 2014.

[6] Gomaa F. Abdelnaby, Tarek A. Dakrory, Shawky H. Arafa† and Salah G. Ramdan :Sensorless Field Oriented Control of Permanent Magnet Synchronous Motor".

[7] SENSORLESS CONTROL OF PMSM USING DSPACE CONTROLLER by C.Sindhura1, K.Victor Sam Moses Babu2 1Department of EEE, Malla Reddy Engineering College for women, Hyderabad, India.

[8] PERMANENT MAGNETS SYNCHRONOUS MOTOR CONTROL THEORY by Marek ˇStulrajter, Valeria Hrabovcova and Marek Franko.

[9] T. T. Liu†, Z. G. Yan†, K. Chen†*, & S. G. Li‡" Fuzzy-PI Control Strategy for PMSM Used in Electric Vehicles" Journal of Mechanical Engineering Research and Developments, SSN: 1024-1752.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:01:46 UTC from IEEE Xplore. Restrictions apply.

2020 IEEE International Conference for Innovation in Technology (INOCON)
Bengaluru, India. Nov 6-8, 2020

Sentiment Analysis of Emoticon Based Neuro Fuzzy System

Ritu Agarwal
Department of Information Technology
Delhi Technological University
Delhi, India
ritu.jeea@gmail.com

Sourabh Shreshth
Department of Information Technology
Delhi Technological University
Delhi, India
sourabhshreshth@gmail.com

Abstract- **The widespread Virtual societal space has become an integral part of human life, in which opinion generated by others plays a significant role in one's actions. Tweets as a point of the source have emerged as a way of expressing users' views on a particular subject. Historically sentiment analysis of tweets focused on polarized views, but in the real world, the opinion on a topic usually quite diverse. Therefore, using the classical classification methods cannot adequately fit the human sentiment, and we require a robust way that can match the broad scale of opinion-making.**

Present work describes an approach that can fit the opinion-making scale of human sentiment by using a fuzzy neuro system to train the model for opinion-making. Emoticon used in tweets provides an additional layer with classical textual analysis to incorporate complex emotions that can be easily conveyed by emoticons. The scored sentiment is then classified using various machine learning algorithms to measure its accuracy.

Keywords- Neuro fuzzy system, Sentiment analysis, Accuracy.

I. INTRODUCTION

The emergence of social media brought a significant change in the way we build an opinion about a subject matter. With current 4.2 billion internet users around the world, approximately 3.03 billion users are active on social media[1]. Therefore building a method to model opinion-making of social media users can play a huge role in various aspects like generating a recommendation system, targeted ad delivery and many more. In this work, we are going to use tweets from twitter as our source of information for building our classification model. Twitter is a social media giant that allows users to post and read short messages, called "tweets". Each tweet is 140 characters in length, with 500 million Tweets are generated every day by 240+ million active users[1]. Users often discuss current affairs and share an opinion on various subjects. The tweet is a great tool to express emotions' on a particular subject.

However, the biggest problem in sentiment analysis of tweet is to determine its polarity by using various textual analysis methods. Various classical methods of classification can determine the tweets as positive, negative and neutral to an extent, but they fail profoundly in front of the vagueness of human mind opinions. Tweets containing a mix of negative and positive words can be wrongly classified as neutral sentiment when we only use textual information in tweets. Hence, only using textual analysis of tweet for sentiment generation of a tweet is not sufficient to understand the real opinion of tweet's author.

Therefore before we generate a classification model using a various classification algorithm, we incorporate emoticons along with textual information of tweets for generating the data for our model. Emoticons are the facial expression of the author in a digital form representing expression like love, joy, surprise, anger, sadness, fear and more[2]. Emoticon helps in better representation of the sarcastic expression. Once we have the right set of input data, we can apply various classification algorithm with a fuzzy neuro system to generate the model and compare the result with existing algorithm.

II. BACKGROUND

A. Neuro Fuzzy System

A Neuro-Fuzzy system is a hybrid network based on the combination of fuzzy logic and neural network having a unique structure and lack of any feedback mechanism comprising of weights, non-fuzzy signal and activation functions[3]. The main idea behind the modelling of such a system is to determine the membership function parameters that can fit a fuzzy inference system. For this purpose, the neural network procedures are used to obtain the desired parameters. Therefore the function of neuro-fuzzy systems is to obtain the parameters of fuzzy systems by training. There are three popular fuzzy rule-based systems, namely Mamdani, Sugeno, Tsukamoto and more. In this paper we are using ANFIS system which based on the Sugeno fuzzy inference system designed in the structure of a five-layer feed forward neural network.

B. Adaptive Neuro Fuzzy Inference System (ANFIS)

Adaptive Fuzzy Inference System (ANFIS) based on the network is a neuro fuzzy based classification technique which is a combination of fuzzy system and neural network[4]. In ANFIS, the antecedent and the consequent parameters is adapted in such a way that the ANFIS architecture reflects both the Sugeno and the Tsukamoto fuzzy models. ANFIS approach is a combination of a fuzzy system and neural network where the primary role of the fuzzy network is to handle the uncertainty of system architecture in consideration to reduce error. Also, the neural network provides s the adaption strength to the parameters in use. Using this hybrid

978-1-7281-9744-9/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:04:48 UTC from IEEE Xplore. Restrictions apply.

method, the initial fuzzy model, along with its input variables, is derived from the rules extracted from the input-output data of the system being modelled. Next, the neural network is used to fine-tune the initial fuzzy model rules in order to generate the system's final ANFIS configuration.

ANFIS structure: A fuzzy inference system of Sugeno's and Takagi type with two inputs and one output rule base looks like:

if x is A and y is B then z is f(x,y)

Where z=f(x, y) is a crisp function, and A and B are the fuzzy sets. When f(x, y) is a constant, it leads to the formation of a zero-order Sugeno fuzzy model where a fuzzy singleton specifies each rule of consequence[5]. A type 3 fuzzy inference method proposed by Sugeno is shown in figure below. Here we can see that ever node output is the output of a rule which is a combination of input variables boosted by a constant term[6].The resultant of the system is the weighted average of each rule's output.

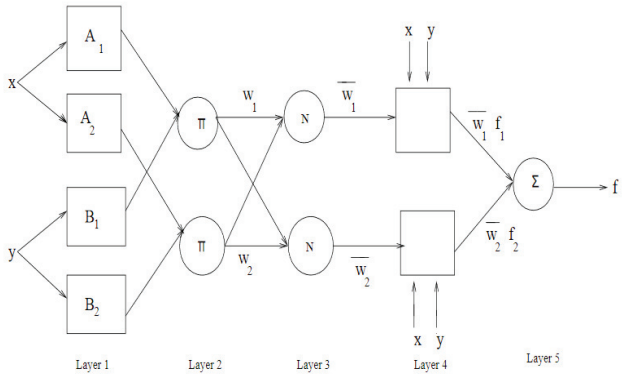


Fig. 1. ANFIS Layer architecture

Each layer in a general ANFIS structure is described as follow:

Layer 1: All nodes in layer 1 of an ANFIS structure is adapted with function:

$$Output_i^1 = \mu_{Anfis_i}(x) \quad (1)$$

Where, x is the input to node i, $Anfis_i$is the linguistic variable in association with the node function and μ_{Anfis_i} is the membership function of $Anfis_i$. $\mu_{Anfis_i}(x)$ is given as:

$$\mu_{Anfis_i}(x) = \frac{1}{1+\left[\left(\frac{x-o_i}{m_i}\right)\times\left(\frac{x-o_i}{m_i}\right)\right]^{n_i}} \quad (2)$$

Where x represents input and $\{m_i , n_i , o_i\}$ is the premise parameter set.

Layer 2: Each node in layer 2 of ANFIS represents a node with fixed value which estimates the firing strength f_i of a rule. The resultant of each node is the product of all the incoming rules i.e. input to it and is represented as:

$$Output_i^2 = f_i = \mu Anfis_i(x) \times \mu\beta_i(y) \quad (3)$$

Where i=1,2.

Layer 3: Each node in layer 4 is a set of fixed node. Each i^{th}node represents the ratio of fire strength of their rule to the sum of fire strengths of every rule of the sytem. The output of the i^{th}node is the normalized force of firing given by:

$$Output_i^3 = \overline{f_i} = \frac{f_i}{f_1+f_2} \quad (4)$$

Where i =1, 2.

Layer4: Every node in layer 4 is an adaptive node with function represented as:

$$Output_i^4 = \overline{f_i}g_i = \overline{f_i}(l_i x + m_i y + n_i), \text{ i= 1,2} \quad (5)$$

Here $\overline{f_i}$ is the output of layer 3 and $\{l_i , m_i , n_i\}$ is the corresponding set of parameters set.

Layer 5: Layer 5 of ANFIS structure is single node that represents the overall output of the system.

$$Output_i^5 = \text{Final output} = \sum \overline{f_i}g_i = \frac{\sum_i f_i g_i}{\sum_i f_i} \quad (6)$$

III. EMOTICON BASED NEURO FUZZY SYSTEM

The research work primarily shows the use of a neuro-fuzzy system in the sentimental analysis of twitter data using emoticon as a core ingredient of the data set and compare the performance with other classifiers. The first step is the data collection of twitter data using tweepy. Tweepy allows us to extract tweets based on our requirements. Here in this paper as we are using emoticons, we can use query to search for tweets using emoticons values. For example, \U0001F602 refers to happy, 1F62D = sad, 1F621= angry, 2764 = love, 1F61C = playful, 1F631 = confused and more. The methodology used can be primarily divided into three steps data preprocessing, Sentiment score generation and neuro fuzy classifier based model generation.

A. Data Preprocessing

The raw collected using tweepy from twitter contains noisy data that affects the accuracy of classifier adversely. Noisy data refers to the data which contains in necessary information that affects system performance. When we collect data using tweepy various noises are username, retweets, slangs and more. Therefore we can apply following techniques for cleaning of raw tweets.

In the preprocessing of data we first have to check if there are missing values in our dataset to make sure our dataset is consistent. We will replace them with a blank space since it might be a _ or - or punctuation with no space from the next word, and we don't want the words to join together. Next step involves removing of Re- punctuations which are not considered suitable for evaluating the sentiment score. Further the text available in tweets are split into sentences using Sentence Tokenizer.For maintaining accurate polarity score we have remove all sentences with less than five words and all the character in sentences are converted into lower case. For simplicity we have shown limited the no of emoticon score in sentences to six only which will be further illustrated. Next we remove stop word in our input data as the sop word represents word that are frequently used like he, they and more that do not carry sentimental analysis value. Stop word does not help

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:04:48 UTC from IEEE Xplore. Restrictions apply.

in discriminating relevant and irrelevant sentences therefore we remove stop word in preprocessing of data. Finally, the remaining word are converted to their base form by a process called as stemming. Stemmed verb are helpful in understanding any tweet subjectivity. The preprocessing steps for better understanding can be illustrated in block diagram as shown below.

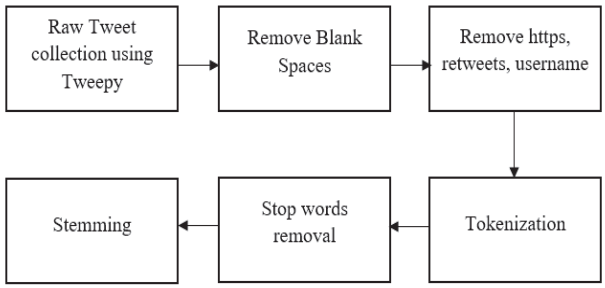


Fig. 2. Preprocessing of tweets

B. *Sentiment Score Generation*

After preprocessing of raw tweets, we can move toward generating sentiment score. Studies conducted on emoticons have estimated intensity of emoticon in terms of integer polarity. In this project, we are selecting the six most commonly used emoticon out of 751 available emoticons for showing sentiment score rest can be visualized in the emoji ranking website[7]. The emoticon selected is represented in the table below with their score.

TABLE I. EMOJI SENTIMENT SCORE

Emoji	Score
😂	0.221
😜	0.445
❤️	0.746
😭	-0.093
😡	-0.173
😕	-0.397

Using emoticon as a critical feature of our dataset, we have to make sure to cover all feature of tweets. Which include general text preprocessing, hashtag and emoticons, besides we also focus on the pattern in parts of speech (POS). The advantage of adding POS to our analysis is incorporating the subjectivity of tweets in our sentiment analysis. We achieve our goal by adding stemmed verbs and noun-adjective pairs in our sentiment score generation algorithm. Using emoticon as a critical feature of our dataset, we have to make sure to cover all feature of tweets. Which include general text preprocessing, hashtag and emoticons, besides we also focus on the pattern in parts of speech (POS). The advantage of adding POS to our analysis is incorporating the subjectivity of tweets in our sentiment analysis. We achieve our goal by adding stemmed verbs and noun-adjective pairs in our sentiment score generation algorithm. Till this point, we are left with emoticons, keywords and hashtag in our cleaned dataset our algorithm decide data to be shifted for each input encountered. Since we are using emoticon score instead of integer polarity based on Emoji Sentiment Ranking an average of all emoji score is set to 0.124833 denoted as averageChangeInSentiment[8].

Algorithm steps for generating sentiment score are as follows:

Step 1: Initialize P as sentiment score, for each emoticon encountered assign sentiment score based upon the score of emoticon in emoji ranking website.

Step 2: For every hashtag, if positive increment by 2*averageChangeInSentiment else decrement by 2*averageChangeInSentiment for negative hashtag.

Step 3: For POS tags, if an adjective precedes a noun then we decrement or increment by 2*averageChangeInSentiment depending upon the polarity of word else we decrement or increment by averageChangeInSentiment.

Step 5: For POS tags, if verb is encountered we check positive or negative on the stem of word and we decrement or increment by 2*averageChangeInSentiment. For adjective repeat the above process without stemming.

Step 6: Save sentiment score in text format for further process.

At the end of above described steps, the resultant textual data is ready to form as input to classifiers for generation of model that can determine the sentiment of tweets

C. *Neuro Fuzzy based classifier*

Neuro-fuzzy system integrates the advantage of learning in ANN and use of fuzzy system if-then rules membership function to a higher degree of accuracy[9]. In this research work, we are using ANFIS system to classify tweets because it has higher accuracy of forecasting and faster convergence in comparison to other classifiers. Once we have input data ready, we can move forward in the application of ANFIS in the next step of our work. Before applying the ANFIS algorithm, we have to develop FIS, i.e. fuzzy inference system then tune it using anfis function. There is numerous FIS model like Grid Partitioning, Subtractive Clustering and more, in this research work, we are using FCM or Fuzzy C-Means

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:04:48 UTC from IEEE Xplore. Restrictions apply.

clustering method. In FCM, we set the number of the cluster at the beginning with weights assigned randomly to them[10]. Centroids are updated by running the algorithm described previously alternatively. The process starts with taking input and output vector generated by the FIS model and checking the pair for error. The data is split into training and testing data which is essential for our model generation and validation. The training data generate the parameters for the membership function to minimize error generated between the actual and desired output. Consequent parameters are defined using the least squared method. If this error is higher than the threshold value, then premise parameters will be modified using the gradient decent process. The process is terminated when the error is lower than the threshold value.

A flowchart of ANFIS training system is shown in figure 3 below.

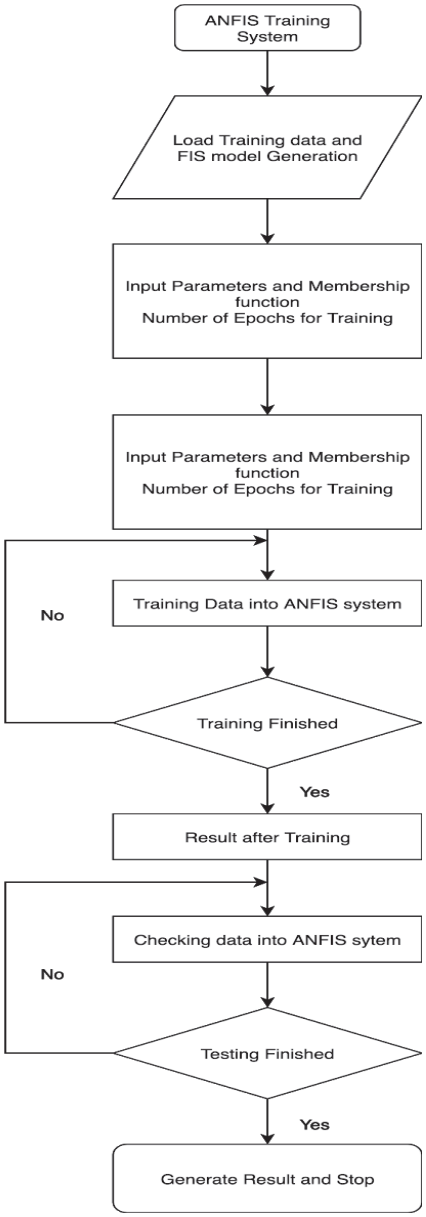


Fig. 3. ANFIS training system.

ANFIS training and learning generates the model and model is tested with testing data. Testing data contains data not included in the training process, which is used to evaluate system performance. System performance is evaluated in terms of accuracy, error curve, generated rules and output. A compact view of the system is shown below.

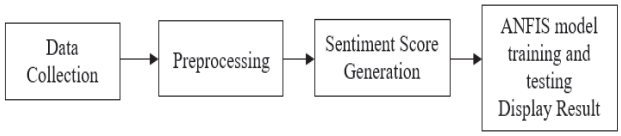


Fig. 4. Emoticon based neuro fuzzy system

IV. Results and Discussion

Using a python library tweepy, we collected recent tweets from tweet stream and filtered the tweets based on emoticons, language and retweets. Twitter exposes a STREAM API which allows the researcher to get access to the tweet stream. Any user can log into the developer account from their twitter account and request access to twitter. On approval, a set of API keys is made available for access. A total of 14123 tweets were collected and processed in preprocessing step to generate our input data set. From input data set all tweets having less than five words, retweets, usernames and stop words are discarded in preprocessing steps. After the stemming process, we were left with nearly 5000 tweets in our final dataset. Analysis of our dataset leads to the conclusion that six emoticons shown in fig are available more than the threshold of 1200 tweets. Therefore all tweets without these six emoji are discarded, and we are left with nearly 3000 tweets. At this point, we are ready with our data and can proceed toward sentiment score generation.

As soft clustering technique does not require extensive training, therefore we are only considered of incorporating all features of tweets in our sentiment score generation methodology. We used a combination of regular expression and control statement in our code to identify POS tags like noun, verb, adjective and their precedence to generate sentiment score required as an input to neuro fuzzy system. Sentiment score generated for each tweet is split into training, and testing data set on labels confused, angry, sad, happy, playful and love about which tweets are classified. Sentiment score is used to train a classification model by using k-means and ANFIS algorithm and later evaluated for efficiency.

Sentiment score generated for each tweet is indexed as input in our final dataset. In this research work, we have six labels namely happy, playful, love, sad, confused and angry to which we are classifying our data based on the proximity of sentiment score generate for each input tweet. The model generated is evaluated against this label by using metrics confusion matrix, ROC curve and accuracy for both classifiers .We run our code for all cases to get following centroid in the table 2 below.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:04:48 UTC from IEEE Xplore. Restrictions apply.

TABLE II. CENTROIDS

Emoticon	FCM+ANFIS	K-means
happy	0.21	0.41
playful	0.46	0.76
love	0.86	2.01
sad	-0.11	0.19
angry	-0.36	-0.13
confused	-0.64	-0.47

K-means algorithm requires non-overlapping of clusters which as expected leads to a fall in the performance of the classifier. From the data distribution with a cluster centroid table, we can see that K-means is very close to the desired prediction of confused and angry sentiment but perform horrendously for other sentiments.

TABLE III. K-MEANS CLASSIFICATION REPORT

	precision	recall	F1-score	support
0	0.83	0.81	0.82	537
1	0.67	0.66	0.67	491
2	0.77	0.78	0.77	532
3	0.71	0.81	0.76	496
4	0.75	0.65	0.70	420
5	0.68	0.66	0.67	471

micro avg	0.73	0.73	0.73	2947
macro avg	0.73	0.73	0.73	2947
weighted avg	0.74	0.73	0.73	2947

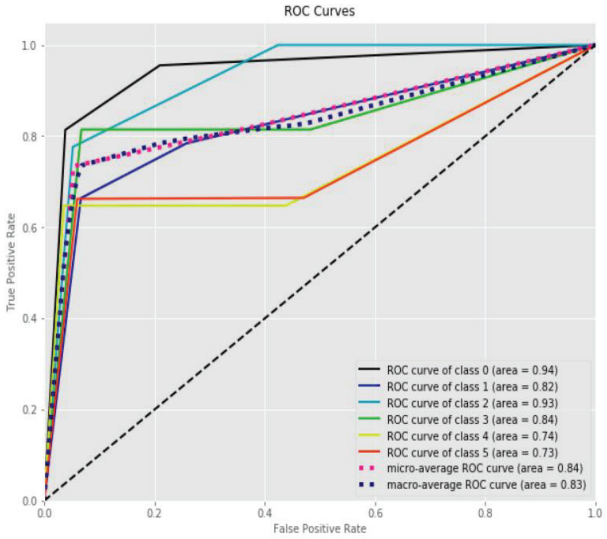


Fig. 5. ROC curve for kmeans

The evaluated overall accuracy of K-means classifier is 73.41% which is near to the estimated value for our dataset. The main reason for the lousy accuracy of k-means boils down to clustering technique, unstructured dataset, lack of incorporating subjectivity of tweets instead of focusing tweet's objectivity only. Some of these shortcoming are handled by neuro fuzzy classifier in our work.

In the ANFIS training system we used FCM clustering for FIS generation. We used genfis3 or fuzzy c-means for FIS creation in ANFIS training system.FIS object generated is then exported to the ANFIS system for tuning of our classifier.RMSE is the de facto metric for measure ANFIS performance, but for the sake of simplicity, we used accuracy metric used in other classifiers for evaluation of classifier.

TABLE IV. ANFIS CLASSIFICATION REPORT

	precision	recall	F1-score	support
0	1.00	1.00	1.00	537
1	0.88	0.80	0.84	491
2	0.83	0.90	0.86	532
3	0.86	0.96	0.91	496
4	0.93	0.87	0.90	420
5	0.89	0.83	0.86	471

micro avg	0.90	0.90	0.90	2947
macro avg	0.90	0.89	0.90	2947
weighted avg	0.90	0.90	0.90	2947

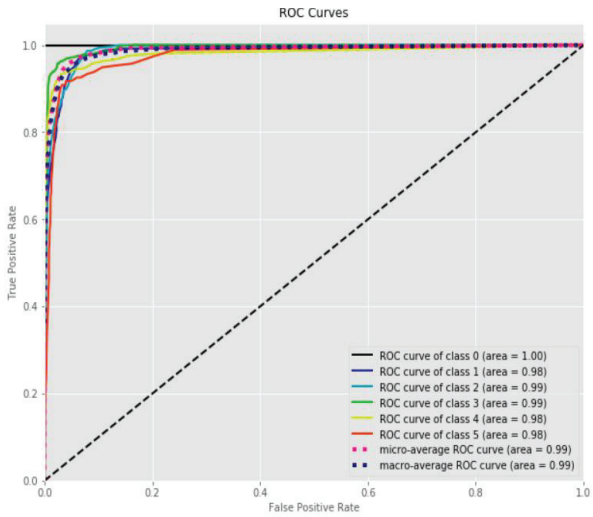


Fig. 6. ROC curve for ANFIS

ANFIS classifier, in comparison to other classifiers, is not dependent on system description .ANFIS possess the advantage of the generation of the efficient model; even system parameters are unknown as in case of data from social media like tweets we use in this project. We begin ANFIS system training by generating fuzzy inference system using FCM clustering. The cluster generated by FCM clustering

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:04:48 UTC from IEEE Xplore. Restrictions apply.

shows that cluster centroids for all the six labels are nearly identical to the desired value. The fuzzy inference system generated is tuned by using ANFIS algorithm.ANFIS system with membership function, rule base converts input data from crisp set to fuzzy values.ANFIS evaluates the rules, combines the output of each rule and finally converts the output data into the crisp form in an artificial neural network structure to generate a model for classification. We are using the evaluation metric the same as in case of k-mean, we can see that ANFIS performs with higher accuracy. The accuracy of the ANFIS system is 87.13% which is much better in comparison to the k-means algorithm.

TABLE V. ACCURACY OF CLASSIFIERS

Classifier	Accuracy
K-Means	73.41
FCM+ANFIS	87.13

V. CONCLUSION

The prime benefit of applying soft computing tools in our research work instead of machine learning algorithm like the random forest is that we can build an effective classifier around the set of input-output value without requiring large no of features and mathematical model. We compared the accuracy of our proposed model in comparison to other well-known classifiers, and the result was satisfactory. One of the most deciding factors of good accuracy is the use of effective preprocessing, which, in our case sentiment score captures the sentiment of tweets. Sentiment score generation algorithm gives a comprehensive approach on how to incorporate unique characteristics of tweets as a whole. Emoticon based neuro-fuzzy system performs with better accuracy and minimal execution time in comparison to other classifiers in our project.

We are observing an accuracy of 87.14% with our system that does not mean we have solved sentiment analysis of social data. High accuracy means our system has certain limitations. Some of them are dense text preprocessing in the form of selecting only tweets having emoticons having clear information in emoticon ranking website, and the use of emoticon prevent substantial penalty of words that can change the meaning of the entire sentence.ANFIS system also poses a limitation of extensive iteration of selecting parameters and number of the epoch that affects the model performance. To improve our system performance some of the work to be done in future scope are listed below:

- Use of bigger dataset to improve system effectiveness.
- Adding subjectivity of tweets in our sentiment generation methodology.
- Strengthening of ANFIS model by combining with other evolutionary techniques like Genetic algorithm.
- Integration of correlation of membership degree of classes in designing of fuzzy inference system

REFERENCES

[1] R. Nagamanjula and A. Pethalakshmi, "A novel framework based on bi-objective optimization and LAN2FIS for Twitter sentiment analysis," Soc. Netw. Anal. Min., vol. 10, no. 1, pp. 1–16, 2020.

[2] L. A. Ramos, A. F. Ramos, M. Melgarejo, and S. Vargas, "Tuning up Fuzzy Inference Systems by using optimization algorithms for the classification of solar flares," Tecciencia, vol. 12, no. 23, pp. 35–46, 2017.

[3] C. Jefferson, H. Liu, and M. Cocea, "Fuzzy approach for sentiment analysis," IEEE Int. Conf. Fuzzy Syst., no. July, 2017.

[4] D. Kalbande, H. Panchal, N. Swaminathan, and P. Ramaraj, "ANFIS based Spam filtering model for Social Networking Websites," Int. J. Comput. Appl., vol. 44, no. 11, pp. 32–36, 2012.

[5] K. Padmaja and N. P. Hegde, "Twitter sentiment analysis using adaptive neuro-fuzzy inference system with genetic algorithm," Proc. 3rd Int. Conf. Comput. Methodol. Commun. ICCMC 2019, no. Iccmc, pp. 498–503, 2019.

[6] S. M. Basha, Y. Zhenning, D. S. Rajput, N. C. S. N. Iyengar, and R. D. Caytiles, "Weighted fuzzy rule based sentiment prediction analysis on tweets," Int. J. Grid Distrib. Comput., vol. 10, no. 6, pp. 41–54, 2017.

[7] P. K. Novak, J. Smailović, B. Sluban, and I. Mozetič, "Sentiment of emojis," PLoS One, vol. 10, no. 12, pp. 1–22, 2015.

[8] A. Chavan, "Emoticon based Sentiment Analysis of Tweets."

[9] M. N. M. Salleh and K. Hussain, "A Review of Training Methods of ANFIS for Applications in Business and Economics," Int. J. u- e- Serv. Sci. Technol., vol. 9, no. 7, pp. 165–172, 2016.

[10] H. Wang and J. A. Castanon, "Sentiment expression via emoticons on social media," Proc. - 2015 IEEE Int. Conf. Big Data, IEEE Big Data 2015, pp. 2404–2408, 2015.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:04:48 UTC from IEEE Xplore. Restrictions apply.

PAPER • **OPEN ACCESS**

Single Dual-X Current Conveyor based BASK/BPSK Modulators

To cite this article: Deva Nand 2021 *IOP Conf. Ser.: Mater. Sci. Eng.* **1020** 012011

View the article online for updates and enhancements.

The 17th International Symposium on Solid Oxide Fuel Cells (SOFC-XVII)
DIGITAL MEETING • July 18-23, 2021
EXTENDED Abstract Submission Deadline: February 19, 2021
SUBMIT NOW

This content was downloaded from IP address 14.139.251.100 on 11/02/2021 at 06:20

Single Dual-X Current Conveyor based BASK/BPSK Modulators

Deva Nand
Department of Electronics and Communication Engineering
Delhi Technological University, Delhi, India.
devkamboj07@gmail.com

Abstract. Two voltage mode digital modulators have been proposed i.e. Binary Amplitude Shift Keying (BASK) modulator and Binary Phase Shift Keying (BPSK) modulator. Both circuits use Dual-X Second Generation Current Conveyor (DXCCII) as the active building block and have the advantages of proper input impedances and use of grounded resistor. Detailed mathematical analysis of both circuits are given. The circuits have been simulated with SPICE using 0.35µm CMOS technology parameters. The supply voltage of ±1.8 V is used. Much satisfactory results have been obtained with the proposed circuits.

Keywords: BASK; BPSK; DXCCII; modulators.

1. Introduction

Modulation is a process where one of the parameters of the carrier signal, which can be amplitude, phase or frequency, is varied linearly according to the variations in the message signal or modulating signal. With the modulation, the information contained in the message signal is transferred to the carrier signal. When modulation is done for the digital signals then it is known as the digital modulation. The device that performs this modulation is known as the digital modulator. The digital signal is a basically a binary signal in the form of 0s and 1s and it is a low pass signal and can't be transmitted through the wireless medium, so in order to transmit the digital signal through wireless medium, modulation is required. In Binary Amplitude Shift Keying (BASK), the binary 1 is represented by the presence of carrier signal while binary 0 is represented by the absence of carrier signal whereas in Binary Phase Shift Keying (BPSK), the binary 1 is represented by the presence of in phase carrier signal while binary 0 is represented by the presence of 180° out of phase carrier signal. Infrared (IR) remote controls, transmitter and receiver in optical fiber communication employ BASK, and broadband modems, satellite communication, mobile phones employ BPSK.

Many new active building blocks have been introduced in recent past to explore above mentioned applications and one such trendy block is second generation current conveyor (CCII) [1] and inverting CCII [2]. Literature survey shows that one another most trending block that can be explored to implement modulation process is dual X second generation current conveyor (DXCCII) [3,4]. As per literature survey, two EXCCII based voltage mode digital modulator circuits are reported in [5] and [6] which provide simultaneous ASK and BPSK modulation processes. A current mode BASK modulator circuit has been reported in [7] making use of two MO-CCCCTA. A voltage mode BASK

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

modulator using only one CCIII+ has been reported in [8]. Voltage mode memristor based BASK and BPSK modulators are presented in [9]. DPCCII based BASK modulator is reported in [10]. Further voltage mode MO-CCCC, digitally programmable CDTA and MO-OTA based BASK and BPSK modulator circuits are reported in [11], [12] and [13] respectively. A DXCCII based modulator topology has also been presented in [6] which performs BPSK modulations but unexplored and unanalyzed.

The newly proposed BASK modulator circuit employs single DXCCII, one NMOS transistor and one grounded resistor while BPSK modulator circuit uses single DXCCII, one NMOS and one PMOS transistor and one grounded resistor. This article includes total of six sections including introduction in Section-1. DXCCII block description is given in Section-2 followed by proposed BASK and BPSK modulators in Section-3 & 4 respectively. Proposals simulation verification is reported in Section-5 and concluded in Section-6.

2. DXCCII Description

The DXCCII, conceptually, is a combination of CCII [1] and ICCII [2]. It is therefore more versatile building block as it has the features of both the current conveyors i.e. CCII and ICCII. The block diagram of DXCCII [3] is shown below in Fig. 1. It has 5 terminals named as Y, Xp, Xn, Zp and Zn. Xp is called as the non-inverting X terminal while Xn is called as the inverting X terminal. Y is an input terminal with high impedance. Xp and Xn are terminals with low impedance while Zp and Zn are terminals with high impedance.

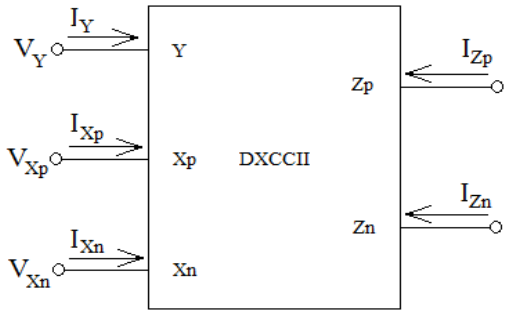


Figure 1. Block diagram of DXCCII [3].

Table 1. (W/L) Ratio of CMOS for DXCCII [3].

Transistor	W(µm)	L(µm)
M_1, M_2	1.4	0.7
M_3, M_7, M_8	2.8	0.7
M_4, M_5	2.4	0.7
M_6, M_9, M_{10}	4.8	0.7
M_{11} to M_{20}	9.6	0.7

Ideally, the voltages and currents relationships of various terminals can be described by the following matrix [3] mentioned below in (1).

$$\begin{bmatrix} I_Y \\ V_{Xp} \\ V_{Xn} \\ I_{Zp} \\ I_{Zn} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} V_Y \\ I_{Xp} \\ I_{Xn} \end{bmatrix} \quad (1)$$

The CMOS implementation of the DXCCII used in the proposed circuits is shown in Fig. 2. The aspect ratio of transistors used in the implementation of DXCCII are given in Table 1 [3].

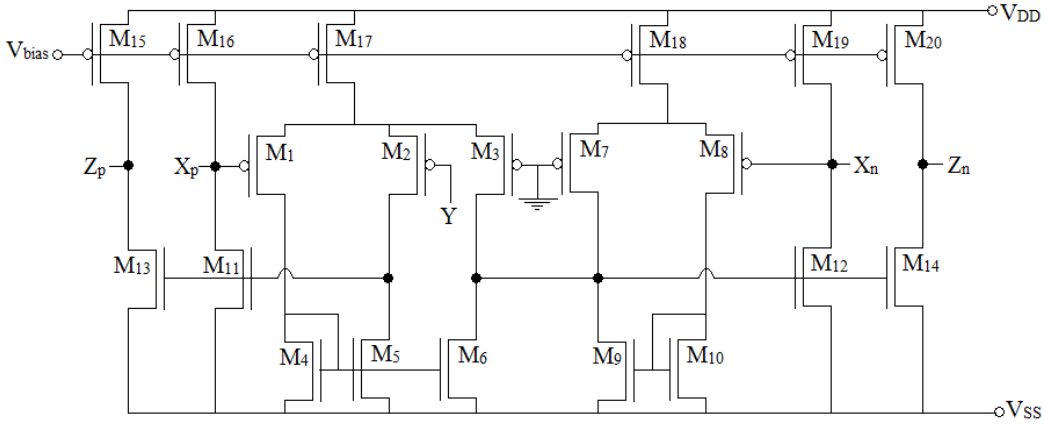


Figure 2. CMOS implementation of DXCCII [3].

3. Proposed BASK Modulator

The proposed BASK modulator is shown in Fig. 3 below.

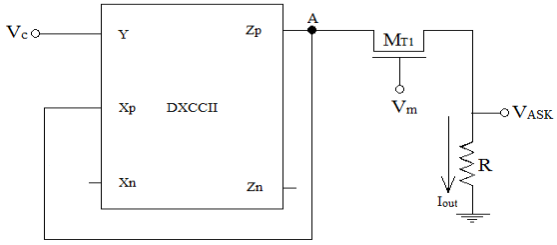


Figure 3. Proposed BASK modulator

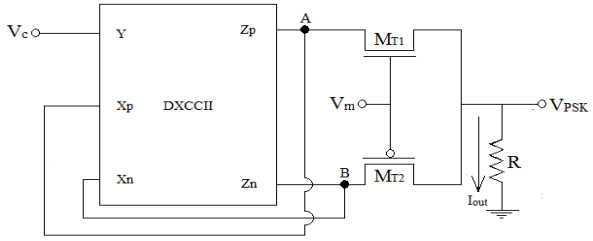


Figure 4. Proposed BPSK modulator

3.1 Operation of the proposed BASK modulator

V_c is the sinusoidal carrier signal. It carries the modulating signal V_m. Voltage V_c is applied at the high input impedance terminal Y of the DXCCII. Voltage V_c is given as

$$V_c = A_c \mathrm{Sin}(2\pi f_c t) \tag{2}$$

Where

A_c is the amplitude of the carrier signal voltage V_c,
t is the time period and f_c is the frequency in hertz (Hz) or $\mathrm{cycles/second}$.

This carrier signal voltage V_c is copied at the Xp terminal of the DXCCII, that is

$$V_{Zp} = V_{Xp} = V_c \tag{3}$$

voltage V_m is the bipolar modulating signal.
voltage V_{ASK} is the BASK modulated output voltage signal.

The M_{T1} is used as a switch. The modulating signal V_m is applied at the high impedance gate terminal of this transistor. When voltage V_m is positive, the M_{T1} turns on and works as a closed switch, and when voltage V_m is negative, the M_{T1} turns off and works as an open switch. The gate voltage of M_{T1} depends on the amplitude of the modulating signal V_m. And the drain voltage of M_{T1} depends on the amplitude of the carrier signal V_c, which is very small as voltage V_c is much smaller in magnitude than voltage V_m. The source voltage of M_{T1} depends on the voltage V_{ASK} developed across the resistor R.

Therefore, whenever M_{T1} is on, it always works in the non-saturation region or the ohmic region as it always holds the condition of working in ohmic region which is given as

$$V_{DSn} < V_{GSn} - V_{THn} \tag{4}$$

where
V_{DSn} is the drain to source voltage of M_{T1}, V_{GSn} is the gate to source voltage of M_{T1},
and V_{THn} is the threshold voltage of M_{T1}, V_{THn} for NMOS is positive.

The on resistance of this M_{T1} used as a switch is denoted by $R_{on,n}$. From [6], for the ohmic region, the resistance $R_{on,n}$ is therefore expressed as

$$R_{on,n} = \frac{1}{(W/L)_n \mu_n C_{ox,n}(V_{GSn} - V_{THn} - V_{DSn})} \tag{5}$$

Where
$(W/L)_n$ is the aspect ratio of $M_{T1,}$ μ_n is the mobility of electron, and $C_{ox,n}$ is the gate oxide capacitance per unit area, of $M_{T1.}$
It can be seen from the equation (5) that the $R_{on,n}$varies with voltage V_{GSn} and voltage V_{DSn}. But, since the on resistance $R_{on,n}$ is much smaller in magnitude than the magnitude of the external resistor R, therefore the effect of variation of $R_{on,n}$ on the BASK modulated output voltage V_{ASK} is negligible. Mathematically, it can be expressed by the following equations written below

$$R_{on,n} \ll R \tag{6}$$

The BASK modulated output voltage is given as

$$V_{ASK} = I_{out} \times R \tag{7}$$

The voltage V_A at node A is equal to the voltage V_{Xp}, that is

$$V_A = V_{Xp} \tag{8}$$

Also, when the M_{T1} is on, then the total resistance $R_{total,n}$ between node A and ground is equal to the sum of the on resistance $R_{on,n}$ of the M_{T1} and the external resistance R, i.e.

$$R_{total,n} = R_{on,n} + R \tag{9}$$

The output current I_{out} is given as

$$I_{out} = \frac{V_A}{R_{total,n}} \tag{10}$$

From equations (8), (9) and (10)

$$I_{out} = \frac{V_{Xp}}{R_{on,n} + R} \tag{11}$$

From equations (6) and (9)

$$R_{total,n} = R_{on,n} + R \cong R \tag{12}$$

Therefore from equations (11) and (12), I_{out} is given as

$$I_{out} \cong \frac{V_{Xp}}{R} \tag{13}$$

Thus from the equations (7) and (13), it can be concluded that the variation in the output current I_{out} because of $R_{on,n}$ is negligible and therefore, the effect on the output voltage V_{ASK} because of the variations in $R_{on,n}$ is also negligible. When the M_{T1} working as a switch is in off condition than the associated resistance is denoted by $R_{off,n}$. $R_{off,n}$ is ideally infinite as the M_{T1} is in non-conducting state. Now when the voltage of the modulating signal V_m is positive (logic high), M_{T1} turns on and the equal amount of the currents developed at Xp and Zp terminals of the DXCCII flow towards the node A, which then get added up at the node A as no current goes to the gate terminal of M_{T1} and then the

net current flows through the transistor M_{T1} and then flows through the resistor R and then goes to the ground. The two currents developed at Xp and Zp terminals of the DXCCII linearly depend on the voltage V_{Xp} as can be observed from the equation (13). As the voltage V_{Xp} is AC in nature therefore the currents developed because of it are also AC in nature and since these two currents add up at the node A and therefore the resultant current is also AC in nature which flows through the resistor R developing an AC voltage V_{ASK} in return across it which is in phase with the carrier signal voltage V_c. When the voltage of the modulating signal V_m is negative (logic low), M_{T1} turns off and therefore no current flows through the resistor R and the voltage V_{ASK} developed across it is zero. This way, the ASK modulated output voltage V_{ASK} is thus obtained across the resistor R.

The overall operation of the proposed BASK modulator can therefore be summarized by these two equations as mentioned below
$V_{ASK} = I_{out} \times R \cong V_c = A_c Sin(2\pi f_c t)$; when V_m is positive or logic high (NMOS M_{T1} is on)
$V_{ASK} = 0$; when V_m is negative or logic low (NMOS M_{T1} is off)

4. Proposed BPSK Modulator

The proposed BPSK modulator is shown in Fig. 4.

4.1 Operation of the proposed BPSK modulator

V_c is the sinusoidal carrier signal. It carries the modulating signal V_m. V_c is applied at the high input impedance terminal Y of the DXCCII. V_c is given as

$$V_c = A_c Sin(2\pi f_c t) \quad (14)$$

Where
A_c is the amplitude of the carrier signal voltage V_c, t is the time period and f_c is the frequency in hertz (Hz) or $cycles/second$. This carrier voltage V_c is copied at the Xp terminal of the DXCCII, that is

$$V_{Zp} = V_{Xp} = V_c = A_c Sin(2\pi f_c t) \quad (15)$$

And the voltage developed at the Xn terminal of DXCCII is the inverted carrier voltage V_c, that is

$$V_{Zn} = V_{Xn} = -V_c = A_c Sin(2\pi f_c t) \quad (16)$$

V_m is the bipolar modulating signal. V_{PSK} is the BPSK modulated output voltage signal.

The M_{T1} and M_{T2} are used as the two switches. The modulating signal V_m is applied at the high impedance gate terminals of these transistors, M_{T1} and M_{T2}. When voltage V_m is positive M_{T1} turns on and works as a closed switch while M_{T2} turns off and works as an open switch. And when voltage V_m is negative M_{T1} turns off and works as an open switch while M_{T2} turns on and works as a closed switch. The gate voltages of M_{T1} and M_{T2} depend on the amplitude of the modulating signal V_m. And the drain voltage of M_{T1} and the source voltage of M_{T2} depend on the amplitude of the carrier signal V_c, which are very small as voltage V_c is much smaller in magnitude than voltage V_m. The source voltage of M_{T1} and the drain voltage of M_{T2} depend on the voltage developed across the resistor R. Therefore, whenever M_{T1} and M_{T2} are on, they always work in the non-saturation region or the ohmic region as they always hold the condition of working in the ohmic region. The conditions for operation in the ohmic region for M_{T1} and M_{T2} are given below in equations (17) and (18) respectively.

For M_{T1}

$$V_{DSn} < V_{GSn} - V_{THn} \quad (17)$$

For M_{T2}

$$V_{DSp} > V_{GSp} - |V_{THp}| \quad (18)$$

where
V_{DSn} is the drain to source voltage of M_{T1}, V_{GSn} is the gate to source voltage of M_{T1},
V_{THn} is the threshold voltage of the M_{T1}. V_{THn} for NMOS is positive,
V_{DSp} is the drain to source voltage of M_{T2}, V_{GSp} is the gate to source voltage of M_{T2}, and
V_{THp} is the threshold voltage of the M_{T2}. V_{THp} for PMOS is negative.

The on resistances of the transistors M_{T1} and M_{T2} used as switches are denoted by $R_{on,n}$ and $R_{on,p}$ respectively. From [6], for the ohmic region, the resistances $R_{on,n}$ and $R_{on,p}$ can be expressed as

$$R_{on,n} = \frac{1}{(W/L)_n \mu_n C_{ox,n}(V_{GSn} - V_{THn} - V_{DSn})} \quad (19)$$

and

$$R_{on,p} = \frac{1}{(W/L)_p \mu_p C_{ox,p}(V_{GSp} - V_{THp} - V_{DSp})} \quad (20)$$

where
$(W/L)_n$ is the aspect ratio of the $M_{T1,}$ μ_n is the mobility of electron, $C_{ox,n}$ is the gate oxide capacitance per unit area, of $M_{T1,}$ $(W/L)_p$ is the aspect ratio of $M_{T2,}$ μ_p is the mobility of hole, $C_{ox,p}$ is the gate oxide capacitance per unit area, of $M_{T2.}$

It can be seen from the above equations (19) and (20), that $R_{on,n}$ varies with the V_{GSn} and V_{DSn}, while $R_{on,p}$ varies with the V_{GSp} and V_{DSp}. But, since both of these on resistances i.e. $R_{on,n}$ and $R_{on,p}$ are much smaller in magnitude than the magnitude of the external resistor R therefore the effect of variations of $R_{on,n}$ and $R_{on,p}$ on the BPSK modulated output voltage V_{PSK} is negligible. Mathematically, it can be expressed by the following two equations written below

$$R_{on,n} \ll R \quad (21)$$

$$R_{on,p} \ll R \quad (22)$$

The BPSK modulated output voltage is given as

$$V_{PSK} = I_{out} \times R \quad (23)$$

The voltage V_A at node A is equal to the voltage V_{Xp}, that is

$$V_A = V_{Xp} \quad (24)$$

The voltage V_B at node B is equal to the voltage V_{Xn}, that is

$$V_B = V_{Xn} \quad (25)$$

Also, only of the transistors is on at a time i.e. either M_{T1} or M_{T2}. When M_{T1} is on, then the total resistance $R_{total,n}$ between node A and ground is equal to the sum of the on resistance $R_{on,n}$ of the M_{T1} and the external resistance R. And when the M_{T2} is on, then the total resistance $R_{total,p}$ between node B and ground is equal to the sum of the on resistance $R_{on,p}$ of M_{T2} and the external resistance R. Mathematically $R_{total,n}$ and $R_{total,p}$ are given as

$$R_{total,n} = R_{on,n} + R \quad (26)$$

$$R_{total,p} = R_{on,p} + R \quad (27)$$

When only M_{T1} is on, then the output current I_{out} is

$$I_{out} = \frac{V_A}{R_{total,n}} = \frac{V_{Xp}}{R_{on,n} + R} \quad (28)$$

From equations (21) and (26)

$$R_{total,n} = R_{on,n} + R \cong R \quad (29)$$

Therefore, from equations (28) and (29)

$$I_{out} \cong \frac{V_{Xp}}{R} \tag{30}$$

Now when only M_{T2} is on, then the output current I_{out} is

$$I_{out} = \frac{V_B}{R_{total,p}} = \frac{V_{Xn}}{R_{on,p} + R} \tag{31}$$

From equations (22) and (27)

$$R_{total,p} = R_{on,p} + R \cong R \tag{32}$$

Therefore from equations (31) and (32), I_{out} is given as

$$I_{out} \cong \frac{V_{Xn}}{R} \tag{33}$$

Thus from the equations (30) and (33), it can be concluded that the variation in the output current I_{out} due of $R_{on,n}$ and $R_{on,p}$ is negligible and therefore, the effect on the output voltage V_{PSK} because of the variations in $R_{on,n}$ and $R_{on,p}$ is also negligible.

When M_{T1} and M_{T2} working as the switches are in off condition than the associated resistances are denoted by $R_{off,n}$ and $R_{off,p}$ respectively. $R_{off,n}$ and $R_{off,p}$ are infinite ideally as the M_{T1} and M_{T2} transistors are in non-conducting state or off state. Now when the voltage of the modulating signal V_m is positive (logic high), M_{T1} turns on and M_{T2} turns off and the currents developed at Xp and Zp terminals of DXCCII flow towards node A, which then get added up at node A as no current goes in to the gate terminal of M_{T1} and then the net current flows through the transistor M_{T1} and then through the resistor R and then goes to the ground. The two currents developed at Xp and Zp terminals of DXCCII linearly depend on the voltage V_{Xp} as can be seen in equation (30). As the voltage V_{Xp} is AC in nature therefore the currents developed because of it are also AC in nature and since these two currents add up at the node A and therefore the resultant current is also AC in nature which flows through the resistor R developing an AC voltage V_{PSK} in return across it which is in phase with the carrier signal voltage V_c. And when the voltage of the modulating signal V_m is negative (logic low), M_{T1} turns off and M_{T2} turns on and the currents developed at Xn and Zn terminals of DXCCII flow towards node B, which then get added up at node B as no current goes in to the gate terminal of M_{T2} and then the net current flows through the M_{T2} and then through the resistor R and then goes to the ground. The two currents developed at Xn and Zn terminals of DXCCII linearly depend on the voltage V_{Xn} and it can be seen in equation (33). As the voltage V_{Xn} is AC in nature therefore the currents developed because of it are also AC in nature and since these two currents add up at the node A and therefore the resultant current is also AC in nature which flows through the resistor R developing an AC voltage V_{PSK} in return across it which is 180° out of phase with the carrier signal voltage V_c.

The operation can be summarized by these two equations as mentioned below

$V_{out} = I_{out} \times R \cong V_c = A_c Sin(2\pi f_c t)$; when V_m is positive or logic high (NMOS M_{T1} is on and PMOS M_{T2} is off)

$V_{out} = I_{out} \times R \cong -V_c = -A_c Sin(2\pi f_c t)$; when V_m is negative or logic low (NMOS M_{T1} is off and PMOS M_{T2} is on)

In conclusion, the output voltage V_{ASK} of the proposed BASK modulator and the output voltage V_{PSK} of the proposed BPSK modulator can mathematically be written in the tabulated form as shown below in the Table 2.

Table 2. Summary of outputs of the proposed BASK and BPSK modulators

Modulating signal (V_m)	**BASK output (V_{ASK})**	**BPSK output (V_{PSK})**
Logic HIGH (1V)	$A_c Sin(2\pi f_c t)$	$A_c Sin(2\pi f_c t)$
Logic LOW (−1V)	0	$-A_c Sin(2\pi f_c t)$

5. Simulation results of the proposed BASK and BPSK modulator

The BASK & BPSK modulators are simulated with SPICE using CMOS implementation of the active block DXCCII [35] and 0.35µm CMOS technology parameters. The supply voltage for the active block DXCCII is taken as ± 1.8V and the voltage V_{bias} is set to 1.05V. The aspect ratios (W/L) of the transistors M_{T1} and M_{T2} used as the switches are same i.e. $30\mu m/0.7\mu m$. The value of external resistor R is selected as 5KΩ. The waveform of the input carrier voltage signal V_c is shown in Fig. 5. The carrier signal V_c is having a peak amplitude A_c of 100mV and frequency f_c of 20MHz. The waveform for the modulating signal V_m is shown in Fig. 6. The waveforms for the ASK modulated output voltage V_{ASK} and PSK modulated output voltage V_{PSK} are shown above in Fig. 7 and Fig. 8 respectively.

The logic high and logic low levels of the modulating signal V_m are defined as follows

$V_m = 1V$; for logic high or logic 1 or bit 1

$V_m = -1V$; for logic low or logic 0 or bit 0

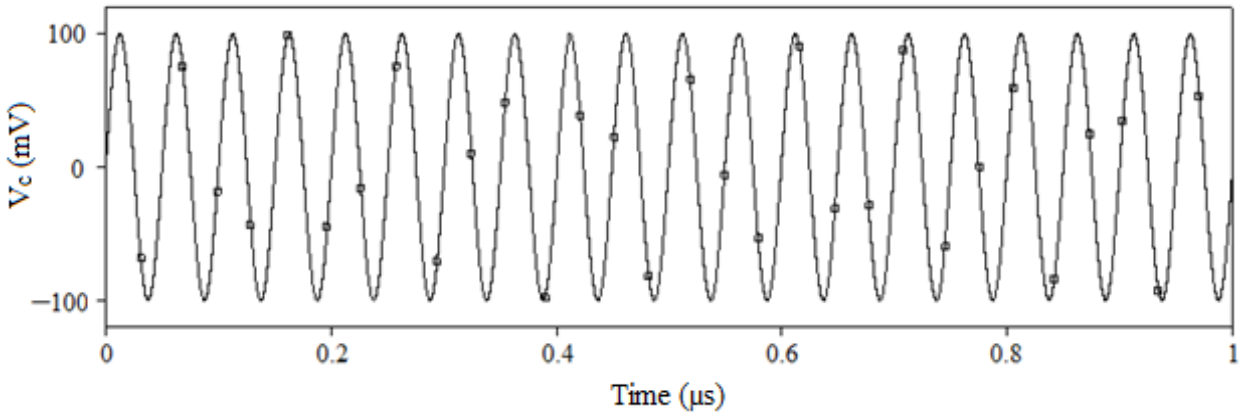


Figure 5. Waveform of the carrier signal V_c

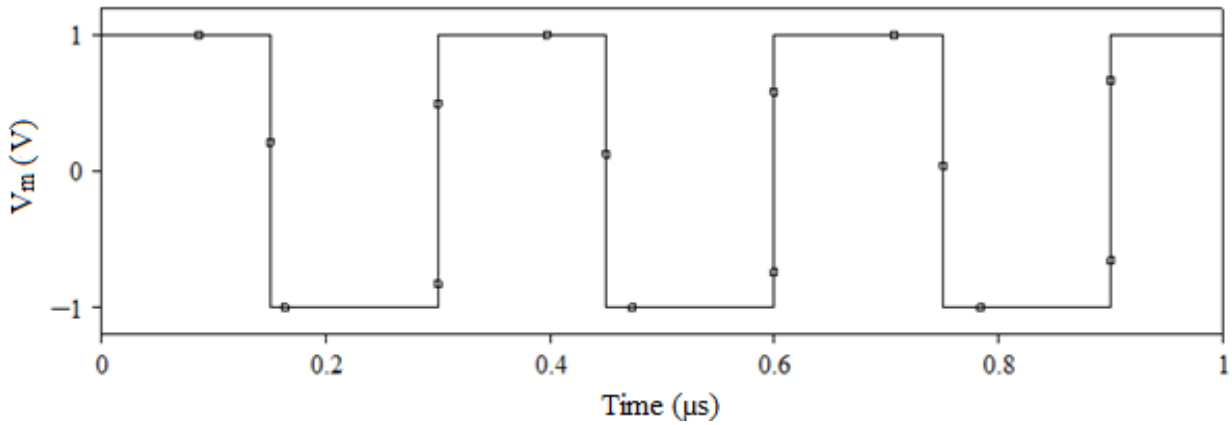


Figure 6. Waveform of the modulating signal V_m

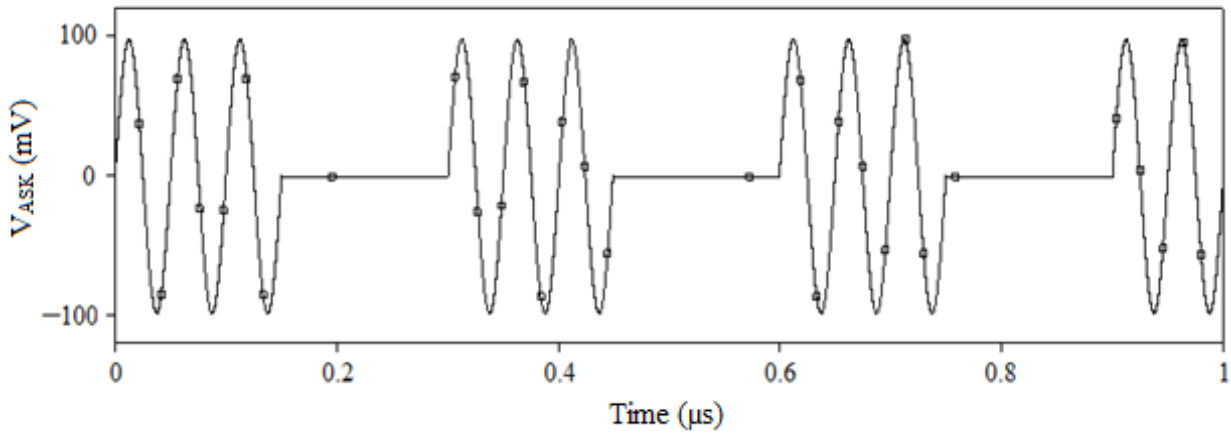


Figure 7. ASK modulated output voltage V_{ASK}

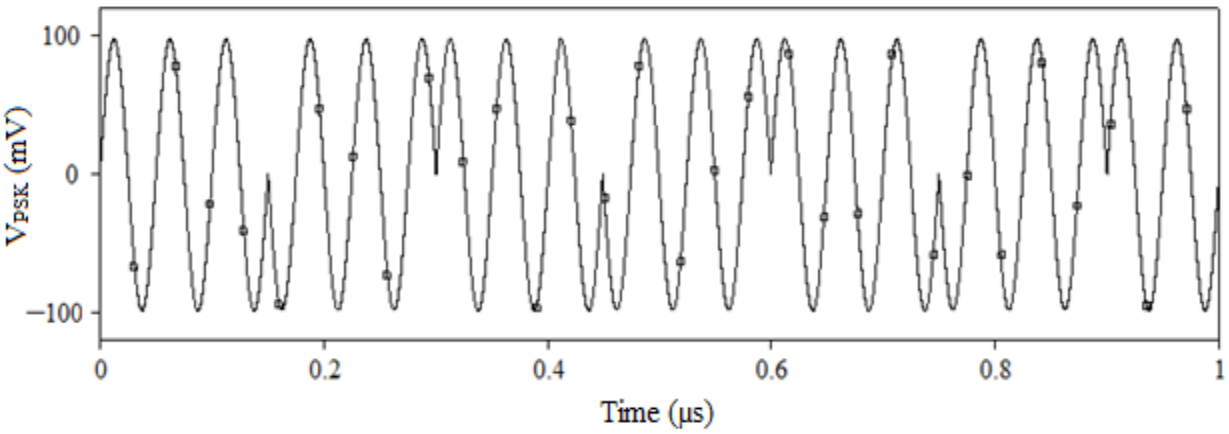


Figure 8. PSK modulated output voltage V_{PSK}

Table 3. Comparison of the various Modulators

Ref. no.	Active elements	No. of floating/ grounded resistors	No. of capacitors	No. of MOS	BASK/ BPSK	Max. carrier frequency (Hz)	Supply voltage
[5]	1 EXCCII 1 MOS	1/0	0	23+1	Both	20M	±1.25
[6]	1 EXCCII 2 MOS	0/3	0	21+2	Both	20M	±1.25
[7]	2 MO-CCCCTA	0/0	3	2×26	ASK	2.11M	±0.9
[8]	1 CCCII+	0/1	2	NA	ASK	20.78M	±1.5
[9]	1 VCCS Memristors	0/1	1	NA	Both	1K	NA
[10]	1 DPCCII	0/2	0	18	ASK	125K	±0.75
[11]	1 MO-CCCC	1/2	3	NA	Both	1.5M	NA
[12]	1 Digitally programmable CDTA	0/2	2	67	Both	34.39M	±0.8
[13]	2 MO-OTA	0/1	3	NA	Both	NA	NA
Proposed	1 DXCCII, 1 MOS	0/1	0	20+1	ASK	20M	±1.8
Proposed	1 DXCCII, 2 MOS	0/1	0	20+2	PSK	20M	±1.8

6. Conclusion

Two voltage mode digital modulators have been proposed namely BASK and BPSK. BASK employs one DXCCII, one grounded resistor and one NMOS transistor while BPSK employs one DXCCII, one grounded resistor, one NMOS and one PMOS transistor. Both the above proposed circuits provide the advantage of having proper input impedances and use of grounded resistor. A comparative study table of various digital modulators available in the open literature is given in table 3. All the proposed circuits have been simulated with SPICE using 0.35µm CMOS technology parameters. The supply voltage of $\pm 1.8V$ is used. Convincing results have been obtained with all the proposed circuits.

References

[1] Sedra, A. S., Smith, K. C., "A second generation current conveyor and its applications", IEEE Transactions on Circuit Theory, 17, pp. 132-134, 1970.
[2] Awad, I. A., Soliman, A. M., "Inverting second generation current conveyor: the missing building blocks, CMOS realizations and applications", Int. J. Electron. (AEÜ), 86, pp. 413-432, 1999.
[3] Zeki, A., Toker, A., "The dual-X current conveyor (DXCCII): a new active device for tunable continuous-time filters", Int. J. Electron. (AEÜ), 89(12), pp. 913-923, 2002.
[4] Mohan, J., Chaturvedi, B., Maheshwari, S., "Novel current-mode all-pass filter with minimum component count", I.J. Image, Graphics and Signal Processing, 12, pp. 32-37, 2013.
[5] Maheshwari, S., "Voltage-mode full-wave precision rectifier and an extended application as ASK/BPSK circuit using a single EXCCII", Int. J. Electron. (AEÜ), 84, pp. 234-241, 2018.
[6] Maheshwari, S., "Realization of simple electronic functions using EXCCII", 26(11), Journal of Circuits, Systems, and Computers, 2017.
[7] Chen, H., Hwang, Y., Ku, Y., "Voltage-mode and current-mode resistorless third-order quadrature oscillator", Appl. Sci., 6, 2016.
[8] Khan A. M., "Generation of ASK and FSK from digitally controlled CCCII+ oscillator", The International Journal of Engineering and Science (IJES), 2(10), pp. 19-22, 2013.
[9] Goknar, I. C., Oncul, F., Minayi, E., "New memristor applications: AM, ASK, FSK and BPSK modulators", IEEE Antennas Wave Propag. Mag., 55(2), pp. 304-313, 2013.
[10] Khan, I. A., Simsim, M. T., "An ASK modulator for RFID applications using low voltage digitally controlled programmable CMOS-CCII", Saudi Int. Electronics, Communication, and Photonics Conf., pp. 1-4, 2013.
[11] Abuelma'Atti M.T., "New ASK/FSK/PSK/QAM wave generator using a single current-controlled multiple output current conveyor", Int. J. Electron. (AEÜ), 89(1), pp. 35-43, 2002.
[12] Xia, Z., Wang, C., Jin, J., Du, S., Lin, H., Yang, H., "Novel AM/FM/ASK/FSK/PSK/QAM signal generator based on a digitally programmable CDTA", Circuits Syst Signal Process, 34(5), pp. 1635-1653, 2014.
[13] Abuelma'Atti, M.T., "New ASK/FSK/PSK/QAM wave generator using multiple-output operational transconductance amplifiers", IEEE Trans. Circuits Syst. I, Fundam. Theory Appl., 48(4), pp. 487-490, 2001.



Deva Nand did his B.Tech in Electronics and Communication Engineering, M.Tech in Microelectronics and VLSI Design from Kurukshetra University, Kurukshetra, India and Ph.D. from Delhi Technological University, Delhi, India. At present he is Assistant Professor in Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India. A life member of ISTE, member of IEEE, USA, life member of IAENG and member of IAASSE. His research interests include analog mixed signal VLSI design and digital system design.

International Journal of Theoretical Physics
https://doi.org/10.1007/s10773-021-04725-z

Teleportation Criteria Based on Maximum Eigenvalue of the Shared $d \otimes d$ Dimensional Mixed State: Beyond Singlet Fraction

Anuma Garg[1] · **Satyabrata Adhikari**[1]

Received: 24 September 2020 / Accepted: 9 January 2021 / Published online: 20 January 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC part of Springer Nature 2021

Abstract
We derive a criteria for the detection of $d \otimes d$ dimensional negative partial transpose (NPT) entangled state useful for teleportation. The newly derived criteria are based on the maximum eigenvalue of the NPT entangled state, which is easier to determine experimentally than to completely reconstruct the state via tomography. We then illustrate our criteria by considering a class of qubit-qubit system and qutrit-qutrit system.

Keywords Quantum teleportation · Teleportation fidelity · Singlet fraction · Eigenvalue · Dembo's bound · Structural physical approximation

1 Introduction

Quantum teleportation is an important topic to study in quantum information science. It plays a vital role in the development of quantum information theory and quantum technologies [1, 2]. Bennett et al. [3] have developed the first protocol of quantum teleportation for two-qubit system. The developed protocol talks about the transfer of information contained in a qubit from a sender (say, Alice) to a receiver (say, Bob). To execute this protocol, firstly Alice and Bob shared an entangled state between them. Then she perform two-qubit Bell-state measurement on particles in her possession. After that, she communicates the measurement result to Bob by sending two classical bits. The receiver Bob then reconstructs the quantum state at his place by applying suitable unitary operation such as $I, \sigma_x, \sigma_y, \sigma_z$ on his qubit according to the measurement outcome sent by the sender Alice.

The first protocol of quantum teleportation can be considered as a basic protocol for other quantum schemes such as quantum repeater [4], quantum gate teleportation [5], port-based teleportation [6]. We should note that the most important ingredient in quantum teleportation

✉ Satyabrata Adhikari
satyabrata@dtu.ac.in

Anuma Garg
anumagarg_phd2k18@dtu.ac.in

[1] Delhi Technological University, Delhi, 110042, India

protocol is the resource state, which is shared between two distant parties because it would not be possible to realize quantum teleportation without shared entangled state. In a realistic situation, it is not possible to keep the shared entangled state in a pure form as the particle has to interact with the environment. Thus, the shared entangled state in general become a mixed entangled state. In this context, the question may arise that whether the generated mixed entangled state is useful as a resource state in quantum teleportation? To answer this question, a quantity known as singlet fraction has been defined [7]. It is defined as the maximum overlap between the quantum state described by the density operator ρ and a maximally entangled state in a finite dimensional Hilbert space. The usefulness of the shared entangled state between two distant partners in a teleportation protocol depends on the value of the singlet fraction. For $d \otimes d$ dimensional system, the shared state is useful in quantum teleportation if the singlet fraction of the shared state is greater than $\frac{1}{d}$.

The singlet fraction also characterizes the nature of the quantum state in the sense that if the given state is separable then the singlet fraction of the given two-qubit mixed state is less than or equal to $\frac{1}{2}$ [7]. In another way, it can be stated as if the singlet fraction of an arbitrary given state is greater than $\frac{1}{2}$ then the state is entangled. But the converse of the statement is not true. This means that there exist two-qubit mixed entangled state whose singlet fraction is less than or equal to $\frac{1}{2}$ and hence not useful in quantum teleportation. In this perspective, Badziag et al. [8] have shown that a dissipative interaction with the environment is sufficient to improve the value of the singlet fraction. They have presented a class of entangled quantum state whose singlet fraction is exactly equal to $\frac{1}{2}$ before interaction with the environment but after the interaction with the environment, the value of singlet fraction improves. Even getting this result also, the question remains that whether interaction with the environment increases the singlet fraction of any two-qubit mixed state? The answer is in affirmative. Verstraete et al. [9] have studied this problem and obtained trace preserving LOCC that enhances the singlet fraction and make its value greater than $\frac{1}{2}$ for any two-qubit mixed entangled state. They have derived a connection between the optimal singlet fraction and partial transpose of a given state. The established relation tells us that the two-qubit state is useful as a resource state for teleportation if and only if the optimal singlet fraction is greater than $\frac{1}{2}$.

Till now, we have discussed about the resource state useful in teleportation for $2 \otimes 2$ dimensional system. We now continue our discussion with higher dimensional system. Generally, it has been proved that to teleport an arbitrary d-dimensional pure state, only maximally entangled pure state in $d \otimes d$ is required [10]. Zhao et al [11] have derived the necessary and sufficient conditions of faithful teleportation of an arbitrary d-dimensional pure state with $m \times d$ and $d \times n$ dimensional entangled resource, where m and n denoting the dimension of the first qubit in $m \times d$ and second qubit in $d \times n$ dimensional entangled resource states respectively. A general expression for the output state of the quantum channel associated with the original teleportation protocol with an arbitrary $d \otimes d$ dimensional mixed resource state has been obtained in [12].

The motivation of this work is two fold: (i) Since partial transposition is a non-physical operation and cannot be implemented in a laboratory so we apply SPA on partial transposition of the given state. Being SPA a completely positive map, the expression of singlet fraction get free from partial transposition operation and thus it can be implemented in an experiment. (ii) The second motivation comes from the problem of estimation of singlet fraction in higher dimensional bipartite system. Since singlet fraction depends on bipartite maximally entangled states but in higher dimensional system, it is very difficult to construct bipartite maximally entangled state so it would not be an easy task to get the experimentally

estimated value of singlet fraction for higher dimensional bipartite system. Thus, it is desirable to establish another criteria for the detection of entangled state useful in teleportation and that must be easy to implement in experiment. Therefore, instead of singlet fraction if the criterion is expressed in terms of the eigenvalue then it requires lesser number of measurements than to completely reconstruct the state via tomography [13]. Thus, in terms of number of required measurements, the criterion based on eigenvalue is more efficient than quantum tomography.

This paper is organized as follows: In Section 2, we state some preliminary result that will be needed in the later sections. In Section 3, we revisit Verstraete et.al. work and modify their result in terms of SPA-PT of a given quantum state. In Section 4, we give the criterion in terms of maximum eigenvalue for the detection of $d \otimes d$ dimensional NPT entangled resource state useful in quantum teleportation. In Section 5, we will show that the criterion based on eigenvalue may fail and thus we provide another criterion based on Dembo's bound. Lastly, we conclude in Section 6.

2 Preliminary Results

Result 1: For any two Hermitian $d \times d$ matrices A and B, we have [14, 15]

$$\lambda_{min}(A)Tr(B) \leq Tr(AB) \leq \lambda_{max}(A)Tr(B) \tag{1}$$

where the eigenvalues of A are arranged as $\lambda_{min} = \lambda_1 \leq \lambda_2 \leq \lambda_3 \leq ..... \leq \lambda_d = \lambda_{max}$.

Result 2: An arbitrary $d \otimes d$ dimensional entangled quantum state ρ is useful for quantum teleportation if [8]

$$F^{max}(\rho) > \frac{1}{d} \tag{2}$$

where $F^{max}(\rho)$ denote the singlet fraction of ρ and it is given by [16]

$$\begin{aligned} F^{max}(\rho) \; = \; & max_{U_A,U_B}\{F[\rho, (U_A \otimes U_B)|\phi_d^+\rangle\langle\phi_d^+| \\ & (U_A^\dagger \otimes U_B^\dagger)]\} \end{aligned} \tag{3}$$

where $|\phi_d^+\rangle = \frac{1}{\sqrt{d}} \sum_{i=0}^{d-1} |ii\rangle$.

It is known that there exist an entangled state ρ for which $F^{max}(\rho) \leq \frac{1}{d}$ but these entangled states may or may not be useful for teleportation.

Result 3: In standard teleportation scheme, the relation between the singlet fraction of ρ in $d \otimes d$ dimensional Hilbert space and maximal achievable teleportation fidelity $f^{tel}(\rho)$ is given by [7]

$$f^{tel}(\rho) = \frac{dF^{max}(\rho) + 1}{d + 1} \tag{4}$$

The teleportation fidelity can be regarded as useful if $f^{tel}(\rho) > \frac{2}{d+1}$, where the expression $\frac{2}{d+1}$ denoting the classical teleportation fidelity.

Result 4: Let us consider an arbitrary two qubit state described by the density operator ρ_{12}

$$\rho_{12} = \begin{pmatrix} e_{11} & e_{12} & e_{13} & e_{14} \\ e_{12}^* & e_{22} & e_{23} & e_{24} \\ e_{13}^* & e_{23}^* & e_{33} & e_{34} \\ e_{14}^* & e_{24}^* & e_{34}^* & e_{44} \end{pmatrix}, \sum_{i=1}^{4} e_{ii} = 1 \tag{5}$$

where $(*)$ denotes the complex conjugate, then the structural physical approximation (SPA) of partial transpose (PT) of ρ_{12} is given by [17, 18]

$$\begin{aligned}\tilde{\rho}_{12} &= [\frac{1}{3}(I \otimes \widetilde{T}) + \frac{2}{3}(\widetilde{\Theta} \otimes D)](\rho_{12}) \\ &= \begin{pmatrix} E_{11} & E_{12} & E_{13} & E_{14} \\ E_{12}^* & E_{22} & E_{23} & E_{24} \\ E_{13}^* & E_{23}^* & E_{33} & E_{34} \\ E_{14}^* & E_{24}^* & E_{34}^* & E_{44} \end{pmatrix}\end{aligned} \tag{6}$$

where

$$\begin{aligned} &E_{11} = \frac{1}{9}(2 + e_{11}),\ E_{12} = \frac{1}{9}e_{12}^*,\ E_{13} = \frac{1}{9}e_{13}, \\ &E_{14} = \frac{1}{9}e_{23},\ E_{22} = \frac{1}{9}(2 + e_{22}),\ E_{23} = \frac{1}{9}e_{14}, \\ &E_{24} = \frac{1}{9}e_{24},\ E_{33} = \frac{1}{9}(2 + e_{33}),\ E_{34} = \frac{1}{9}e_{34}^*, \\ &E_{44} = \frac{1}{9}(2 + e_{44}) \end{aligned} \tag{7}$$

$\widetilde{T}$ is the SPA of transpose map T and $\widetilde{\Theta}$ denotes the SPA of inversion map Θ. The inversion map Θ is defined as $\Theta(\rho)$=-ρ and $D(\rho) = \frac{I_2}{2}$ denotes the polarization. SPA-PT for two qubit photonic system using single-photon polarization qubits and linear optical devices has been demonstrated in [19].

Result 5: For any $n \otimes n$ Hermitian positive semi-definite operator R_n with eigenvalues $\lambda_1 \leq \lambda_2 \leq .....\lambda_n$, Dembo's bound [20, 21] is given by

$$\begin{aligned} &\frac{c + \eta_1}{2} + \sqrt{\frac{(c - \eta_1)^2}{4} + (b^*)^T b} \leq \lambda_n(R_n) \\ &\leq \frac{c + \eta_{n-1}}{2} + \sqrt{\frac{(c - \eta_{n-1})^2}{2} + (b^*)^T b} \end{aligned} \tag{8}$$

where $R_n = \begin{pmatrix} R_{n-1} & b \\ (b^*)^T & c \end{pmatrix}$, η_1 is the lower bound on minimal eigenvalue of R_{n-1}, η_{n-1} is the upper bound on maximal eigenvalue of R_{n-1} and b is a vector of dimension $n - 1$.

3 Revisiting Maximal Singlet Fraction of Mixed Two-Qubit State

Verstraete et al. [9] have derived the optimal trace-preserving local operation together with classical communication and have shown that it optimally increases the singlet fraction of a mixed quantum state ρ_{12} and hence maximize its teleportation fidelity. They have studied the case of two-qubit system and proved that if the state is entangled then it is always possible to increase the singlet fraction above $\frac{1}{2}$ and thus make teleportation fidelity greater than $\frac{2}{3}$. Since their result is based on the partial transposition operation so it is not possible to realize it in the experimental setup. In this section, we revisit their result and apply structural physical approximation (SPA) of partial transposition. By doing this, the final expression of the singlet fraction get modified and since SPA of partial transposition is a completely positive map so the singlet fraction can be estimated experimentally. However, we find that

in this case the value of the singlet fraction is not always greater than $\frac{1}{2}$. Therefore, the result of this section motivate us further to investigate for new teleportation criteria that will be studied in the following section.

The expression of optimal singlet fraction after LOCC for a two-qubit state ρ_{12} is given by [9]

$$F^{opt}_{LOCC}(\rho_{12}) = \frac{1}{2} - Tr(X^{opt}\rho^{\Gamma}_{12}) \tag{9}$$

where without any loss of generality, we denote Γ as the partial transposition with respect to the second subsystem, and X^{opt} is described by a 4×4 matrix, which will be of rank one and it is expressed in the form as

$$X^{opt} = (A \otimes I_2)|\psi\rangle_{12}\langle\psi|(A^{\dagger} \otimes I_2) \tag{10}$$

with $|\psi\rangle_{12} = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$, A denotes a filter described by a 2×2 matrix and I_2 represents a 2×2 identity matrix.

It is clear from equation (9) that if ρ_{12} is an entangled state then $F^{opt}_{LOCC}(\rho_{12})$ always greater than $\frac{1}{2}$ and hence one can reach the conclusion that every mixed two-qubit entangled state is useful in teleportation. But an important point to note here is that the partial transposition Γ is not a completely positive map so ρ^{Γ}_{12} is not a physically realizable operation. Thus the value of $F^{opt}_{LOCC}(\rho_{12})$ cannot be realized in experiment. To overcome this problem, we use SPA-PT of ρ_{12} and re-express $Tr(X^{opt}\rho^{\Gamma}_{12})$ given in (9) as [17, 18]

$$Tr(X^{opt}\rho^{\Gamma}_{12}) = 9Tr(X^{opt}\widetilde{\rho}_{12}) - 2 \tag{11}$$

where $\widetilde{\rho}_{12}$ is the SPA-PT of ρ_{12}.

Using (11), equation (9) can be re-expressed as

$$F^{opt}_{LOCC}(\rho_{12}) = \frac{5}{2} - 9Tr(X^{opt}\widetilde{\rho}_{12}) \tag{12}$$

Now it is possible to estimate the value of $F^{opt}_{LOCC}(\rho_{12})$ experimentally but we have to pay cost in a way that the obtained value of $F^{opt}_{LOCC}(\rho_{12})$ may not be always greater than $\frac{1}{2}$. If it may happen that $Tr(X^{opt}\widetilde{\rho}_{12}) < \frac{2}{9}$ then only $F^{opt}_{LOCC}(\rho_{12})$ is greater than $\frac{1}{2}$. This implies that if there exist state for which $Tr(X^{opt}\widetilde{\rho}_{12}) \geq \frac{2}{9}$ then $F^{opt}_{LOCC}(\rho_{12}) \leq \frac{1}{2}$. To illustrate, let us consider a two-qubit state described by the density operator

$$\sigma_{12} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & b & f & 0 \\ 0 & f^* & d & 0 \\ 0 & 0 & 0 & e \end{pmatrix}, \quad b + d + e = 1 \tag{13}$$

where $*$ denote the complex conjugation.

The density matrix σ_{12} has been studied by many authors in different contexts [22–27]. The state σ_{12} is an entangled state and its concurrence is given by [22, 23]

$$C(\sigma_{12}) = 2|f| \tag{14}$$

Using Result 4, we can obtain the SPA-PT of σ_{12} as

$$\widetilde{\sigma}_{12} = \begin{pmatrix} \frac{2}{9} & 0 & 0 & \frac{f}{9} \\ 0 & \frac{2+b}{9} & 0 & 0 \\ 0 & 0 & \frac{2+d}{9} & 0 \\ \frac{f^*}{9} & 0 & 0 & \frac{2+e}{9} \end{pmatrix} \tag{15}$$

Now if we consider the filter A of the form as $\begin{pmatrix} a & 0 \\ 0 & 1 \end{pmatrix}, 0 \leq a \leq 1$, then X^{opt} is given by

$$X^{opt} = \begin{pmatrix} \frac{a^2}{2} & 0 & 0 & \frac{a}{2} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \frac{a}{2} & 0 & 0 & \frac{1}{2} \end{pmatrix} \tag{16}$$

The optimal singlet fraction of σ_{12} is given by

$$F^{opt}_{LOCC}(\sigma_{12}) = \frac{5}{2} - 9Tr(X^{opt}\widetilde{\sigma}_{12}) \tag{17}$$

where $Tr(X^{opt}\widetilde{\sigma}_{12})$ is given by

$$Tr(X^{opt}\widetilde{\sigma}_{12}) = \frac{2a^2 + 2aRe(f) + 2 + e}{18} \tag{18}$$

The inequality $\frac{2a^2+2aRe(f)+2+e}{18} \geq \frac{2}{9}$ holds if the filtering parameter a satisfies

$$\frac{-Re(f) + \sqrt{Re(f)^2 - 2e + 4}}{2} \leq a \tag{19}$$

Therefore, $F^{opt}_{LOCC}(\sigma_{12})$ is less than equal to $\frac{1}{2}$ iff (19) holds.

Let us now consider a particular case where we can set the values of the state parameter as: $b = 0.2$, $d = 0.4$, $e = 0.4$ and $f = 0.25 + 0.1i$. Using these values, the density matrix given in (13) reduces to

$$\sigma^{(1)}_{12} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0.2 & 0.25 + 0.1i & 0 \\ 0 & 0.25 - 0.1i & 0.4 & 0 \\ 0 & 0 & 0 & 0.4 \end{pmatrix} \tag{20}$$

The SPA-PT of $\sigma^{(1)}_{12}$ is given by

$$\tilde{\sigma}^{(1)}_{12} = \begin{pmatrix} \frac{2}{9} & 0 & 0 & \frac{0.25+0.1i}{9} \\ 0 & \frac{2.2}{9} & 0 & 0 \\ 0 & 0 & \frac{2.4}{9} & 0 \\ \frac{0.25-0.1i}{9} & 0 & 0 & \frac{2.4}{9} \end{pmatrix} \tag{21}$$

The value of $Tr(X^{opt}\widetilde{\sigma}^{(1)}_{12})$ is given by

$$Tr(X^{opt}\widetilde{\sigma}^{(1)}_{12}) = \frac{2a^2 + 0.5a + 2.4}{18} \tag{22}$$

Thus, the optimal singlet fraction of $\sigma^{(1)}_{12}$ is given by

$$F^{opt}_{LOCC}(\sigma^{(1)}_{12}) = \frac{2.6 - 2a^2 - 0.5a}{2}, 0.78 \leq a \leq 1 \tag{23}$$

Figure 1 illustrate the fact that $F^{opt}_{LOCC}(\sigma^{(1)}_{12})$ is always less than or equal to $\frac{1}{2}$.

This motivates us to investigate for new teleportation criteria which may go beyond singlet fraction and identify not only two-qubit entangled state but also higher dimensional NPT entangled state, which may be useful in quantum teleportation. We will study this in the later sections.

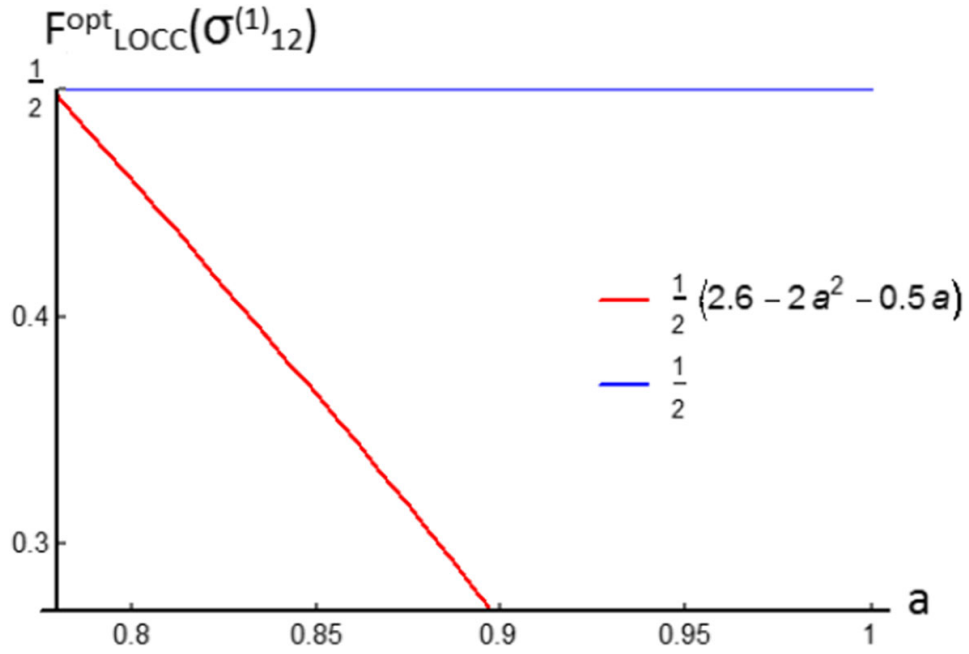


Fig. 1 Plot of optimal singlet fraction obtained after LOCC operation versus the filtering parameter a

4 Teleportation Criteria in Terms of Maximum Eigenvalue

In this section, we derive a criterion for the usefulness of the shared $d \otimes d$ dimensional NPT entangled states in quantum teleportation. In particular, for $2 \otimes 2$ dimensional entangled states, the derived criterion may be useful in a situation when singlet fraction calculated before sending a qubit through the local environment is less than or equal to $\frac{1}{2}$. This means that if we don't use local environment to increase the value of singlet fraction, our criterion can still detect whether the shared resource state is useful for teleportation or not.

Lemma 1 *The maximum eigenvalue of an arbitrary $d \otimes d$ dimensional quantum state ρ is always greater than or equal to the singlet fraction of ρ. Mathematically, it can be expressed as*

$$\lambda_{max}(\rho) \geq F^{max}(\rho) \tag{24}$$

where $\lambda_{max}(\rho)$ denote the maximum eigenvalue of ρ.

Proof Let us start with the definition of the singlet fraction given in (3), which can be re-expressed as

$$\begin{aligned} F^{max}(\rho) &= max_{U_A,U_B} Tr[\rho(U_A \otimes U_B)|\phi_d^+\rangle\langle\phi_d^+|(U_A^\dagger \otimes U_B^\dagger)] \\ &= max_{U_A,U_B} Tr[(U_A^\dagger \otimes U_B^\dagger)\rho(U_A \otimes U_B)|\phi_d^+\rangle\langle\phi_d^+|] \\ &\leq \{max_{U_A,U_B}\lambda_{max}[(U_A \otimes U_B)\rho(U_A^\dagger \otimes U_B^\dagger)]\} \\ &\quad \{Tr[|\phi_d^+\rangle\langle\phi_d^+|]\} \\ &= max_{U_A,U_B}\lambda_{max}[(U_A \otimes U_B)\rho(U_A^\dagger \otimes U_B^\dagger)] \\ &= \lambda_{max}(\rho) \end{aligned} \tag{25}$$

Hence proved. □

The inequality in the third step is a consequence of the result 1 and the last equality follows from a well known fact that the two quantum states $(U_A \otimes U_B)\rho(U_A^\dagger \otimes U_B^\dagger)$ and ρ have same set of eigenvalues [28]. Therefore, we have provided the alternative proof of this result that has already been obtained in [29].

Lemma 2 *If $\lambda_{max}(\rho)$ denotes the maximum eigenvalue of $d \otimes d$ dimensional quantum state ρ, then*

$$\frac{1}{d^2} \leq \lambda_{max}(\rho) \leq 1 \tag{26}$$

Let us consider the noisy singlet state of the form

$$\rho_p = p|\phi^+\rangle\langle\phi^+| + (1-p)\frac{I \otimes I}{d^2}, 0 \leq p \leq 1 \tag{27}$$

where $|\phi^+\rangle = \frac{1}{\sqrt{d}} \sum_{i=0}^{d-1} |ii\rangle$.

The maximum eigenvalue of the density matrix ρ_p is given by

$$\begin{aligned} \lambda_{max}(\rho_p) &= \lambda_{max}[p|\phi^+\rangle\langle\phi^+| + (1-p)\frac{I \otimes I}{d^2}], 0 \leq p \leq 1 \\ &= p + \frac{1-p}{d^2} \end{aligned} \tag{28}$$

Now, since it is known that the state ρ_p is separable if and only if $0 \leq p \leq \frac{1}{d+1}$ [30] so we can say that the state ρ_p is separable if and only if $\lambda_{max}(\rho_p) \leq \frac{1}{d}$.

Therefore, we can state the following theorem.

Theorem *An arbitrary $d \otimes d$ dimensional isotropic quantum state ρ_p shared between two distant partners is separable if and only if*

$$\lambda_{max}(\rho_p) \leq \frac{1}{d} \tag{29}$$

It should be noted that the above condition is necessary and sufficient condition only for isotropic state.

Corollary 1 *An entangled quantum state ρ shared between two distant partners is useful for quantum teleportation if*

$$\lambda_{max}(\rho) > \frac{1}{d} \tag{30}$$

Corollary 2 *The upper bound of the maximum achievable teleportation fidelity from a given bipartite state ρ in $d \otimes d$ dimensional Hilbert space is given by*

$$f^{tel}(\rho) \leq \frac{\lambda_{max}(\rho)d + 1}{d+1} \tag{31}$$

4.1 Examples

We are now in a position to discuss a few examples in which our criterion detects mixed entangled states useful in quantum teleportation while singlet fraction failed to do so. In this, we will study a few quantum states ρ for which singlet fraction $F_{max}(\rho)$ and maximum

eigenvalue $\lambda_{max}(\rho)$ satisfies the inequality

$$F_{max}(\rho) \leq \frac{1}{d} < \lambda_{max}(\rho) \tag{32}$$

When The left part of the inequality (32) holds, then We are uncertain about the usefulness of state ρ for a resource state in quantum teleportation except for d=2, where we can affirmatively say that state ρ is not useful for a resource state in quantum teleportation. If the right part of the inequality (32) holds true, then we can say that the state ρ can be used in teleportation. This means that when singlet fraction is unable to detect the useful states for quantum teleportation, then the maximum eigenvalue can serve the purpose. Also we should note that for the above quantum state ρ, interaction with the environment is not taken into account.

4.1.1 Example 1

Let us now recall the quantum state described by the density operator $\sigma_{12}^{(1)}$ given in (20). Firstly, we need to check whether the state is entangled. Let us consider a witness operator of the form [31]

$$W_1 = \frac{1}{4}(I \otimes I + I \otimes \sigma_z + \sigma_z \otimes I - \sigma_x \otimes \sigma_x + \sigma_y \otimes \sigma_y) \tag{33}$$

where $\sigma_x, \sigma_y, \sigma_z$ are denoting pauli matrices.

The expectation value of the operator W_1 with respect to the state $\sigma_{12}^{(1)}$ is given by

$$Tr(W_1\sigma_{12}^{(1)}) = -0.55 \tag{34}$$

Since W_1 is a witness operator and the expectation value is negative for the state $\sigma_{12}^{(1)}$ so the state $\sigma_{12}^{(1)}$ is an entangled state. Now we are in a position to say whether the entangled state is useful for teleportation by calculating its maximum eigenvalue. The maximum eigenvalue of the state is $\lambda_{max}(\sigma_{12}^{(1)}) = 0.587 > \frac{1}{2}$. Thus the state is useful in quantum teleportation according to corollary 1. This example is important in the sense that the value of the singlet fraction (after applying SPA-PT operation) is unable to detect the state for possible application of the state $\sigma_{12}^{(1)}$ as a resource state in quantum teleportation but on the other hand, corollary 1 help us to reach the correct conclusion.

4.1.2 Example 2

Let us consider a quantum state described by the density matrix ρ_1, which is given by [8]

$$\rho_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{3-2\sqrt{2}}{2} & \frac{1-\sqrt{2}}{2} & 0 \\ 0 & \frac{1-\sqrt{2}}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \sqrt{2}-1 \end{pmatrix} \tag{35}$$

To detect whether the state ρ_1 is entangled, let us construct a witness operator as [31]

$$W_2 = \frac{1}{4}(I \otimes I + \sigma_z \otimes I + \sigma_x \otimes \sigma_x + \sigma_y \otimes \sigma_y) \tag{36}$$

where $\sigma_i's (i = x, y, z)$ are pauli matrices.

$Tr(W_2\rho_1)$ can be calculated as

$$Tr(W_2\rho_1) = -1.406 \tag{37}$$

The negative value of $Tr(W_2\rho_1)$ indicates that the state ρ_1 is entangled. Also it can be calculated that the singlet fraction of ρ_1 is $\frac{1}{2}$. Since $F^{max}(\rho_1) = \frac{1}{2}$ so it can be concluded that the state ρ_1 is not useful as a resource state for teleportation. But it is known that all entangled two-qubit mixed states are useful for teleportation. Hence, the inference from singlet fraction that the state ρ_1 is not useful as a resource state for teleportation is not correct. Let us now apply corollary 1 to detect whether the state ρ_1 is useful for teleportation. To this end, let us calculate the eigenvalues of ρ_1 and they are given by {0.5858,0.4142,0,0}. The maximum eigenvalue is found out to be $\lambda_{max}(\rho_1) = 0.5858$. Since $\lambda_{max}(\rho_1) > 1/2$, we can conclude that the state ρ_1 is useful for teleportation.

4.1.3 Example 3

Let us take another quantum state from $3 \otimes 3$ dimensional Hilbert space described by the density matrix ρ_2

$$\rho_2 = \begin{pmatrix} \frac{1-a}{2} & 0 & 0 & 0 & 0 & 0 & 0 & -0.22 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2}-a & -0.22 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0.22 & a & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -0.22 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{a}{2} \end{pmatrix} \tag{38}$$

where $0.35 \leq a \leq 0.369$.

The witness operator that detect the state described by the density operator ρ_2 is given by [32]

$$\begin{aligned} W_3 = & -\frac{3}{2}A_3 \otimes A_3 + \frac{3}{2}S_3 \otimes S_3 - \frac{2}{\sqrt{3}}D_2 \otimes D_1 + \frac{2}{3}I \otimes D_1 \\ & -\frac{1}{2}D_1 \otimes D_1 + \frac{5}{6}D_2 \otimes D_2 + \frac{1}{2}iA_1 \otimes S_1 + \frac{1}{2}iS_1 \otimes A_1 \\ & -\frac{1}{2}S_2 \otimes S_2 + \frac{1}{2}A_2 \otimes A_2 + \frac{2}{3}D_1 \otimes I - \frac{2}{3}I \otimes I \\ & +\frac{2\sqrt{3}}{9}D_2 \otimes I - \frac{2}{\sqrt{3}}D_1 \otimes D_2 + \frac{2\sqrt{3}}{9}I \otimes D_2 \end{aligned} \tag{39}$$

where S_1, S_2 and S_3 are three symmetric Gell-Mann matrices, A_1, A_2 and A_3 are three anti-symmetric Gell-Mann matrices and D_1 and D_2 are two diagonal Gell-Mann matrices given by

$$S_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, S_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, S_3 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \tag{40}$$

$$A_1 = \begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, A_2 = \begin{pmatrix} 0 & 0 & -i \\ 0 & 0 & 0 \\ i & 0 & 0 \end{pmatrix}, A_3 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix} \tag{41}$$

$$D_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, D_2 = \begin{pmatrix} \frac{1}{\sqrt{3}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{3}} & 0 \\ 0 & 0 & \frac{-2}{\sqrt{3}} \end{pmatrix} \tag{42}$$

We note that $Tr(W_3\rho_2) = 0.44 - 3a$, where $0.35 \leq a \leq 0.369$. Thus, $Tr(W_3\rho_2) < 0$ for $0.35 \leq a \leq 0.369$. Hence the state ρ_2 is an entangled state.

Let us calculate the singlet fraction of ρ_2. To do this, we need maximally entangled basis states in $3 \otimes 3$ dimensional Hilbert space. The maximally entangled basis for two-qutrit system is given by [33]

$$\begin{aligned}
|B_0\rangle &= \frac{1}{\sqrt{3}}[|00\rangle + |22\rangle - e^{i\frac{\pi}{3}}|11\rangle] \\
|B_1\rangle &= \frac{1}{\sqrt{3}}[|01\rangle + |20\rangle - e^{i\frac{\pi}{3}}|12\rangle] \\
|B_2\rangle &= \frac{1}{\sqrt{3}}[|02\rangle + |21\rangle - e^{i\frac{\pi}{3}}|10\rangle] \\
|B_3\rangle &= \frac{1}{\sqrt{3}}[|11\rangle + |00\rangle - e^{i\frac{\pi}{3}}|22\rangle] \\
|B_4\rangle &= \frac{1}{\sqrt{3}}[|12\rangle + |01\rangle - e^{i\frac{\pi}{3}}|20\rangle] \\
|B_5\rangle &= \frac{1}{\sqrt{3}}[|10\rangle + |02\rangle - e^{i\frac{\pi}{3}}|21\rangle] \\
|B_6\rangle &= \frac{1}{\sqrt{3}}[|11\rangle + |22\rangle - e^{i\frac{\pi}{3}}|00\rangle] \\
|B_7\rangle &= \frac{1}{\sqrt{3}}[|20\rangle + |12\rangle - e^{i\frac{\pi}{3}}|01\rangle] \\
|B_8\rangle &= \frac{1}{\sqrt{3}}[|21\rangle + |10\rangle - e^{i\frac{\pi}{3}}|02\rangle]
\end{aligned} \tag{43}$$

Then the singlet fraction of ρ_2 can be calculated using the maximally entangled basis (43) as

$$\begin{aligned}
F^{max}(\rho_2) &= max_{B_i}\langle B_i|\rho_2|B_i\rangle, i = 0, 1, ..., 8 \\
&= \frac{1.22 - a}{3}
\end{aligned} \tag{44}$$

Figure 2 shows that $F^{max}(\rho_2)$ decreases as the state parameter a increases. The singlet fraction $F^{max}(\rho_2)$ is always less than $\frac{1}{3}$ when the state parameter a lying in the interval [0.35,0.369]. Therefore, according to singlet fraction criterion, the state described by the density operator ρ_2 may or may not be useful in quantum teleportation.

Let us now calculate the eigenvalues of ρ_2. The maximum eigenvalue of ρ_2 is given by

$$\begin{aligned}
\lambda_{max}(\rho_2) &= \frac{1}{4} + \frac{1}{2}\sqrt{0.4436 - 2a + 4a^2}, \\
&0.35 \leq a \leq 0.369
\end{aligned} \tag{45}$$

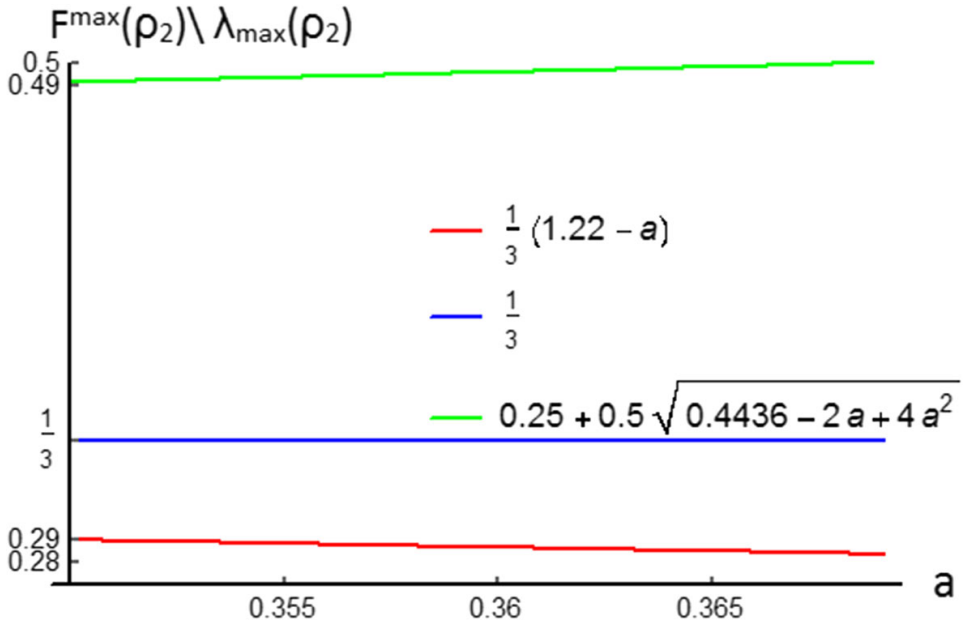


Fig. 2 Plot of singlet fraction $(F^{max}(\rho_2))/(\lambda_{max}(\rho_2))$ versus the state parameter a

We have also shown in Fig. 2 that $\lambda_{max}(\rho_2)$ is always greater than $\frac{1}{3}$ when $a \in [0.35, 0.369]$. Thus, our criterion detect that the state ρ_2 is useful in quantum teleportation.

5 Teleportation Criteria in Terms of Upper Bound of the Maximum Eigenvalue in Dembo's Bound

In this section, we will study those cases where maximum eigenvalue criterion for teleportation fails i.e. the case where maximum eigenvalue of a given quantum state satisfies the inequality

$$\lambda_{max}(\rho) \leq \frac{1}{d} \tag{46}$$

To overcome this problem, we have provided another criterion which is based on Dembo's bound to detect whether the state useful for teleportation or not.

Let us consider a qutrit-qutrit system described by the density operator

$$\rho_3 = \begin{pmatrix} \frac{a}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.015 \\ 0 & \frac{a}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1-a}{2} & 0 \\ 0.015 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1-a}{2} \end{pmatrix}, 0.5 \leq a \leq 0.65 \tag{47}$$

We find that $Tr(W_3\rho_3) = -1.53 + 2a < 0$, for $0.5 \leq a \leq 0.65$. Thus, the witness operator W_3 detect the state ρ_3 as an entangled state. But the question is whether the entangled state ρ_3 is useful in quantum teleportation.

Eigenvalues of ρ_3 are given by: $[0, 0, 0, 0, 0, \frac{1-a}{2}, \frac{a}{2}, 0.125(2-\sqrt{16a^2-16a+4.0144}), 0.125(2+\sqrt{16a^2-16a+4.0144})]$. The maximum eigenvalue is given by $\lambda_{max}(\rho_3) = 0.125(2+\sqrt{16a^2-16a+4.0144})$. We can observe that $\lambda_{max}(\rho_3) \leq \frac{1}{3}$ When $0.5 \leq a \leq 0.65$. Therefore, our maximum eigenvalue criterion fails to detect whether the state ρ_3 is useful in teleportation. It motivate us to search for maximal bound of maximum eigenvalue that can be greater than $\frac{1}{d}$ for $d \otimes d$ dimensional system.

To start with our search, let us consider the upper bound of maximal eigenvalue of the $d \otimes d$ dimensional quantum state ρ under investigation. The upper bound may be denoted as $\lambda^D_{max}(\rho)$ and it is given by R.H.S of the inequality (8)

$$\lambda^D_{max}(\rho) = \frac{c+\eta_{d^2-1}}{2} + \sqrt{\frac{(c-\eta_{d^2-1})^2}{2} + (b^*)^T b} \tag{48}$$

where $R_{d^2} = \begin{pmatrix} R_{d^2-1} & b \\ (b^*)^T & c \end{pmatrix}$, η_1 is the lower bound on minimal eigenvalue of R_{d^2-1}, η_{d^2-1} is the upper bound on maximal eigenvalue of R_{d^2-1} and b is a vector of dimension d^2-1. We are now in a position to provide another criterion in terms of upper bound of maximal eigenvalue of the $d \otimes d$ dimensional quantum state ρ.

Corollary 3 *An arbitrary $d \otimes d$ dimensional mixed NPT entangled state ρ shared between two distant partners is useful as a resource state in quantum teleportation if*

$$\lambda^D_{max}(\rho) > \frac{1}{d} \tag{49}$$

5.1 Example 4

Recalling, the qutrit-qutrit system described by the density operator ρ_3 given in (47). The maximal bound of maximum eigenvalue of $3 \otimes 3$ dimensional density matrix ρ_3 is given by $\lambda^D_{max}(\rho_3)$. For the 9×9 order density matrix ρ_3, we have $c = 0.1750$, $(b^*)^T = \begin{pmatrix} 0.015 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$ and $\eta_8 = 0.325$. Using the above values, we find that $\lambda^D_{max}(\rho_3) = 0.357$. Thus using corollary 3, we are able to show that the state ρ_3 is useful in teleportation.

5.2 Example 5

Let us consider another qutrit-qutrit NPT entangled state, which is given by the density matrix [34]

$$\rho_\alpha = \frac{2}{7}|\phi_3^+\rangle\langle\phi_3^+| + \frac{\alpha}{7}\sigma_+ + \frac{5-\alpha}{7}\sigma_-, 4 < \alpha \leq 5 \tag{50}$$

where $|\phi_3^+\rangle = \frac{1}{\sqrt{3}}\sum_{i=0}^{2}|ii\rangle$, $\sigma_+ = \frac{1}{3}(|01\rangle\langle 01| + |12\rangle\langle 12| + |20\rangle\langle 20|)$, $\sigma_- = \frac{1}{3}(|10\rangle\langle 10| + |21\rangle\langle 21| + |02\rangle\langle 02|)$. For the density matrix ρ_α, we have $c = \frac{2}{21}$, $(b^*)^T = \begin{pmatrix} \frac{2}{21} & 0 & 0 & 0 & \frac{2}{21} & 0 & 0 & 0 \end{pmatrix}$ and $\eta_8 = \frac{5}{21}$. In this case, $\lambda^D_{max}(\rho_\alpha) = 0.3346$, which is greater than $\frac{1}{3}$. Therefore, in this case also we can use corollary 3 to conclude that the state described by the density operator ρ_α is useful in quantum teleportation.

6 Conclusion

To summarize, we have modified the relationship between the optimal singlet fraction and partial transpose of a given state. The modification is required because partial transposition is a non-physical operation and thus it cannot be implemented in the laboratory. We use SPA-PT method to overcome this difficulty and have shown that the modified value of optimal singlet fraction can be estimated in an experiment. Further, we have proposed two criteria based on maximum eigenvalue of the given state and have studied our criteria for the detection of $d \otimes d$ dimensional NPT entangled states useful in quantum teleportation in the given two cases: (i) $F_{max}(\rho) \leq \frac{1}{d} < \lambda_{max}(\rho)$ or (ii) $\lambda_{max}(\rho) \leq \frac{1}{d} < \lambda^{D}_{max}(\rho)$. Our criteria can be realized in an experiment because maximum eigenvalue can be estimated experimentally [13, 35, 36].

Acknowledgments A. G. would like to acknowledge the financial support from CSIR. This work is supported by CSIR File No. 08/133(0035)/2019-EMR-1.

References

1. Nielsen, M.A., Chuang, I.L.: Quantum Computation and Quantum Information. Cambridge University Press, Cambridge (2000)
2. Wilde, M.M.: Quantum Information Theory. Cambridge University Press, Cambridge (2013)
3. Bennett, C.H., Brassard, G., Crepeau, C., Jozsa, R., Peres, A., Wootters, W.K.: Phys. Rev. Lett. **70**, 1895 (1993)
4. Briegel, H.J., dur, W., Cirac, J.I., Zoller, P.: Phys. Rev. Lett. **81**, 5932 (1998)
5. Gottesman, D., Chuang, I.L.: Nature **402**, 390 (1999)
6. Ishizaka, S., Hiroshima, T.: Phys. Rev. Lett. **101**, 240501 (2008)
7. Horodecki, M., Horodecki, P., Horodecki, R.: Phys. Rev. A **60**, 1888 (1999)
8. Badziag, P., Horodecki, M., Horodecki, P., Horodecki, R.: Phys. Rev. A **62**, 012311 (2000)
9. Verstraete, F., Verschelde, H.: Phy. Rev. Lett. **90**, 097901 (2003)
10. Albeverio, S., Fei, S.M., Yang, W.L.: Phys. Rev. A **66**, 012301 (2002)
11. Zhao, M.J., Li, Z.G., Fei, S.M., Wang, Z.X., Jost, X.L.: J. Phys. A: Math. Theor. **44**, 215302 (2011)
12. Bowen, G., Bose, S., Bose, S.: Phys. Rev. Lett. **87**, 267901 (2001)
13. Ekert, A.K., Alves, C.M., Oi, D.K.L., Horodecki, M., Horodecki, P., Kwek, L.C.: Phys. Rev. Lett. **88**, 217901 (2002)
14. Lasserre, J.B.: IEEE Trans. on Automatic Control **40**, 1500 (1995)
15. Kumari, A., Adhikari, S.: Phys. Rev. A **100**, 052323 (2019)
16. Liang, Y.-C., Yeh, Y.-H., Mendonca, P.E.M.F., Teh, R.Y., Reid, M.D., Drummond, P.D.: Rep. Prog. Phys. **82**, 076001 (2019)
17. Bae, J.: Rep. Prog. Phys. **80**, 104001 (2017)
18. Adhikari, S.: Phys. Rev. A **97**, 042344 (2018)
19. Lim, H.-T., Kim, Y.-S., Ra, Y.-S., Bae, J., Kim, Y.-H.: Phys. Rev. Lett. **107**, 160401 (2011)
20. Dembo, A.: IEEE Trans. Info. Th. **34**, 352 (1988)
21. Park, D., Lee, B.G.: IEEE Signal Process. Lett. **10**, 267 (2003)
22. Bruss, D., Macchiavello, C.: Found. Phys. **33**, 1617 (2003)
23. Wootters, W.K.: Phys. Rev. Lett. **80**, 2245 (1998)
24. Connor, K.M.O., Wootters, W.K.: Phys. Rev. A **63**, 052302 (2001)
25. Verstraete, F., audenaert, K., Moor, B.D.: Phys. Rev. A **64**, 012316 (2001)
26. Munro, W.J., James, D.F.V., White, A.G., Kwiat, P.G.: Phys. Rev. A **64**, 030302 (2001)
27. Ishizaka, S., Hiroshima, T.: Phys. Rev. A **62**, 022310 (2000)
28. Ziman, M., Stelmachovic, P., Buzek, V.: Fortschr. Phys. **49**, 1123 (2001)
29. Juan, G.R., Ming, L., Ming, F., Qing, L.: Commun. Theor. Phys. **53**, 265 (2010)
30. Horodecki, M., Horodecki, P.: Phys. Rev. A **59**, 4206 (1999)

31. Riccardi, A., Chruscinski, D., Macchiavello, C.: Phys. Rev. A **101**, 062319 (2020)
32. Jafarizadeh, M.A., Behzadi, N.: Eur. Phys. J. D. **55**, 729 (2009)
33. Karimipour, V., Memarzadeh, L.: Phys. Rev. A **62**, 012311 (2000)
34. Horodecki, P., Horodecki, M., Horodecki, R.: Phys. Rev. Lett. **82**, 1056 (1999)
35. Keyl, M., Werner, R.F.: Phys. Rev. A **64**, 052311 (2001)
36. Tanaka, T., Ota, Y., Kanazawa, M., Kimura, G., Nakazato, H., Nori, F.: Phys. Rev. A **89**, 012117 (2014)

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

2020 IEEE International Conference for Innovation in Technology (INOCON)
Bengaluru, India. Nov 6-8, 2020

Universal Biquad Fractional order Filter Using Single LT1228 IC

Gopal Singh
Department of Electrical Engineering
Delhi Technological University
Delhi, India
gopal_mt2k18@dtu.ac.in

Garima
Department of Electrical Engineering
Delhi Technological University
Delhi, India
garima.dtu@gmail.com

***Abstract*— In this paper, a voltage mode (VM) fractional order filter, capable of providing fractional low pass filter response (FLPF), fractional high pass filter response (FHPF), fractional band pass filter response (FBPF) and fractional band reject response (FBRF) along with tunability has been presented. The modified circuit employs a single commercially available Integrated Circuit, namely, LT1228 along with two fractional order capacitors (designed using Oustaloup approximation). PSPICE simulation results of all the filters mentioned above have also been presented.**

Keywords— LT1228, fractional order capacitor, fractional order filter, multi input single output

I. INTRODUCTION

In integer order analog filters available in open literature [1-3], the transition from pass band to stop band depends on integer multiples of a fixed slope (±20n decibels/decade). This usually results in the order of a filter being more than the minimum required to realize a given attenuation characteristics. In fractional order filters on the other hand, an attenuation of 20(n+α) where α varies from 0 to 1, provides more precise control over transition from pass band to stop band because of additional degree of freedom 'α' provided by the fractional order 'α' in the filter transfer function. With the availability of some fractional order immittances (both fractional order capacitors and fractional order inductors) as experimental prototypes, the possibility of these fractional order immittances becoming available as standard passive elements is ever increasing. As a result, research work on fractional order analog circuits and other related topics [4-17] has accelerated during the last decades.

Fractional order filters have been realized using different active elements like operational amplifier [4-9], current conveyors [10], current feedback followers [11], operational transconductance amplifiers [12] and various other active elements. Since the fractional order capacitor of any non-integer order is not available commercially [13-14], most of the paper presented on fractional order circuits use fractional order capacitor, designed using some integer order approximation of the term s^{α} resulting in a Foster like RC network approximating the fractional order capacitor. Out of the various filters realized with different active elements, filters realized with commercially available IC LT1228 are more versatile as they can be easily implemented using current feedback amplifier (CFA) and operational transconductance amplifier (OTA). A detailed review of filters realized using LT1228 IC [15-17] reveals the following:

(i) Single input multiple output (SIMO) type second order filter utilizing two LT1228 IC and two OTAs have been presented [15].

(ii) Single input single output (SISO) type first order all pass filter utilizing single LT1228 IC structure [16].

(iii) Multi input single output (MISO) type universal second order filter using single LT1228 IC structure [17].

From the above discussion, it thus emerges that, very little work has been done on realization of voltage mode (VM) multi input single output (MISO) fractional order filter using single LT1228 IC. These MISO type of structures, are very useful to get different responses without disturbing the structure. In this paper, thus, we have, generalized the design of an existing MISO type VM biquad filter [1] realized with LT1228 IC by replacing the integer order capacitors with fractional order capacitors and examined the tunability of the different parameters of the realized filters with the fractional order parameter 'α' and the bias current I_{bias}. Also, the stability of the realized filters has been studied in detail.

II. CIRCUIT DESCRIPTION

The LT1228 IC schematic symbol along with its equivalent circuit is shown in fig. 1.

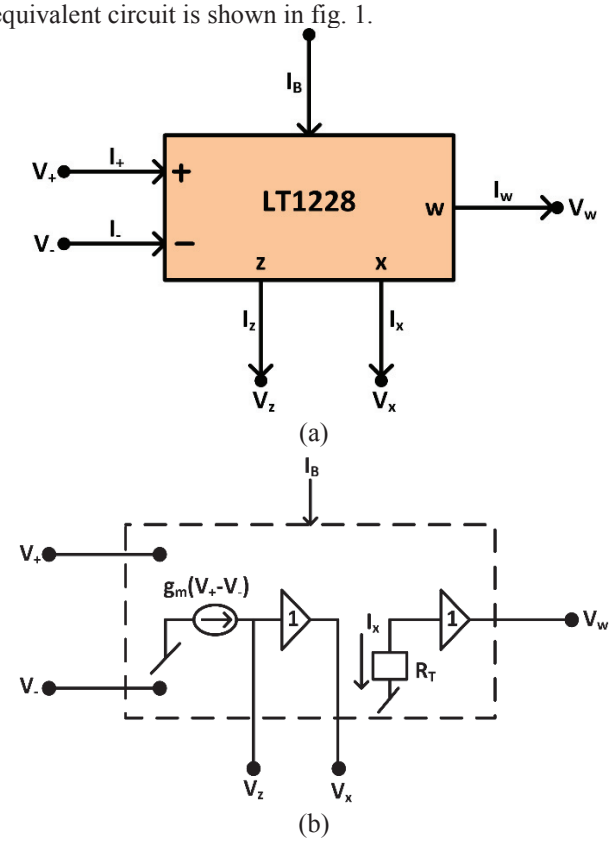


Figure 1 a) Schematic symbol b) Equivalent circuit of LT1228 IC

Ideally, its characteristics equation can be described as (1):

978-1-7281-9744-9/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:05:40 UTC from IEEE Xplore. Restrictions apply.

$$\begin{bmatrix} I_+ \\ I_- \\ I_z \\ V_x \\ V_w \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ g_m & -g_m & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & R_T & 0 \end{bmatrix} \begin{bmatrix} V_+ \\ V_- \\ V_z \\ I_x \\ I_w \end{bmatrix} \quad (1)$$

where V_x, V_w, V_+, V_- and V_z are the output and input port voltages, I_{V+}, I_{V-}, I_Z, I_x and I_w are the output and input port currents, g_m is the transconductance gain whose relation with I_{bias} is given in equation (2) and R_T is the transresistance gain which is ideally very high. Also, impedances at terminals V_+, V_- and Z are high while impedances at terminals X and W are low.

LT1228 is a commercially available IC manufactured by Linear Technology Inc. [18]. It is a combination of CFA and OTA which has an external bias current to control its gain externally (as given in equation (2)) which is further used for tuning the filter responses.

$$g_m = 10 I_{bias} \quad (2)$$

where I_{bias} is the external bias current.

The modified filter circuit, obtained by replacing the integer order capacitors of the filter circuit presented in [1] with fractional order capacitor (C_α) whose driving point impedance is given by $1/Cs^\alpha$, is shown in fig. 2. Like the integer order filter circuit presented in [1], which belongs to the class of multi input single output (MISO) type of configuration, the proposed circuit also belongs to the MISO class, wherein by selecting appropriate input signal, various filter responses along with electronic tunability can be obtained.

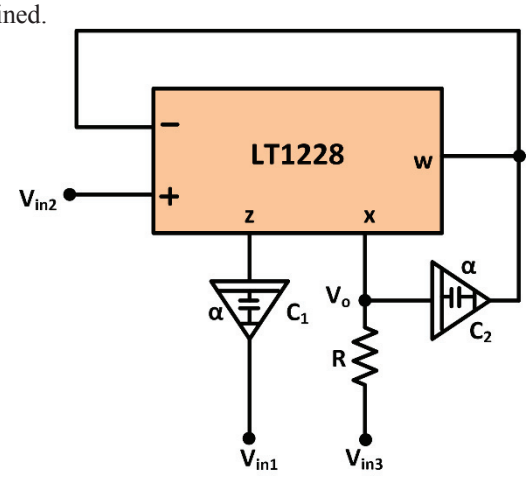


Figure 2 Modified filter circuit

Analyzing the fractional filter circuit as shown above the following output function is obtained:

$$V_0(s) = \frac{s^{2\alpha} V_{in1}(s) + s^{\alpha}(g_m/C_1)V_{in2}(s) + (g_m/C_1C_2R)V_{in3}(s)}{D(s)} \quad (3)$$

where $D(s) = s^{2\alpha} + (g_m/C_1) + (g_m/C_1C_2R)$ (4)

For any fractional order filter, with transfer function T(s), the various important parameters need to be determined [4].

A. Fractional low pass filter (FLPF)

By selecting $V_{in} = V_{in3}$ and $V_{in1} = V_{in2} = 0$ we get the required FLPF transfer function:

$$\frac{V_0(s)}{V_{in}(s)} = \frac{(g_m/C_1C_2R)}{D(s)} \quad (5)$$

The DC gain of FLPF is 1 while the high frequency gain is 0. Its cutoff frequency, maximum frequency and right phase frequency for any order alpha can be obtained by equations (5a), (5b) and (5c).

$$2X^3 + 3\cos\left(\frac{\alpha\pi}{2}\right)X^2 + 2[1 + k\cos(\alpha\pi)]X + k\cos\left(\frac{\alpha\pi}{2}\right) = 0 \quad (5a)$$

$$Y^4 + 2\cos\left(\frac{\alpha\pi}{2}\right)Y^3 + (1 + 2k\cos(\alpha\pi)Y^2 + 2k\cos\left(\frac{\alpha\pi}{2}\right)Y - k^2 = 0 \quad (5b)$$

$$\omega_{rp} = \left(\frac{-a\cos\left(\frac{\alpha\pi}{2}\right) - \sqrt{a^2\cos^2\left(\frac{\alpha\pi}{2}\right) - 4ab\cos(\alpha\pi)}}{2\cos(\alpha\pi)} \right)^{\frac{1}{\alpha}} \quad (5c)$$

where $X = \frac{\omega_m^\alpha}{a}, Y = \frac{\omega_h^\alpha}{a}, k = \frac{b}{a}, a = \frac{g_m}{C_1}$ and $b = \frac{1}{C_2R}$

B. Fractional high pass filter (FHPF)

If we select $V_{in} = V_{in1}$ and $V_{in2} = V_{in3} = 0$ we get the required FHPF transfer function:

$$\frac{V_0(s)}{V_{in}(s)} = \frac{s^{2\alpha}}{D(s)} \quad (6)$$

Similarly, the DC gain of FHPF is 0 while the high frequency gain is 1. The important frequency parameters for any order alpha can be calculated using relation (6a):

$$\omega_{mFLPF} \cdot \omega_{mFHPF} = \omega_{hFLPF} \cdot \omega_{hFHPF} = \omega_{rpFLPF} \cdot \omega_{rpFHPF} = (ab)^{\frac{1}{\alpha}} \quad (6a)$$

C. Fractional band pass filter (FBPF)

By selecting $V_{in} = V_{in2}$ and $V_{in1} = V_{in3} = 0$ we get the required FBPF transfer function:

$$\frac{V_0(s)}{V_{in}(s)} = \frac{s^{\alpha}(g_m/C_1)}{D(s)} \quad (7)$$

Both high frequency gain and DC gain is 0. For any order alpha, its 3dB frequency, maximum frequency and right phase frequency parameters can be calculated using equation (7a), (7b) and (7c).

$$\left(X^2 - k\right)\left(X^2 + \cos\left(\frac{\alpha\pi}{2}\right)X + k\right) \quad (7a)$$

$$Y^4 + 2\cos\left(\frac{\alpha\pi}{2}\right)Y^3 - Y^2 + 2k\cos(\alpha\pi)Y^2 - \left(8k\cos^2\left(\frac{\alpha\pi}{2}\right) + 8\sqrt{k}\cos\left(\frac{\alpha\pi}{2}\right)\right)Y^2 + 2k\cos(\alpha\pi)Y + k^2 = 0 \quad (7b)$$

$$\omega_{rp} = \left(\frac{-a \pm \sqrt{a^2 - 4ab\cos^2\left(\frac{\alpha\pi}{2}\right)}}{2\cos\left(\frac{\alpha\pi}{2}\right)} \right)^{\frac{1}{\alpha}} \quad (7c)$$

D. Fractional band reject filter (FBRF)

If we select $V_{in} = V_{in1} = V_{in3}$ and $V_{in2} = 0$ we get the required FBRF transfer function:

$$\frac{V_0(s)}{V_{in}(s)} = \frac{s^{2\alpha} + (g_m/C_1C_2R)}{D(s)} \quad (8)$$

Both DC gain and high frequency gain is 1. For any order alpha, its minimum frequency and 3dB frequency parameters can be determined using equation (8a) and (8b).

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:05:40 UTC from IEEE Xplore. Restrictions apply.

$$\left(X^2-k\right)\left(\begin{array}{l} X^4\cos\left(\frac{\alpha\pi}{2}\right)+X^3+\left[\begin{array}{l}4k\cos\left(\frac{\alpha\pi}{2}\right)\\ -2k\cos(\alpha\pi)\cdot\cos\left(\frac{\alpha\pi}{2}\right)\end{array}\right]X^2 \\ +kX+k^2\cos\left(\frac{\alpha\pi}{2}\right)\end{array}\right) \quad (8a)$$

$$Y^4-2\cos\left(\frac{\alpha\pi}{2}\right)Y^3-\left(1-2k\cos(\alpha\pi)\right)Y^2-2k\cos\left(\frac{\alpha\pi}{2}\right)Y+k^2=0 \quad (8b)$$

By putting alpha=1, all the important frequency parameter for different filters namely FLPF, FHPF, FBPF and FBRF have been given in Table 1. For values of α other than 1, these important parameters of different filter can be obtained using equation (5)-(8).

Table I Parameters when α=1

Types of filters	ω_m	ω_h	ω_{rp}
FLPF	$a\sqrt{(k-1)}$	$\left(a\sqrt{\left(k\pm\frac{\sqrt{8k^2-4k+1}}{2}-\frac{1}{2}\right)}\right)$	$\sqrt{ab}$
FHPF	$\frac{b}{\sqrt{(k-1)}}$	$\frac{ab}{\left(a\sqrt{\left(k\pm\frac{\sqrt{8k^2-4k+1}}{2}-\frac{1}{2}\right)}\right)}$	$\sqrt{ab}$
FBPF	$a\sqrt{k}$	$a\left(\frac{\sqrt{4k+1}}{2}\pm\frac{1}{2}\right)$	0,∞
FBRF	$a\sqrt{k}$	$a\left(\frac{\sqrt{4k+1}}{2}\pm\frac{1}{2}\right)$	$\sqrt{ab}$

III. Stability Analysis

Fractional order system stability has been studied in detail [19] and its stability graph is plotted in W-plane as shown in fig. 3. For any fractional order system, stability depends on the coeffficients of equation (4).

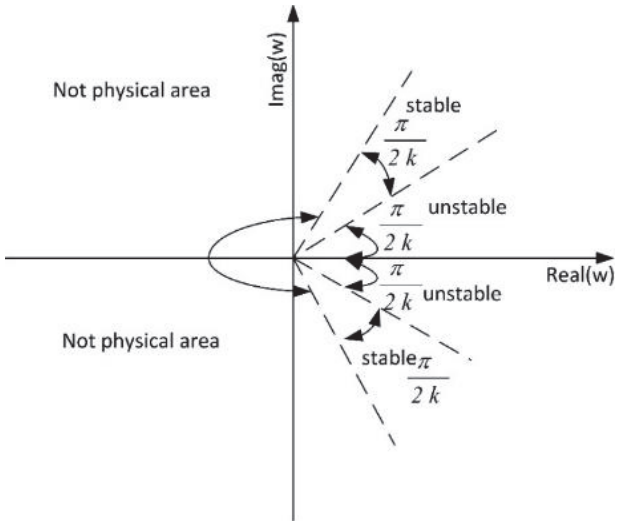


Figure 3 W-plane

In this paper, considering all the coefficients of equation positive, two different cases arises that are given in table II and has been adopted from [4]. Stability graph of different filters namely FLPF, FHPF, FBPF, and FBRF are shown in fig 4-7 respectively.

Table II Pole frequency (ω_0) and pole quality factor (Q) for different cases

Cases	Relations	Condition for stability and roots	ω_0, Q
1	$a \geq 4b$	$\alpha<2$, $r_{1,2}=\frac{-a\pm\sqrt{(a)^2-4(ab)}}{2}=g_{1,2}e^{j\pi}$	$\omega_{01,2}=g_{1,2}^{1/\alpha}$, $Q=\frac{-1}{2\cos(\pi/\alpha)}$
2	$a<4b$	$\alpha<\frac{2\delta}{\pi}$, $\delta=\cos^{-1}\left(\frac{-a}{2\sqrt{ab}}\right)>\frac{\pi}{2}$, $r_{1,2}=\sqrt{ab}e^{\pm j\delta}$	$\omega_0=\left(\sqrt{ab}\right)^{1/\alpha}$, $Q=\frac{-1}{2\cos(\delta/\alpha)}$

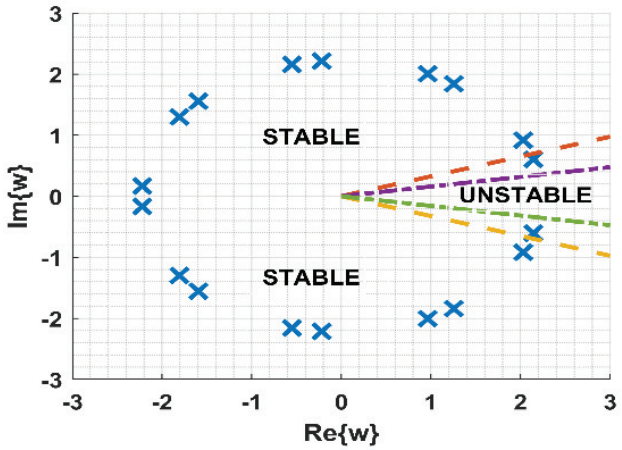


Figure 4 Stability plot of FLPF C=0.382μF and α=0.9

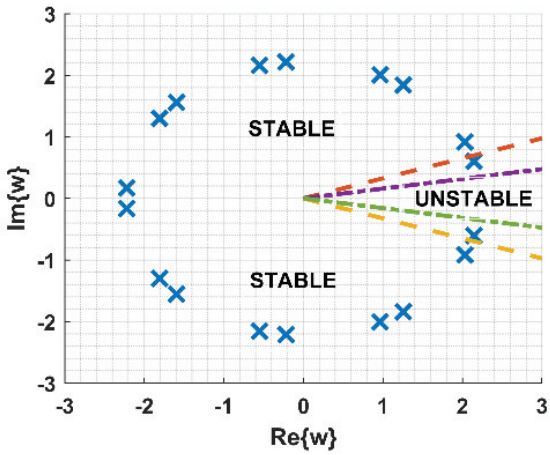


Figure 5 Stability plot of FHPF C=0.382μF and α=0.9

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:05:40 UTC from IEEE Xplore. Restrictions apply.

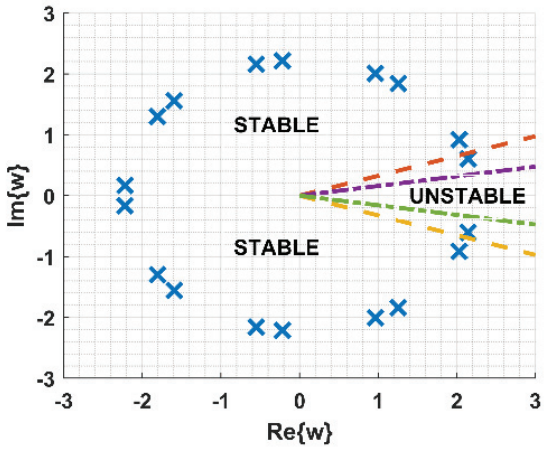


Figure 6 Stability plot of FBPF C=0.382μF and α=0.9

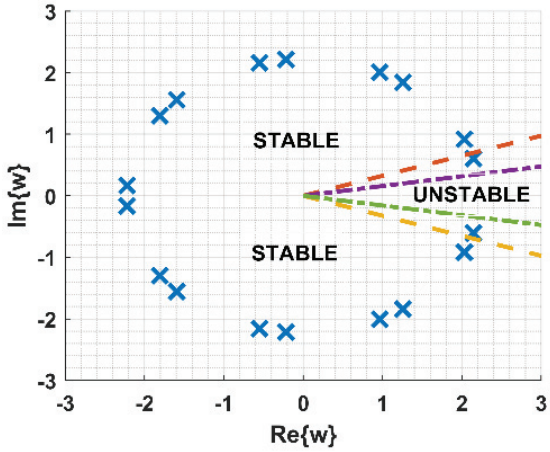


Figure 7 Stability plot of FBRF C=0.382μF and α=0.9

IV. Simulation Results

The workability of the modified circuit have been validated through PSPICE simulation using the macro model of IC LT1228. The DC power supply voltage is taken as ± 12v and value of passive component R_1=1kΩ. The simulation results obtained matches with theoretical analysis and it is further divided into two categories:

i. Tunability with α:

It can be achieved by replacing each ordinary capacitors with fractional order capacitors (C_1=C_2=0.382uF/(rad/sec)$^{(1-\alpha)}$) designed using oustaloup method for α=0.6, α=0.7, α=0.8, α=0.9 of order 8 resulting in foster type network as shown in fig 8.

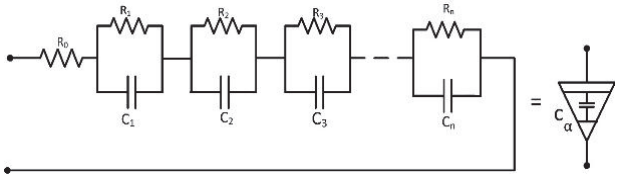


Figure 8 Fractional order Capacitor

All the filters designed in this section have values I_{bias}=0.226mA (R_{bias}=100kΩ) as shown in figure 9-12. The different frequency parameters for filters namely FLPF, FHPF, FBPF and FBRF for α=0.6 to α=1.0 have been listed in Table III. The difference between simulation values and theoretical values is very small and the error is as minimum as 4.4%.

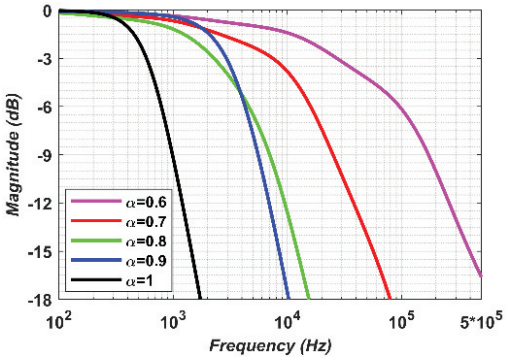


Figure 9 Magnitude response of FLPF for different α

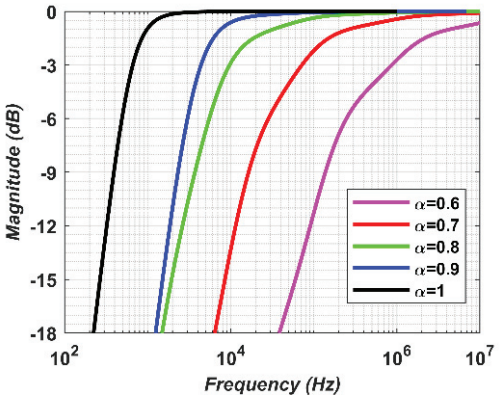


Figure 10 Magnitude response of FHPF for different α

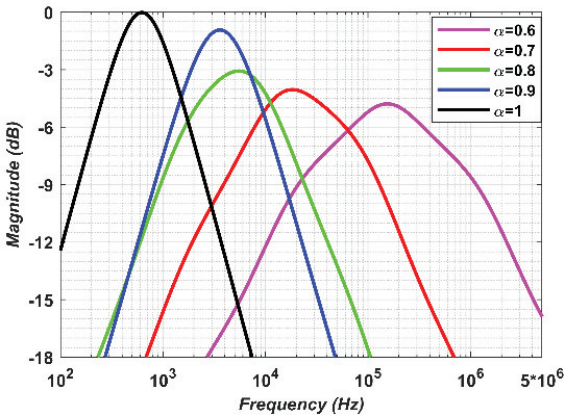


Figure 11 Magnitude response of FBPF for different α

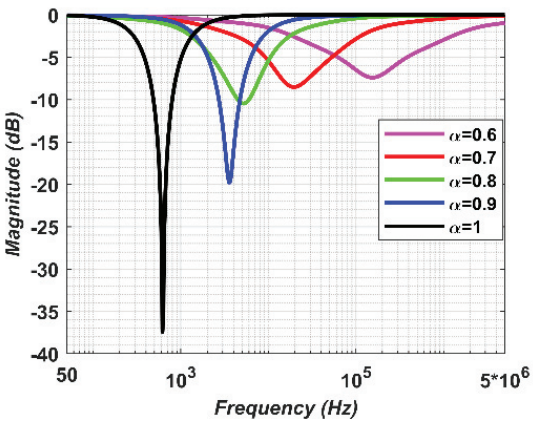


Figure 12 Magnitude response of FBRF for different α

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:05:40 UTC from IEEE Xplore. Restrictions apply.

Table III Cutoff frequency (f_h) and Maximum/Minimum frequency (f_m) for different fractional filters

	Cutoff Frequency (f_h) (in kHz)				Maximum/Minimum Frequency (f_m) (in kHz)			
α	**FLPF**		**FHPF**		**FBPF**		**FBRF**	
	PSPICE	Theoretical	PSPICE	Theoretical	PSPICE	Theoretical	PSPICE	Theoretical
0.6	28.857	30.244	841.954	807	154.882	156.3	154.882	156.3
0.7	7.285	7.01	70.58	68.030	20.73	21.77	20.82	21.77
0.8	2.380	2.45	9.635	10.124	5.219	5.01	5.248	5.01
0.9	1.15	1.108	2.13	2.23	1.54	1.572	1.55	1.572
1.0	0.585	0.587	0.659	0.668	616.6	0.626	0.616	0.626

ii. Tunablity with I_{bias}:

It can be achieved by changing the values of g_m and R by the same factor so that the ratio (g_m/R) remains constant which implies its cutoff frequency (ω_0) will remain same but quality factor (Q) changes while using standard capacitor as given in equation (10)-(11). All the filters designed in this section have values I_{bias1}=0.02825mA, I_{bias2}=0.0565mA, I_{bias3}=0.113mA, I_{bias4}=0.226mA and I_{bias5}=0.452mA respectively as shown in figure 13-16.

$$\omega_0 = \sqrt{\frac{g_m}{C_1 C_2 R}} \tag{10}$$

$$Q = \sqrt{\frac{C_1}{C_2 g_m R}} \tag{11}$$

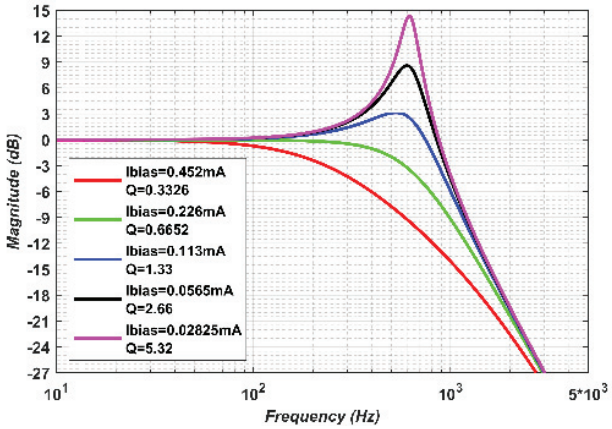


Figure 13 Magnitude response of FLPF for different I_{bias}

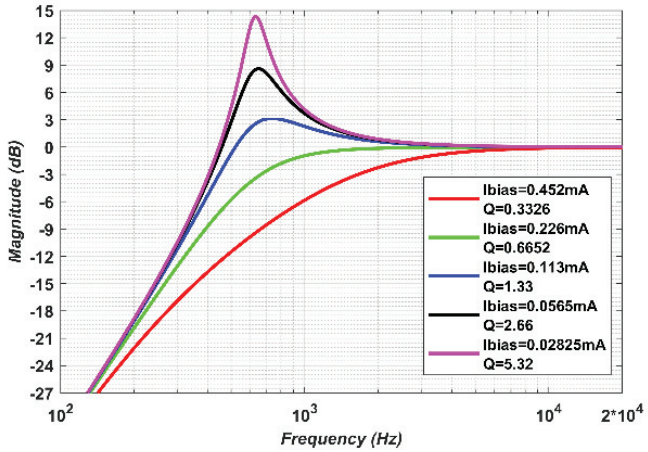


Figure 14 Magnitude response of FHPF for different I_{bias}

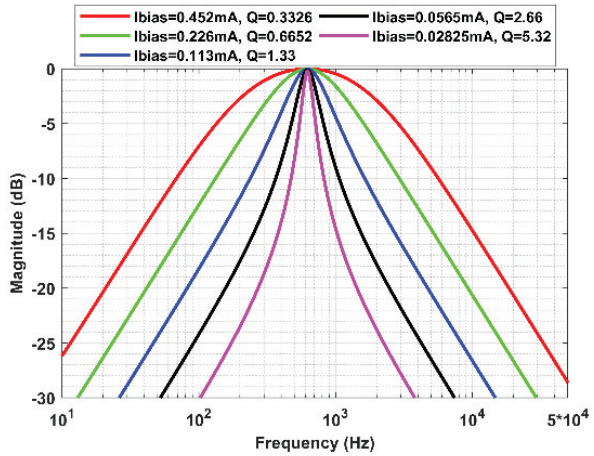


Figure 15 Magnitude response of FBPF for different I_{bias}

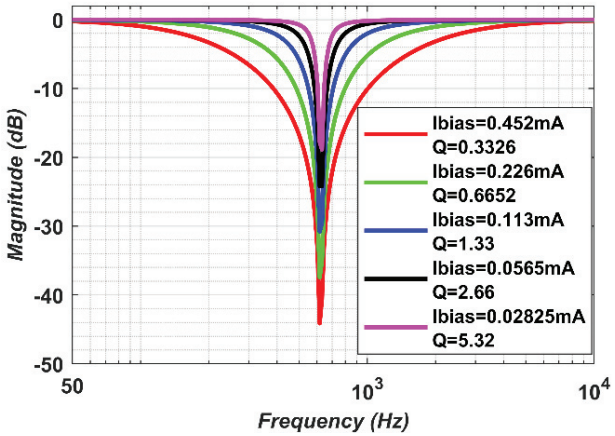


Figure 16 Magnitude response of FBRF for different I_{bias}

V. CONCLUSION

In our paper, the main focus is to develop different techniques for electronic tuning of different types of filter. The first method suggests us how to tune with different values of alpha and second method suggests us how to tune with different values of Ibias. Some of the constraints or motivation for future work are:

i. When fractional order system gains momentum, two different valued fractional order capacitors can be used in order to obtain more precisely tuning of parameters.

ii. While tuning with different values of I_{bias}, the quality factor is inversely proportional to product of g_m and R and cutoff frequency is directly proportional to (g_m/R) ratio for standard capacitor.

REFERENCES

[1] S. Klungtong, D. Thanapatay, and W. Jaikla, "Three-input single-output voltage-mode multifunction filter with electronic controllability based on single commercially available IC," Active and Passive Electronic Components, 2017.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:05:40 UTC from IEEE Xplore. Restrictions apply.

[2] C. M. Chang, B. M. Al-Hashimi, and J. N. Ross, "Unified active filter biquad structures," IEE Proceedings Circuits, Devices and Systems, 151, no. 4, 2004, pp. 273-277.
[3] S. A. Mahmoud, M. A. Hashiesh, and A. M. Soliman, "Low-voltage digitally controlled fully differential current conveyor," IEEE Transactions on Circuits and Systems I, Regular Papers, 52, no. 10, pp. 2055-2064, 2005.
[4] A. G. Radwan, A. S. Elwakil, and A. M. Soliman. "On the generalization of second-order filters to the fractional-order domain." Journal of Circuits, Systems, and Computers, 18, no. 02, pp. 361-386, 2009.
[5] A. S. Ali, A. G. Radwan, and A. M. Soliman, "Fractional order Butterworth filter: active and passive realizations," IEEE journal on emerging and selected topics in circuits and systems, 3, no. 3, pp. 346-354, 2013.
[6] A. G. Radwan, A. S. Elwakil, and A. M. Soliman, "Fractional-order sinusoidal oscillators: design procedure and practical examples," IEEE Transactions on Circuits and Systems I: Regular Papers, 55, no. 7, pp. 2051-2063, 2008.
[7] L. A. Said, A. G. Radwan, A. H. Madian, and A. M. Soliman, "Fractional order oscillator design based on two-port network," Circuits, Systems, and Signal Processing, 35, no. 9, pp. 3086-3112, 2016.
[8] 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, 34, no. 5, pp. 1461-1480, 2015.
[9] T. J. Freeborn, B. Maundy, and A. Elwakil. "Fractional-step Tow-Thomas biquad filters," Nonlinear Theory and Its Applications, IEICE, 3, no. 3, pp. 357-374, 2012.
[10] J. Koton, J. Jerabek, N. Herencsar, and D. Kubanek, "Current conveyors in current-mode circuits approximating fractional-order low-pass filter," In 2017 European Conference on Circuit Theory and Design (ECCTD), 2017, pp. 1-4.
[11] L. Langhammer, R. Sotner, J. Dvorak, J. Jerabek, and J. Polak. "Fully-differential tunable fractional-order filter with current followers and current amplifiers," In 2017 27th International Conference Radioelektronika (RADIOELEKTRONIKA), 2017, pp. 1-6.
[12] J. Jerabek, R. Sotner, J. Dvorak, L. Langhammer, and J. Koton, "Fractional-order high-pass filter with electronically adjustable parameters," In 2016 International Conference on Applied Electronics, 2016, pp. 111-116.
[13] A. Oustaloup, F. Levron, B. Mathieu, and F. M. Nanot, "Frequency-band complex noninteger differentiator: characterization and synthesis," IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications, 47, no. 1, pp. 25-39, 2000.
[14] J. Valsa, and J. Vlach "RC models of a constant phase element," International Journal of Circuit Theory and Applications, 41, no. 1, pp. 59-67, 2013.
[15] M. Olšák, and D. Biolek, "Controlled filters with LT1228 OTA amplifiers," In Proceedings of 23rd the International Conference Telecommunications and Signal Processing (TSP), 2000, pp. 191-194.
[16] A. Chaichana, S. Siripongdee, and W. Jaikla, "Electronically Adjustable Voltage-mode First-order Allpass Filter Using Single Commercially Available IC," In IOP Conference Series: Materials Science and Engineering, 559, no. 1, 2019.
[17] S. Siripongdee, and W. Jaikla, "Universal filter using single commercially available IC: LT1228," In MATEC Web of Conferences, 95, 2017.
[18] http://www.linear.com/product/LT1228.
[19] A. G. Radwan, A. M. Soliman, A. S. Elwakil, and A. Sedeek, "On the stability of linear systems with fractional-order elements," Chaos, Solitons & Fractals, 40, no. 5, pp. 2317-2328, 2009.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:05:40 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/2059-6014.htm

Unveiling contemporary dimensions in the internationalisation of family firms through bibliometric analysis and thematic analysis

Sonal Thukral and Apoorva Jain
Delhi School of Management, Delhi Technological University, Delhi, India

Received 25 September 2020
Revised 1 December 2020
Accepted 1 December 2020

Abstract

Purpose – For sustaining a competitive advantage in the integrated world economy, it has become imperative for family firms to internationalise their operations in overseas markets. However, despite the growing set of literature, results are still inconclusive with respect to family firms' internationalisation. Thus, this study aims to address this gap by systematically reviewing 142 articles (1991–2019) to help researchers in identifying and unfolding the unexplored themes in the underlying area.

Design/methodology/approach – For systematically reviewing articles, the study uses a three-step methodology following PRISMA guidelines, bibliometric analysis and thematic analysis. Descriptive statistics of 142 research articles are obtained through bibliometric analysis while thematic analysis is carried out to create themes or clusters of various factors relating to family firms' internationalisation.

Findings – The current review uncovers the evolving trends in the research streams, most productive authors, top journals and articles, co-citation analysis, as well as the major themes surrounding the family firms' internationalisation literature. Results from bibliometric analysis indicate that family firms' internationalisation is an upcoming research area. Also, the review indicates an opportunity for scholars from developing nations to make significant contributions in the underlying research stream.

Research limitations/implications – Results from bibliometric and thematic analysis will help academicians and researchers in accumulating a holistic understanding relating to family firms' internationalisation and understanding the upcoming trends in family firms' research, thereby guiding the future research scope. Also, it will assist the family firms' leaders and managers in understanding the important dynamics in overseas markets and various factors to be considered while planning their internationalisation.

Originality/value – Undertaking a systematic literature review presents readers with a state-of-the-art understanding of the underlying research topic. To the best of the knowledge, to date, the study is the first to conduct the review of literature through bibliometric analysis with the help of R Studio software in the field of family firms' internationalisation. Also, the study is the first to review more than 100 research articles in the underlying area. Finally, the study proposes a comprehensive framework integrating the major themes and facets relating to family firms' internationalisation.

Keywords Systematic review, Family firms, Bibliometric analysis, Internationalisation, PRISMA

Paper type Literature review

1. Introduction

Globally, China stands at the first position pertaining to the number and concentration of family-owned businesses, followed by the USA and India (Credit Suisse Research Institute, 2018). For family firms, sustaining competitive advantage in the integrated world economy has become all the more challenging. This in itself necessitates the need for family firms to

Review of International Business and Strategy
© Emerald Publishing Limited
2059-6014
DOI 10.1108/RIBS-09-2020-0121

explore the ways in which their competitiveness and uniqueness could be enhanced, one of the ways is by focussing on their *degree of internationalisation*. Owing to such market competitiveness, family firms have started exploring various foreign market options; e.g. majority of the businesses of *Godrej Group*, which is more than 120 years old Indian family firm, comes from marking its footprints at the global market places, with internationalisation as the focal point of its decision-making process. Past research studies on family firms' internationalisation render us with an inconclusive view about whether the peculiar family characteristics and resources, fosters or hinders their degree of internationalisation (Pukall and Calabro, 2014; Arregle *et al.*, 2017; De Massis *et al.*, 2018). Family firms are often faced with divergent objectives between their desires of being the conventional local producer on one side, while pursuing their internationalisation strategies on other hand, ultimately hindering their international expansion (Chrisman *et al.*, 2005). With respect to their internationalisation strategies, results are still inconclusive as to whether family firms prefer entering foreign markets through FDI (Pongelli *et al.*, 2019); or through indirect entry modes like exports (Kontinen and Ojala, 2010a; Monreal-Pérez and Sánchez-Marín, 2017); or whether they prefer to enter through joint ventures (Kuo *et al.*, 2012); or whether they exhibit a greater preference for wholly-owned subsidiaries (Kao and Kuo, 2017; Kuo *et al.*, 2012). Such differing internationalisation strategies are due to the heterogeneity residing within them with respect to various dimensions, namely, the extent of family involvement, family resources, risk orientation, etc., thus grabbing the attention of various researchers in recent years to understand the impact of such heterogeneity (Arregle *et al.*, 2012; Sanchez-Bueno and Usero, 2014; Gomez-Mejia *et al.*, 2010).

While internationalisation aids in the acquisition of overseas resources and customers, it also brings uncertainty to the business operations due to unfamiliar market dynamics, rendering it a mixed gamble (Alessandri *et al.*, 2018; Martin *et al.*, 2013). Extant literature has identified various factors, namely, family firms' long term orientation (Claver *et al.*, 2009); social capital and networks (Arregle *et al.*, 2007); family control (Zahra, 2003; Fernandez and Nieto, 2006); the presence of non-family managers (Banalieva and Eddleston, 2011); Socioemotional Wealth *(hereafter SEW)* (Berrone *et al.*, 2012) and managerial capabilities (Boeker and Karichalil, 2002), as either facilitating or restraining family firms' degree of internationalisation. Though, earlier family firms restricted their innovation operations to internal research and development activities, however, recent technological advancements are persuading them to undertake collaborative actions with their peers abroad for the acquisition of requisite resources, expertise and technology (Deloitte Family Business Center, 2018). Family-run *Harilela Group*, founded by six Indian brothers in 1959 has been able to manage its strong presence in both domestic and foreign markets even with the tight family control. This is mainly because the group placed greater emphasis on inculcating the up-to-date skills and technology in the organisation, as well as human resources by organising effective training programmes (Ward, 2006). Although, internationalisation aids in diversifying the overall risk of an organisation, however, family firms' desire to preserve SEW, family control and reputation (Jimenez, 2019); long term survival (Ray *et al.*, 2018); their high degree of risk aversion (Evert *et al.*, 2018) and their limited resource pool (Dou *et al.*, 2019), still prevents them from taking risky and uncertain decisions like internationalisation.

Drawing from such inconclusive results, the study aims to use the three-step methodology for systematically reviewing 142 articles, authored by 289 scholars which are published in 58 peer-reviewed journals in the past 29 years (from 1991–2019). With the ever-evolving and developing a set of literature on family firms' internationalisation, it becomes imperative to understand the already existing work to help researchers in identifying and

unfolding the unexplored themes in the underlying area. Also, with the advancement in the research community, researchers are now focussing on conducting interdisciplinary research work, thereby systematic review following PRISMA guidelines (Moher *et al.*, 2009) help us in reconciling the varying views of researchers in multiple disciplines. Further, descriptive statistics of 142 research articles are obtained through bibliometric analysis, conducted through *R Studio* software. Furthermore, thematic analysis is carried out to create themes or clusters of various factors relating to family firms' internationalisation. Finally, the study proposes a comprehensive framework integrating all the major themes and facets relating to family firms' internationalisation.

Our study contributes in the following ways:

- Reviewing the existing literature through thematic analysis accompanied by bibliometric analysis gives us an overall view of various antecedents facilitating or restraining the family firms' internationalisation. To the best of our knowledge, to date, our study is the first to conduct the review of literature through bibliometric analysis with the help of *R Studio* software in the field of family firms' internationalisation. Such analysis assists in uncovering the most impactful perspectives in the existing family firms' internationalisation literature, the evolving trends in the research streams, most productive authors, top journals and articles and co-citation analysis; thus rendering our study different from recent review articles (Pukall and Calabro, 2014; Kontinen and Ojala, 2010b). Also, unlike recent articles catering specifically to family ownership or succession or professionalism dimensions (Kubíček and Machek, 2019; Hiebl and Mayrleitner, 2019; Sánchez-Marín *et al.*, 2020), our study instead helps researchers in adopting a holistic approach while understanding the internationalisation patterns of family firms. Moreover, to the best of our knowledge, our study is the first to review more than 100 research articles in the underlying area;
- Our study proposes a comprehensive framework integrating all the major themes relating to family firms' internationalisation. The framework unveils various aspects requiring researchers' attention in the coming years with respect to family firms' internationalisation, namely, development of socioemotional wealth measurement scale, degree of innovation, separation of "family ownership" and "family members involvement in management", optimal mix in the family firms' ownership structure, as well as their foreign market entry mode choice; and
- Finally, the study will assist academicians, researchers and family managers in accumulating a holistic understanding relating to family firms' internationalisation. Results from bibliometric and thematic analysis will help researchers in understanding the upcoming trends in family firms' research. The review of the existing literature will help them in identifying the unexplored areas in the underlying topic, thus guiding the future research scope. Also, it will assist the family firms' leaders and managers in understanding the important dynamics in overseas markets and various factors to be considered while planning their internationalisation.

2. Family firms' definition

Despite the growing literature on family firms' internationalisation, to date, there has not been any unanimity and agreement relating to family firms' definition. Hernandez-Linares *et al.* (2018) conducted a bibliometric analysis of 258 definitions on family firms published

over a period of 50 years from 1964–2013. The aim of the study was to uncover the trends in the definition of family firms and to identify the key terms used in underpinning the concept of family firms. The results of the study indicated that in many instances, family firms were not defined clearly and when defined, the emphasis was on the business side of the family business. The definitions have often not been able to operationalise the term "family" due to the nebulous nature of what makes up the family. During the first phase of analysis, i.e. 1964–1987, it was observed that "family ownership" was the key element in defining the family firms followed by "management" i.e. family members' involvement in management. During the second phase i.e. 1988–2002, although both "ownership" and "management" were again the two elements defining family firms, "management" seems to hold a key element during this period. Finally, in the last phase i.e. 2003–2013, it appears that both "ownership" and "management" are necessary elements in defining the family firms. In addition, it was observed that "governance" and "continuity" hold key significance during this phase, suggesting the rising importance of succession in defining the family businesses.

Donnelley (1964) pioneered in defining the family businesses, wherein he defined family business as one which "has been closely identified with at least two generations of a family and when this link has had a mutual influence on the company policy and on the interests and objectives of the family". However, over the years, this definition has been evolving, with researchers using varying contexts and criteria while defining family firms. Scholars have used multiple definitions due to the existence of heterogeneity amongst family firms. The definition of family firms also differs based upon the national context, for instance, the notion of the extended family is strong in the case of Asian and African firms while West-European and North-American firms exhibit more of a nuclear family character (Hernandez-Linares *et al.*, 2018). Hence, these varying contexts justify the diversity in the definition of family firms appearing in the literature. In general, the majority of researchers tend to integrate both the ownership, as well as management aspect wherein they define a family firm as "a firm where the family owns the majority of stock and exercises full management control" (Graves and Thomas, 2006; Gallo and Sveen, 1991; Shi *et al.*, 2019; D'Angelo *et al.*, 2016); while others define family firms only on the basis of ownership with varying degrees of control (Basly, 2007; Boellis *et al.*, 2016). Also, some researchers define family firms on the basis of firms' perceptions wherein the respondent firms were asked "if they consider themselves as family firms or not" (Okoroafo, 1999; Davis and Harveston, 2000); while some on the basis of the number of generations involved in family firms operations (Zahra, 2003). The application of such varying definitions often makes it difficult to integrate and compare the results of different studies. It is, thus, imperative for researchers to arrive at a unanimous and a common definition to render us with comparable analysis and understanding relating to family firms' internationalisation.

3. Methodology

The study attempts to use the three-step methodology for systematically reviewing research articles- Systematically selecting the articles following PRISMA guidelines (Moher *et al.*, 2009); Bibliometric Analysis to unveil the evolving trends in the underlying research stream and; Thematic Analysis for identifying the major themes and patterns in the underlying literature. PRISMA guidelines use the four-phase model (identification, screening, eligibility and included) to make the selection of the articles transparent and unbiased, thereby rendering us with more exhaustive and reliable results (Figure 1). Systematic Literature Review has an edge over the traditional reviewing techniques as it reviews the literature in a systematic and reproducible manner (Cook *et al.*, 1997). With the advancement in the research community, researchers are now focussing on conducting interdisciplinary

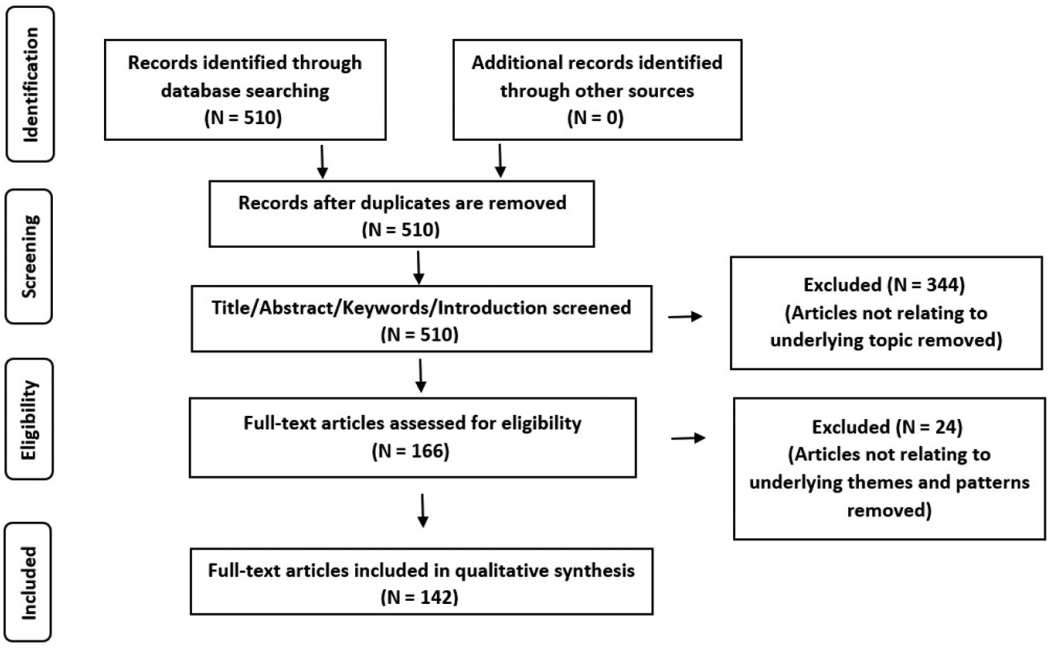


Figure 1. Methodology for data collection following PRISMA guidelines

research work (Parris and Peachey, 2013), thus, the systematic review would assist in reconciling the varying views of researchers in multiple disciplines.

3.1 Article selection process

3.1.1 Database selection. The study extracted the research articles from the online database SCOPUS®, having the exhaustive coverage of articles from the social sciences and humanities field. The database aided us in funnelling down and selecting the journals in the concerned area with the final sample of 142 research articles.

3.1.2 Inclusion criteria. Our study incorporated the following inclusion criteria:

- In line with the past research studies (Pukall and Calabro, 2014; Kontinen and Ojala, 2010b), we searched for the combination of the following words in the title, abstract and keywords – ("family firms*" OR "family businesses*" OR "family involvement" OR "family-owned" OR "family-controlled") AND ("international*" OR "global*" OR "foreign direct investment" OR "FDI" OR "mode of entry" OR "international sales" OR "exports" OR "international commitment");
- we considered only peer-reviewed research articles with the final publication stage;
- we kept the time period criteria open-ended, wherein we got the articles for a time span of 29 years (from 1991–2019);
- the relevant disciplines considered were business, management and accounting; social sciences; economics, econometrics and finance; decisions sciences; arts and humanities; psychology and energy; and
- articles published only in the English language were considered.

The incorporation of PRISMA guidelines initially yielded us 510 research articles following the above-mentioned inclusion criteria. After examining the title, keywords and abstract and in some cases introduction (wherever necessary), we obtained 166 research articles relating to family firms' internationalisation. We then went through the entire research articles to

further funnel down the relevant articles in the underlying area, which ultimately rendered us with the final sample of 142 research articles *(Figure 1)*.

3.2 Bibliometric analysis

The descriptive statistics for the above extracted 142 research articles were obtained through bibliometric analysis carried out with the help of *R Studio* Software (RStudio Team, 2020) (Aria and Cuccurullo, 2019). Extensive and voluminous data including the citation information about authors and documents, as well as bibliographical information relating to the authors' affiliation, the language of the document, etc., is processed by R Studio software with the help of the "bibliometrix" package (Aria and Cuccurullo, 2017) and, in turn, rendered us with various textual and visual analysis of such complex and fragmented data set. Thus, the objective is to uncover the most impactful perspectives surrounding the existing family firms' internationalisation literature, the evolving trends in the research streams, most productive authors, top journals and articles and networks extraction in the form of co-citation analysis.

3.3 Thematic analysis

Taking from (Braun and Clarke, 2006), the qualitative analysis of the above-identified 142 research articles was implemented through thematic analysis which assisted in identifying and analysing the patterns or themes in the underlying literature. After going through all the articles, a summary chart including the articles' core ideas, content and arguments was chalked out. Thematic analysis was then carried out by classifying and organising these articles wherein, homogenous factors, ideas and arguments were been clustered together to form a theme. The major themes that were identified in the underlying literature were classified into – family control, peculiar family characteristics, business environment, internationalisation theories and models and, internationalisation process.

4. Descriptive statistics through bibliometric analysis

A descriptive statistics and network relationship for 142 research articles were obtained through bibliometric analysis implemented through R studio. The software processed large volumes of bibliographical and citation data and gave us the following results:

4.1 General descriptive statistics

Table 1 represents the general descriptive statistics relating to the articles under study. In total there are 142 research articles, authored by 289 scholars during a time span of 29 years (from 1991–2019) which have been published in 58 different journals. Single-authored documents are written by 19 scholars while multi-authored documents are written by 270 scholars. Co-authors per Article Index is calculated as the average number of co-authors per

Table 1. General descriptive statistics

Documents (articles)	142
Sources (journal)	58
Time period	1991–2019
Authors	289
Authors of single-authored documents	19
Authors of multi-authored documents	270
Documents per author	0.491
Authors per document	2.04
Co-authors per document	2.77
Collaboration index	2.21

article. Here, co-authors per document (2.77) is greater than authors per document (2.04), thus indicating that scholars prefer conducting studies jointly with other scholars rather than doing them individually. The extent of collaborative practices between the authors is calculated by Collaboration Index which in our study is 2.21. The collaboration index (CI) is calculated as "total authors of Multi-authored articles/Total Multi-authored articles". That is to say, the index calculates co-authors per article using only the multi-authored article set (Aria and Cuccurullo, 2019). The index basically depicts the alignment in thoughts, ideas and arguments between various authors. In recent years, the greater emphasis on interdisciplinary research has encouraged scholars to collaborate their studies with other scholars.

4.2 Publication trend

Figure 2 depicts the number of research articles on family firms' internationalisation that have been published over the years starting from 1991 till 2019 (the year 2019 has been taken till the month of August). Though scholars have been studying the underlying research stream since 1991, the field gained traction since the year 2012–2013 with the sudden increase in the number of research articles. The annual percentage growth rate in the number of articles is witnessed to be 10.50%, indicating the growing importance and relevance of the underlying field. The above metrics indicate that family firms' internationalisation is a growing and evolving research area, thereby offering more avenues of future research.

4.3 Author influence

Table 2 renders us with the list of the top 10 authors who have made significant contributions in the field of family firms' internationalisation. It presents the top 10 most productive authors in terms of the number of articles that they have published, their H-Index and the total number of citations that they have received. *Calabro, A.* tops the list with 11 research articles and 396 total citations, who started his publications on the family business in the year 2009. Also, because of his maximum number of articles and citations, he also tops the H-Index list, which is based on the number of citations and the authors' most cited papers. The list is followed by *Kraus, S.*, with 6 research articles, 137 total citations and an H-Index of 5. *Kontinen, T.* takes the third place with 5 research articles, but since his total citations (201) are more than *Kraus, S.* (137), he is able to attain the same H-Index of 5, like that of Kraus.

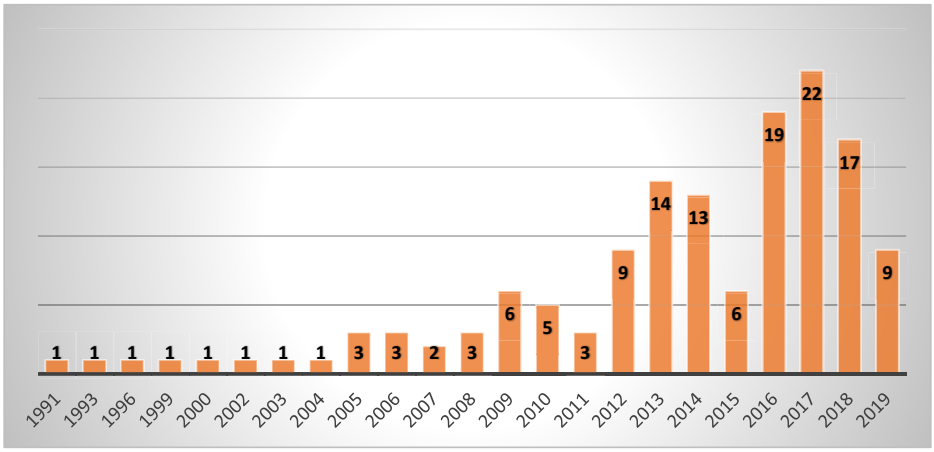


Figure 2.
Publication trend in the area of internationalisation of family firms

4.3.1 Authors' dominance ranking. Table 3 depicts the authors' dominance ranking, wherein, the authors' have been ranked on the basis of the number of times they appear as the first authors in the multi-authored articles. The above table indicates that both *Kontinen, T.* and *Arregle, J.L.* have secured the place of the first author in all their multi-authored articles. Similarly, out of 4 research articles, both *Claver, E.* and *Graves, C.* have appeared as first authors in 3 of the articles, etc. Thus, dominance ranking indicates the extent to which the research teams are been dominated by some specific authors.

4.4 Leading journals and articles

4.4.1 The top 10 contributing journals. FBR- Family business review, GSJ- Global strategy journal, JBR- Journal of business research, IJGSB- International journal of globalisation and small business*, IBR- International business review, JSBED- Journal of small business and enterprise development, RIBS- Review of international business and strategy, JFBM- Journal of family business management, JIBS- Journal of international business studies, EBER- Entrepreneurial business and economics review.*

Table 4 presents the list of top 10 contributing journals in terms of the number of articles published over the years in the field of family firms' internationalisation. *Family Business Review* attains the first place with 11 research articles, followed by *Global strategy Journal*

Table 2. The top 10 contributing authors (number of articles, H-index, total citations)

Authors	no. of articles	H-index	Total citations	Publication year start
CALABRO A	11	8	396	2009
KRAUS S	6	5	137	2014
KONTINEN T	5	5	201	2010
OJALA A	5	5	201	2010
CLAVER E	4	3	200	2007
GRAVES C	4	3	371	2004
MAJOCCHI A	4	3	48	2016
QUER D	4	3	200	2007
RIENDA L	4	3	200	2007
ARREGLE JL	3	2	138	2012

Table 3. Authors' dominance ranking

Authors	Dominance factor (DF)	Total articles	Single authored	Multi authored	First authored	Rank by DF
KONTINEN T	1.0000000	5	0	5	5	1
ARREGLE JL	1.0000000	3	0	3	3	1
CLAVER E	0.7500000	4	0	4	3	3
GRAVES C	0.7500000	4	0	4	3	3
CESINGER B	0.6666667	3	0	3	2	5
CALABR A	0.5454545	11	0	11	6	6
DE MASSIS A	0.3333333	3	0	3	1	7
MAJOCCHI A	0.2500000	4	0	4	1	8
RIENDA L	0.2500000	4	0	4	1	8
KRAUS S	0.1666667	6	0	6	1	10

Year	FBR	GSJ	JBR	IJGSB	IBR	JSBED	RIBS	JFBM	JIBS	EBER
1991	1									
1992										
1993	1									
1996	1									
1999	1									
2000	2									
2004				1						
2005	1			2						
2006	1					1			1	
2008	1					1				
2009	1		1	1						
2010						1			1	
2012						1		1		
2013					1	1				
2014	1		1	1	1			1		
2015			1							1
2016			3	2	2			2	1	
2017					1	1	6			2
2018		8	1					1	1	1
2019			1		1				1	
Total	*11*	*8*	*8*	*7*	*6*	*6*	*6*	*5*	*5*	*4*

Notes: *FBR - Family Business Review, GSJ - Global Strategy Journal, JBR - Journal of Business Research, IJGSB - International Journal of Globalisation and Small Business, IBR - International Business Review, JSBED - Journal of Small Business and Enterprise Development, RIBS - Review of International Business and Strategy, JFBM - Journal of Family Business Management, JIBS - Journal of International Business Studies, EBER - Entrepreneurial Business and Economics Review*

Table 4. The top 10 contributing journals

with 8 research articles, *Journal of Business Research* with 8 articles, *International Journal of Globalisation and Small Business* with 7 articles, etc. *Family Business Review* is witnessed to be the most consistent and oldest journal dealing with family business research. We also came across some upcoming journals that have shown their recent interests in the issues relating to family firms' internationalisation such as *Global Strategy Journal*, *International Business Review*, *Journal of International Business Studies, Review of International Business and Strategy* and *Entrepreneurial Business and Economic Review*, indicating the growing and evolving literature in this field.

4.4.2 The top 10 frequently-cited articles. Table 5 gives the list of most frequently-cited articles/manuscripts in the underlying area. With 464 total citations and 51.56 total citations per year, article "Diversification decisions in family-controlled firms" authored by Gomez-Mejia, L.R.; Makri, M.; Kintana, M.L. in 2010 tops the list, followed by "International expansion of US manufacturing family businesses: the effect of ownership and involvement" authored by Zahra, S.A. in 2003, etc. Out of the top 10 cited articles, 5 are published in *Family Business Review* indicating the consistency and quality of the journal.

4.5 Publication distribution by region

Table 6 presents the list of the top 10 contributing countries in terms of the number of articles been published by the corresponding authors from these countries. The USA tops the list with 16 research articles, followed by Italy with 15 articles, Spain with 12 articles, etc. The majority of these top countries are developed nations, indicating the need and

Table 5.
The top 10 frequently-cited articles

Author(s)	Article title	Journal	Total citations	Total citations per year
Gomez-Mejia *et al.* (2010)	Diversification decisions in family-controlled firms	*Journal of Management Studies*	464	51.56
Zahra (2003)	International expansion of US manufacturing family businesses: the effect of ownership and involvement	*Journal of Business Venturing*	417	26.06
Fernandez and Nieto (2006)	Impact of ownership on the international involvement of SMEs	*Journal of International Business Studies*	267	20.54
Fernández and Nieto (2005)	Internationalisation strategy of small and medium-sized family businesses: some influential factors	*Family Business Review*	261	18.14
Gallo and Sveen (1991)	Internationalizing the family business: facilitating and restraining factors	*Family Business Review*	172	6.14
Graves and Thomas (2008)	Determinants of internationalisation pathways of family firms: an examination of family influence	*Family Business Review*	157	14.27
Bhaumik *et al.* (2010)	Does the ownership structure of emerging-market firms affect their outward FDI? The case of the Indian automotive and pharmaceutical sectors	*Journal of International Business Studies*	149	16.56
Graves and Thomas (2006)	Internationalisation of Australian family businesses: a managerial capabilities perspective	*Family Business Review*	146	11.23
Sciascia *et al.* (2012)	The role of family ownership in international entrepreneurship: exploring non-linear effects	*Small Business Economics*	134	19.14
Gallo and Pont (1996)	Important factors in family business internationalisation	*Family Business Review*	130	5.65

opportunity for scholars from developing nations to contribute to the underlying research stream. Also, Europe is identified as making significant contributions relating to family firms' internationalisation, as 7 out of the top 10 countries, are from Europe.

Table 7 presents the total and average citations of the top 10 contributing countries, with the USA again taking first place with 1,246 total citations. However, its average citations (77.9) are lower than that of Spain (83.9), indicating the greater quality of research work stemming from Spain in comparison to the USA. Although scholars from Italy have published more articles (15) than Spain (12), the articles from Spain received higher total and average citations because of their greater significance on quality research work. Thus, it is imperative to not just consider the number of publications but also their quality in terms of total and average citations received, while ranking the countries in terms of their productivity.

4.5.1 Country collaboration. Figure 3 depicts the collaborative practices and networks between the countries while undertaking research in the area of family firms' internationalisation. The scholars are undertaking international collaborations to exhibit an exhaustive and comprehensive picture of the issues revolving around the world, facilitate the exchanges of ideas and facts in the area of study. The above figure depicts that the USA has maximum collaborations with other countries (represented by the biggest circle), followed by Italy, UK, Spain, Germany, etc. These collaborations indicate the strengthening of networks and relationships amongst the countries within the research community.

Country	No. of articles
USA	16
Italy	15
Spain	12
Germany	7
UK	7
Finland	6
France	6
Taiwan	5
Australia	4
Poland	4

Table 6. Corresponding authors' countries

Country	Total citations	Average article citations
USA	1246	77.9
Spain	1007	83.9
Italy	401	26.7
Australia	303	75.8
UK	276	39.4
Germany	220	31.4
France	218	36.3
Finland	204	34.0
Singapore	96	96.0
Austria	68	34.0

Table 7. Total citations per country

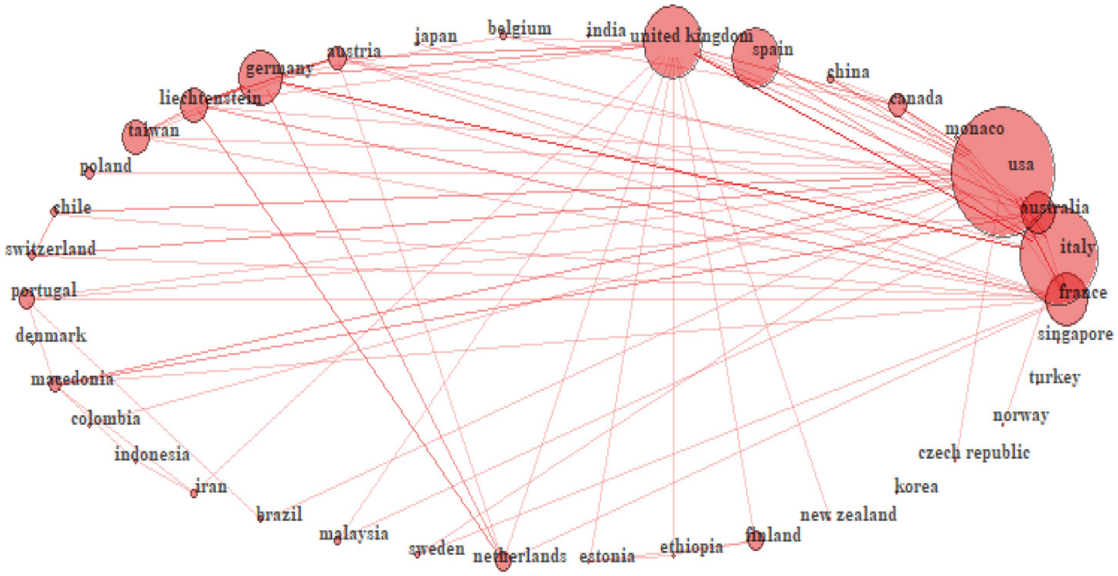


Figure 3.
Country collaboration network

4.6 Co-citation analysis

Figure 4 presents the co-citation network. Taking from (Aria and Cuccurullo), "co-citation of two articles refers to when both the articles are cited in a third article". The network aids in identifying the number of co-citations between the articles. As Kontinen and Ojala (2010a), Zahra (2003) and Graves and Thomas (2008), are represented with bigger circles, it indicates that they have maximum connections with other articles and have been highly cited with them in a third document. The dark lines represent the strong connections between the articles and as these lines become lighter and dotted, it indicates the weak and indirect

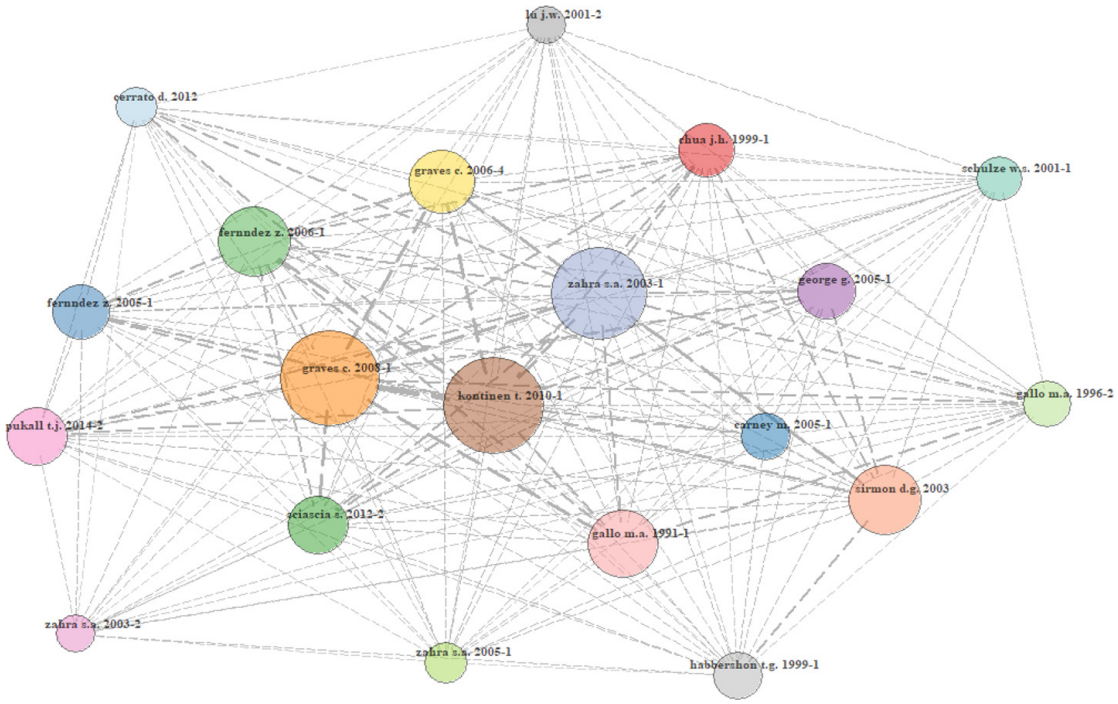


Figure 4.
Co-citation network

relationship between the articles. These connections and networks between the articles represent the similarity between them, either in terms of themes, techniques, etc.

5. Theoretical background

Internationalisation is characterised as being the process of consistent involvement and participation in the overseas markets. With the advent of globalisation, family firms are realizing the significance of making their entrance abroad, and thus expanding internationally. Over the years, scholars are making greater attempts to explore if internationalisation poses opportunities or challenges to the family firms.

A review of literature in the following sub-sections has been attempted by classifying and organising the entire literature of family firms' internationalisation (142 articles) into five major themes. The themes have been identified by conducting thematic analysis wherein homogeneous factors, ideas and arguments were been clustered together to form a theme. For e.g. all the factors relating to firms' ownership, as well as involvement in management and board were clustered into a theme, namely, *Family Control*; all the family-specific traits that distinguish a family firm from the non-family firms such as SEW, bifurcation bias, family generation, family ties, family risk perception and their long term orientation were clustered into a theme, namely, *Peculiar Family Characteristics*; all the factors relating to firms' internal and external environment including internal resources and home and host country factors were clustered into a theme, namely, *Business Environment*; various theories and models explaining the family firms' internationalisation were clustered into *Internationalisation Theories and Models*; and finally, the family firms' internationalisation strategies and their foreign market entry modes could be clustered into a theme, namely, *Internationalisation Process*.

5.1 Family control

Despite its significance, results are still inconclusive and unclear with respect to the relationship between family control and family firms' degree of internationalisation. On one hand, it is often observed that family firms have the fear of losing their family control and power while expanding their operations abroad (Tsao *et al.*, 2018). As a result, firms with highly concentrated family ownership and having high involvement in firms' management or board are viewed as risk-averse, thereby avoid taking uncertain decisions like internationalisation (Jimenez *et al.*, 2019; Ray *et al.*, 2018; Calabro *et al.*, 2013). It is often said that the involvement of non-family managers becomes salient to overcome the shortfall in family members' expertise, as they help in creating distinct and inimitable knowledge-based resources, experiences and network ties (Alayo *et al.*, 2019; Calabro *et al.*, 2017) and provides greater risk-sharing to family firms (Claver *et al.*, 2009), thereby fostering the pace of family firms' internationalisation. For example, Burman family-owned firm, Dabur, decided to professionalise its operations in the 1990s. While Dabur maintained the family ownership in the organisation, they decided to professionalise their managerial operations by recruiting members from outside, as the company reached its fifth generation. By 2011, such professionalism aided Dabur in establishing 17 overseas plants (Srinivasan *et al.*, 2019). Another school of thought posits that the rise in family members involved in management, ownership and governance altogether, leads to the efficient flow of information, knowledge and ideas between the owners and the managers, thereby, stimulating their internationalisation process (Tsao *et al.*, 2018; Zahra, 2003).

It is often contended that such obstacles and resistance by family firms to expand overseas is mitigated with the presence of foreign investors in the firms' ownership structure. Such local foreign partners bring in additional risk-bearing capacity and aids in

accumulating the high-quality network relationships and resources, ultimately helping them to efficiently penetrate into foreign markets (Ray *et al.*, 2018; Bhaumik *et al.*, 2010). Likewise, the presence of corporate block-holders in the family firms' ownership structure, stimulates their degree of internationalization, as these block-holders provide access to quality resources, network relationships, managerial capabilities, ultimately reducing the family firms' risk aversion (Fernandez and Nieto, 2006; Fernández and Nieto, 2005). The above review implies that a family firm must undertake certain modifications in their attitudes and orientation to broaden up their international horizon.

5.2 Peculiar family characteristics

With the advent of globalisation, all the firms are realizing the significance of making their entry abroad. However, as observed already, family firms carry out and execute their internationalisation process differently from non-family firms due to their peculiar family characteristics such as Socioemotional Wealth (SEW), family generation, family ties and bifurcation bias. Due to the dread of losing SEW i.e. "the family members' social needs or various other non-economic benefits accruing beyond the financial rewards from running a business", family firms exhibit a lower degree of internationalisation, as it leads to sharing decision-making authority with the foreign partners and suppliers (Pongelli *et al.*, 2019; Gomez-Mejia *et al.*, 2010). At any moment, they are willing to assume higher risks only if undertaking such risks provides them the opportunity to preserve firms' non-economic assets (Gomez-Mejia *et al.*, 2010; Tsao *et al.*, 2018; Dou *et al.*, 2019; Evert *et al.*, 2018; Monreal-Pérez and Sánchez-Marín, 2017; Zahra, 2003). They have a general preconceived notion that the rewards from internationalisation are not worthy enough against the risk associated with the changes in the operational structure and the loss of SEW (Alessandri *et al.*, 2018).

The existence of multiple generations fosters the family firms' internationalisation process, mainly because, multiple generation family members possess diverse knowledge and information, which is paramount for their internationalisation (Dou *et al.*, 2019; Gallo and Sveen, 1991), as opposed to Alayo *et al.* (2019), who claims that the involvement of multiple generations results in divergent managerial perspectives, ideas, vision and mission, thus disrupting cooperation and coordination. Family firms controlled by the founding generation display greater commitment in preserving their SEW and family name to successfully hand it over to the future generations, thereby exhibit lower inclination towards uncertain decisions like internationalisation (Fang *et al.*, 2018; Menéndez-Requejo, 2005). In contrast, the younger generations in the family firms are better able to locate efficient international opportunities due to up-to-date skills, knowledge, technology and capabilities (Calabro *et al.*, 2016; Colli *et al.*, 2013; Fernández and Nieto, 2005), willingness to assume greater risks (Dou *et al.*, 2019; Menéndez-Requejo, 2005) and exhibit lower attachment in preserving family assets (Fang *et al.*, 2018), thereby stimulating their internationalisation. However, some scholars rather believe, that as the founder generation grows old (Davis and Harveston, 2000) and as the level of knowledge-based resources in the firm rises (Fang *et al.*, 2018), the founder generation accumulates necessary experience, skills and expertise, thereby propelling them to successfully internationalise their operations abroad (Okoroafo, 1999; Davis and Harveston, 2000; Fang *et al.*, 2018). It is imperative to understand that the degree of family firms' internationalisation is affected by the firms' intergenerational dynamics and not by the succession *per se*. Thus, the greater the levels of competence-based trust between the incumbent and successor, the higher is the possibility to arrive at a common goal towards the internationalisation of the firm (Shi *et al.*, 2019; Calabro *et al.*, 2016).

It is also observed that family firms suffer from high degrees of bifurcation bias i.e. considering family assets superior to non-family ones. They exhibit greater sensitivity and attachment towards the firm and the family members, deterring them to hire external professional managers, which is found critical to their internationalisation (Kano and Verbeke, 2018; Majocchi *et al.*, 2018). For example, family-run Jet Airways was forced to suspend all its operations in April 2019, as it experienced a lack of external professionalism due to the majority of family members sitting on the board of the airline (Bhattacharyya, 2019). Firms with the authoritarian family structure exhibit a greater degree of bifurcation bias, ensuring that the topmost positions in the family firms to be occupied with the family members only, even if they are not capable of handling the responsibilities and duties of such positions (Arregle *et al.*, 2019). It is observed that the family firms would be able to achieve efficient international expansion only if they are willing to give away with their bifurcation bias (Kano and Verbeke, 2018). The review of the above literature ultimately reckons us with the significance of peculiar family characteristics in influencing their internationalisation decisions. The understanding of these characteristics become paramount for the family business leaders, as they play a key role in distinguishing family firms' internationalisation from their non-family counterparts.

5.3 Business environment

5.3.1 Internal business environment. In every firm, there are certain elements residing within an organisation that influences its decision-making process, either positively or negatively. In family firms too, very often they exhibit greater willingness to internationalise, however, their operations are restricted by the specific environment in which they thrive. Many of these restrictions include namely, higher risk aversion behaviour (Monreal-Pérez and Sánchez-Marín, 2017; Graves and Thomas, 2006), combined with their limited financial resources (Ramón-Llorens *et al.*, 2017; Claver *et al.*, 2009; Graves and Thomas, 2008; Graves and Thomas, 2006; Gallo and Pont, 1996); lack of experienced human resources (Cesinger *et al.*, 2014; Kontinen and Ojala, 2010a; Bhaumik *et al.*, 2010; Graves and Thomas, 2006; Gallo and Pont, 1996); fewer network ties (Graves and Thomas, 2004) and obsolete technology (Gallo and Pont, 1996), ultimately restricting their *ability* to undertake risky and resource-demanding activities like internationalisation. It is, therefore, imperative to distinctly assess the impact of both the family firms' *willingness* and *ability* on their degrees of internationalisation (Dou *et al.*, 2019; Evert *et al.*, 2018). It is often observed that family firms are reluctant in hiring non-family external professional managers to finance their international expansion as it will dilute their family control and might invade their privacy (Graves and Thomas, 2006; Ramón-Llorens *et al.*, 2017; Monreal-Pérez and Sánchez-Marín, 2017; Bhaumik *et al.*, 2010; Claver *et al.*, 2009; Graves and Thomas, 2008). However, for family firms, it is important to realise that presence of external non-family managers and external capital strengthens family firms with supplementary intangible resources, financial resources and a high degree of professionalism, which is imperative for their internationalisation (Majocchi *et al.*, 2018; D'Angelo *et al.*, 2016; Claver *et al.*, 2009; Calabro *et al.*, 2009). Researchers have also come up with a contrasting point of view wherein they witness that family firms internationalise more because of the presence of experienced family members on the firms' board (Dou *et al.*, 2019; Casillas and Acedo, 2005); heterogeneity in the top management team with the diversified educational background (Tsao *et al.*, 2018); higher degrees of innovation (Corsi and Prencipe, 2018; Braga *et al.*, 2017; Davis and Harveston, 2000); rapid decision-making skills (Marinova and Marinov, 2017); strong network ties with international partners (Graves and Thomas, 2006; Marinova and Marinov, 2017; Cesinger *et al.*, 2016; Zain and Kassim, 2012); high degrees of trust amongst

family members (Colli *et al.*, 2013) and their ability to efficiently capitalise on limited resources (Graves and Thomas, 2006).

5.3.2 External business environment. For businesses to be successful, it is imperative to have a clear and constant understanding of the fluctuations happening in their external environment. Due to the uncertainties associated with foreign markets, family firms generally prefer to expand their operations in less distant countries which could be handled with limited managerial expertise and capabilities (Graves and Thomas, 2006; Cesinger *et al.*, 2014; Gomez-Mejia *et al.*, 2010; Graves and Thomas, 2008; Rexhepi *et al.*, 2017). In fact, the negative effects of such psychic distance are found to be more pronounced in family firms, as they lack experienced human resources having diverse international market knowledge (Cesinger *et al.*, 2014; Kontinen and Ojala, 2010a). Even if the family firms have less information and knowledge about developed markets, still they prefer internationalising their operations in economies characterised with highly secured market mechanisms rather than countries with poorly regulated market systems (Hernandez *et al.*, 2018). In contrast, Bhaumik and Driffield (2011), in their study witnessed that Indian family firms are inclined in making their OFDI in developing economies rather than in developed economies because it is easy to replicate their business models in such economies due to the similarity in terms of market dynamics and institutional systems (Arregle *et al.*, 2019; Bhaumik and Driffield, 2011). Also, family firms would exhibit a greater tendency to internationalise their operations, if the foreign business partner is also a family-owned firm, as it will help in mitigating various cultural and operational barriers (Swinth and Vinton, 1993; Gallo and Pont, 1996).

It is observed that once family firms have initiated their expansion in international markets, they do not lag behind the non-family firms in terms of the progression in their extent of internationalisation (Graves and Thomas, 2004). However, successful internationalisation calls for the integration of the foreign market experience of both the non-family and family members (Majocchi *et al.*, 2018). Thus, a comprehensive understanding of the factors prevailing in the external business environment will assist family firms in identifying profitable opportunities abroad, ultimately shaping their internationalisation strategies.

5.4 Internationalisation theories and models

Over the years, literature has witnessed the evolution in the international business theories, explaining the transition in the nature and pattern of the firms' internationalisation over time. Uppsala model has been considered as the most recognised internationalisation theory. With respect to this, it is observed that family firms tend to follow the Uppsala Model of internationalisation, as they are characterised by limited knowledge and experience about foreign markets. They initiate their internationalisation with geographically close countries requiring lower degrees of resources, knowledge and experience and gradually they tend to expand their operations in more distant countries after accumulating sufficient knowledge, skills and experience (Cesinger *et al.*, 2014; Claver *et al.*, 2007; Kontinen and Ojala, 2010a). This model was then revised to the network theory which states the importance of building strong network ties with the external partners for successful internationalisation. However, these existing internationalisation theories often fail to fully describe the internationalisation process of family firms, due to their peculiarities related to SEW (Cesinger *et al.*, 2016).

Drawing on the resource-based view (RBV) perspective, it is observed that family firms who are able to generate inimitable, rare, valuable and non-substitutable resources are able to enhance their export intensity (Alonso and O'Brien, 2017; Casillas and Acedo, 2005). The resource dependency theory argues that the acquisition of external resources like non-family

members augments the family firm's resource pool by bringing in the external international market knowledge which is vital for accomplishing the firms' strategically important decisions like internationalisation (Calabro *et al.*, 2009; Calabro *et al.*, 2017).

Taking into consideration the stewardship theory perspective, there is a high degree of competence-based trust in family firms, such that family members act as good stewards of the company. Instead of personal goals, all the family members always prioritise firms' goals, mission and objectives, thereby, helping them to arrive at a common goal towards the internationalisation of the firm (Calabro *et al.*, 2016; Zahra, 2003). When a firm is owned, as well as managed by family members, it follows the stewardship theory perspective, as both the principal, as well as agent are the family members, and hence there is no conflict of interest (Avrichir *et al.*, 2016). However, when a family firm is owned by family members but is managed by non-family members, agency theory comes into play, as non-family members tend to take risky decisions for their short term gains which are not in the best interest of the family owners (Avrichir *et al.*, 2016; Mitter *et al.*, 2014). In contrast, another set of scholars believe behavioural agency theory to be better explaining the conduct of family firms rather than agency theory. The theory advocates that it is not the general rule for the firms' managers and owners to be certainly risk-averse, rather there is a specific premise or ground for the firms to behave in a particular manner. Thus, this model explains that the premise for the family firms to be risk-averse is due to their desire to preserve their SEW (Evert *et al.*, 2018).

By virtue of Kinship Theory (Okoroafo, 1999), it is contended that the likelihood of family firms' internationalisation intensifies with the presence of family-owned firms abroad, enabling them to better share their core values and outrightly alleviating all the barriers approaching their internationalisation process (Swinth and Vinton, 1993; Gallo and Pont, 1996). Also, Social Capital Theory explains that family firms possess greater degrees of social capital, wherein, high degrees of trust and strong bond amongst family members leads to enhanced communication in the family firms. As a result, family members are better able to share the accumulated knowledge and experience throughout the firm, which, in turn, stimulates their degree of internationalisation (Jimenez *et al.*, 2019; D'Angelo *et al.*, 2016). Drawing from upper echelons theory, it is argued that the specific characteristics of firms' managers such as their knowledge and skills, shape their decision-making ability. Eventually, it is observed that when CEOs in family firms possess a good educational base, they are able to process diverse market information efficiently, ultimately leading to greater levels of internationalisation (Ramón-Llorens *et al.*, 2017). The comprehensive knowledge of these theories and models assist us in linking and corroborating the internationalisation process and patterns of family firms with the underlying theoretical concepts and ideas. Furthermore, these theories and models offer conceptual explanation and rationale behind varying internationalisation strategies adopted by family firms during their internationalisation process.

5.5 Internationalisation process

Family ownership impacts the choice of foreign market entry decisions of family firms. Drawing on the SEW perspective, family firms are more concerned about preserving their family wealth and control, and thus are reluctant in sharing their decision-making authority with foreign partners. As a result, while expanding abroad, family firms have the inclination towards captive offshoring strategy i.e. operating in foreign markets through FDI (Pongelli *et al.*, 2019). Another view says that family firms prefer making early entry through indirect entry modes like exports, as it is considered as low-risk entry mode, in comparison to FDI (Kontinen and Ojala, 2010a; Monreal-Pérez and Sánchez-Marín, 2017). In cases, where firms

lack international market experience, family firms exhibit greater preference, in comparison to non-family firms, to enter the foreign market through joint ventures rather than wholly-owned subsidiaries. This is because, foreign partners are already familiar and possess adequate knowledge and experience about the foreign market, ultimately helping the family firms to penetrate into overseas markets (Kuo *et al.*, 2012). However, when the degree of internal and external uncertainties reduces (Kao and Kuo, 2017) and when family firms acquire sufficient international market experience (Kuo *et al.*, 2012), they exhibit a greater preference for wholly-owned subsidiaries so as to avoid any external party interference (Kao and Kuo, 2017; Kuo *et al.*, 2012). Thus, in conclusion, it is observed that family firms have such differing internationalisation strategies due to the heterogeneity residing within them with respect to various dimensions, namely, the extent of family involvement, family resources, risk orientation, etc. Also, family firms shall focus on having a formally documented strategic plan to determine the appropriate foreign market entry mode, ultimately leading to successful internationalisation.

6. Discussion and future research directions

Over the years, there has been a growing and ever-evolving literature on various dimensions relating to family firms' internationalisation. Researchers are showing greater interest in exploring the ways in which family firms' are internationalising their operations abroad. However, results are still inconclusive with respect to family firms internationalisation, wherein, some scholars say that they internationalise less (Gomez-Mejia *et al.*, 2010; Fernandez and Nieto, 2006; Jimenez *et al.*, 2019), while others say they internationalise more (De Massis *et al.*, 2018; Zahra, 2003) and others witness no significant difference between family and non-family firms' degree of internationalisation (Cerrato and Piva, 2012; Eddleston *et al.*, 2018). Such inconclusiveness offers scholars an opportunity to further test the relationship and contribute to the research stream. Also, in line with Chua *et al.* (2012), we observe that the focal point of the bulk of research studies has been on comparing family firms' performance and behaviour with that of non-family firms, with an implicit assumption of homogeneity amongst various family firms. This fallacious assumption renders limited generalisability, thus exhibiting the need and opportunity for the researchers to undertake an in-depth exploration of the level and type of heterogeneity residing within family firms. With regard to the definition of family firms, there hasn't been any consistency as there is no single definition followed by the researchers (Hernandez-Linares *et al.*, 2018). For comparable analysis, it is imperative for the researchers to arrive at a unanimous definition, thus, in line with Graves and Thomas (2006), Gallo and Sveen (1991), Shi *et al.* (2019);, D'Angelo *et al.* (2016), we also propose that both the ownership, as well as involvement in management shall be taken into consideration while defining family firms to arrive at more reliable and realistic conclusions. Results from bibliometric analysis render us with the viewpoint that family firms' internationalisation is an upcoming research area offering greater opportunities for future research scope, as there has been a sudden increase in the number of research articles in the recent past (Figure 2). Also, Table 4 indicates that although *Family Business Review* is the most consistent and oldest journal dealing with family business research, there are some upcoming journals that have shown recent interests in the issues relating to family firms' internationalisation such as *Global Strategy Journal*, *International Business Review*, *Journal of International Business Studies*, *Review of International Business and Strategy* and *Entrepreneurial Business and Economic Review*; which further provides an opportunity to the scholars to tap upon these upcoming journals in the underlying area. Although, according to a report by Credit Suisse Research Institute (2018), globally, China and India stand at the first and third position, respectively,

pertaining to the number and concentration of family-owned businesses, however, Table 6 indicates that the majority of top 10 productive countries in the underlying area are developed nations, indicating the need and opportunity for the scholars from developing nations like India and China to make significant contributions in the underlying research stream.

We also conducted a thematic analysis to identify the major themes and patterns evolving in the study of family firms' internationalisation. The analysis aided us in classifying and organising the articles, wherein, homogenous factors, ideas and arguments in the research articles were clustered together to form a theme. This rendered us with the creation of 5 major themes, i.e. family control, peculiar family characteristics, business environment, internationalisation theories and models and internationalisation process. With the help of these themes, the study proposes a holistic framework offering the future direction to the scholars, to explore the reasons pertaining to various facets and dimensions relating to family firms' internationalisation. The linkages between various themes and patterns in the underlying area are depicted in Figure 5.

6.1 Family control

In general, it is believed that family firms are not very much agile in integrating and expanding their domestic business abroad due to their risk-averse behaviour and their desire to preserve their family control (Jimenez *et al.*, 2019; Ray *et al.*, 2018; Calabro *et al.*, 2013). In line with Liang *et al.* (2014), we observe that while examining the extent of family firms' internationalisation, the majority of the past research studies have neglected to

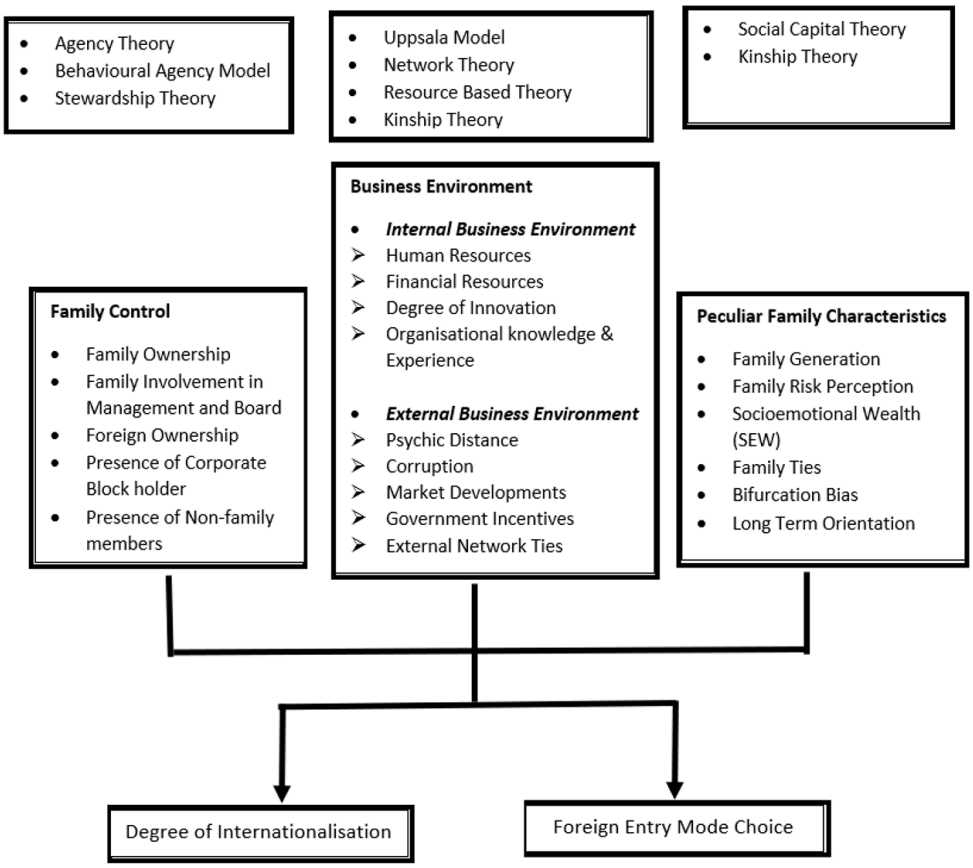


Figure 5.
Theoretical framework for internationalisation of family firms

differentiate between the characteristics of the degree of "family ownership" and the degree of "family involvement in firms' management". Different facets of family control i.e. family ownership and family members' involvement in management have differing impacts on their degree of internationalisation. Thus, as a first suggestion, we recommend the future scholars to explore and analyse the distinct impact of "family ownership" and "family members' involvement in management" on their degree of internationalisation and to come up with the reasons for such differences.

Also, Gallo and Sveen (1991), in their study presented that being a family firm in itself does not imply that it cannot expand its operations overseas, however, a family firm must undertake certain modifications in their attitudes and orientation to broaden up their international horizon. With regard to this, in recent years we have witnessed the presence of foreign investors (Calabro *et al.*, 2013) and corporate block holders (Fernandez and Nieto, 2006), in the family firms' ownership structure, as they bring in additional risk-bearing capacity and access to quality resources. Thus, in this regard, our second recommendation to the scholars is to decide upon the optimal mix in the family firms' ownership structure such that it aids them in fostering their internationalisation on one hand and at the same time maintaining their family legacy on the other hand.

6.2 Peculiar family characteristics

There are some peculiar family characteristics, which hinders their degree of internationalisation, like their desire to preserve their SEW (Jimenez *et al.*, 2019); high degree of risk aversion (Evert *et al.*, 2018); limited resource pool (Dou *et al.*, 2019) and their high degree of bifurcation bias (Kano and Verbeke, 2018); while some fosters their degree of internationalisation, like their long term orientation (Zahra, 2003); strong family ties (Marinova and Marinov, 2017) and trust-based intergenerational dynamics (Shi *et al.*, 2019). With regard to family firms' intergenerational dynamics, it is still inconclusive whether the involvement of multiple generations fosters their degree of internationalisation (Dou *et al.*, 2019; Gallo and Sveen, 1991); or multiple generations hinder their internationalisation (Alayo *et al.*, 2019) or whether younger generations stimulate their degree of internationalisation (Alayo *et al.*, 2019) or is it the founding generation which leads to successful internationalisation (Okoroafo, 1999; Fang *et al.*, 2018). Such inconclusiveness renders an opportunity for scholars to further test this relationship and contribute towards a better understanding of family firms' intergenerational dynamics. Also, with regard to SEW, it is observed that due to the family firms' desire to preserve SEW, family name and reputation, they exhibit lower inclination towards risky decisions like internationalisation (Jimenez *et al.*, 2019; Zahra, 2003). However, to date, except for few studies (Berrone *et al.*, 2012; Debicki *et al.*, 2016), the limited contribution has been made towards the development and validation of formal scale or technique for the measurement of family firms' SEW. As SEW is the prime factor that distinguishes family firms from non-family firms, thus in the future, it is imperative for scholars to contribute towards the development of SEW measurement scale or technique in the context of family businesses.

6.3 Business environment

Various factors in family firms' internal and external business environment either facilitate or restrict their international expansion. Amongst these, limited financial resources (Ramón-Llorens *et al.*, 2017); lack of experienced human resources (Cesinger *et al.*, 2014; Kontinen and Ojala, 2010a); and psychic distance (Cesinger *et al.*, 2014), lowers their degree of internationalisation; while, strong network ties with international partners (Graves and Thomas, 2006); high degrees of trust amongst family members (Colli *et al.*, 2013) and rapid

decision-making skills (Marinova and Marinov, 2017), stimulate their degree of internationalisation. With the rapid advancement in the technological environment, it is nearly impossible for the family firms to sustain their competitive advantage without enhancing their degree of innovativeness. However, except for a few studies, Corsi and Prencipe (2018), Braga *et al.* (2017), there is scant literature exploring the impact of innovation on family firms' degree of internationalisation. This ultimately provides scholars with the future research scope to further explore this relationship in detail. Also, while some studies place greater emphasis on geographical proximity, wherein, they claim that family firms prefer expanding in less distant countries (Graves and Thomas, 2006; Cesinger *et al.*, 2014); while others emphasise more on network relationships, wherein, they observe that family firms prefer entering those foreign markets where they have strong network ties with foreign business partners (Swinth and Vinton, 1993; Gallo and Pont, 1996). Thus, it would also be interesting for scholars to explore in-depth the factors that impact the family business leaders' mindset when they take their internationalisation decisions.

6.4 Internationalisation theories and models

It is important to understand the role of various internationalisation models and theories in explaining the family firms' internationalisation process and patterns. From literature, we identified various theories that have been used by the researchers in the past while explaining the family firms' internationalisation such as, Uppsala model (Cesinger *et al.*, 2014; Claver *et al.*, 2007); network theory (Cesinger *et al.*, 2016); resources based view theory (Alonso and O'Brien, 2017; Casillas and Acedo, 2005); Stewardship theory (Calabro *et al.*, 2016; Zahra, 2003); agency theory (Avrichir *et al.*, 2016); behavioural agency theory (Evert *et al.*, 2018); Kinship theory (Okoroafo, 1999); social capital theory (Jimenez *et al.*, 2019); upper echelons theory (Ramón-Llorens *et al.*, 2017). However, in line with Cesinger *et al.* (2016), we also observe that these existing internationalisation theories might fail to fully describe the internationalisation process and patterns of family firms, due to their peculiarities related to SEW. Thus, it becomes important for future researchers to explore whether these existing internationalisation theories are able to best explain the nature of family firms' internationalisation or they need to be modified to take into account the family firms' outlook.

6.5 Degree of internationalisation

It has been observed that it is nearly inconclusive as to whether family firms' internationalise more or less, wherein, some scholars say that they internationalise less (Gomez-Mejia *et al.*, 2010; Fernandez and Nieto, 2006; Jimenez *et al.*, 2019), while others say they internationalise more (De Massis *et al.*, 2018; Zahra, 2003) and others witness no significant difference between family and non-family firms' degree of internationalisation (Cerrato and Piva, 2012; Eddleston *et al.*, 2018). Thus, it is important for researchers to analyse the impact of the factors explained above to understand their role in family firms' degree of internationalisation.

6.6 Foreign entry mode choice

Recently, researchers are showing greater interest in understanding the family firms' choice of mode of entering the foreign markets (Pongelli *et al.*, 2019; Monreal-Pérez and Sánchez-Marín, 2017). However, despite being the growing field of literature, results are still inconclusive as to whether family firms prefer entering foreign markets through FDI (Pongelli *et al.*, 2019); or through indirect entry modes like exports (Kontinen and Ojala, 2010a; Monreal-Pérez and Sánchez-Marín, 2017); or whether they prefer to enter through

joint ventures (Kuo *et al.*, 2012); or whether they exhibit a greater preference for wholly-owned subsidiaries (Kao and Kuo, 2017; Kuo *et al.*, 2012). Thus, such inconclusiveness offers an opportunity for the family business scholars to further explore the family firms' choice of foreign market entry decisions.

Various factors discussed above can be combined to yield a theoretical framework that would assist academicians, researchers and family managers in obtaining a holistic view about various antecedents facilitating or restraining the family firms' internationalisation. It would, thus, pave the way for the researchers in identifying the unexplored areas in the underlying topic and thereby making significant contributions to the research community. Also, it will assist the family firms' leaders and managers in understanding the important dynamics in overseas markets and various factors to be considered while planning their internationalisation.

7. Conclusion

The study undertakes systematic review of literature of 142 articles, authored by 289 scholars which are published in 58 peer-reviewed journals in the past 29 years (from 1991–2019), relating to various aspects pertaining to family firms' internationalisation. For this purpose, the study uses a three-step methodology- Systematically selecting the articles following PRISMA guidelines; Bibliometric Analysis to unveil the evolving trends in the underlying research stream and; Thematic Analysis for identifying the major themes and patterns in the underlying literature. To the best of our knowledge, to date, our study is the first to conduct the review of literature through bibliometric analysis with the help of *R Studio* software in the field of family firms' internationalisation. Such analysis assists us in uncovering the most impactful perspectives in the existing family firms' internationalisation literature, the evolving trends in the research streams, most productive authors, top journals and articles and co-citation analysis; thereby rendering our study different from recent review articles (Pukall and Calabro, 2014; Kontinen and Ojala, 2010b). Moreover, the present study is the first to review more than 100 research articles in the underlying area. The results from bibliometric analysis clearly indicate that family firms' internationalisation is an upcoming research area offering greater opportunities for future research scope, as there has been a sudden increase in the number of research articles. Also, notwithstanding, the first and third position by China and India (respectively) in the top list of family firms across the globe, there is a dearth of significant contribution in the area of internationalisation of family firms. It thereby indicates the need and opportunity for scholars from these nations to make significant contributions in the underlying research stream. Thematic analysis is carried out to identify the major themes and patterns evolving in the family firms' internationalisation research stream. The analysis rendered us with the creation of five major themes, i.e. family control, peculiar family characteristics, business environment, internationalisation theories and models and internationalisation process. Finally, on the basis of these themes, the study proposes a theoretical framework that would assist academicians, researchers and family business managers in accumulating a comprehensive and holistic view about various antecedents facilitating or restraining the family firms' internationalisation.

8. Limitations of the study

Our study contains certain limitations. Firstly, we have carried out a systematic review of the literature which might have led to some subjective biases, as different researchers have different mindsets and thought processes, ultimately leading to different results. In the future, this can be taken care of by using more statistically robust techniques like meta-analysis that would facilitate developing a more unbiased and clear understanding of the

underlying area. Secondly, although we have used multiple relevant combinations of words while selecting the research articles from the Scopus database, however, we might have overlooked some words leading to the exclusion of some relevant articles. Thirdly, the focus of our study was primarily on reviewing those research articles which have explored and analysed various antecedents either facilitating or restraining the family firms' internationalisation, thus, taking the degree or extent of internationalisation as the dependent variable. A more comprehensive review might consider analysing the research articles which have further explored the impact of family firms' internationalisation or their performance, thus taking internationalisation as the independent variable.

References

*Alayo, M., Maseda, A., Iturralde, T. and Arzubiaga, U. (2019), "Internationalization and entrepreneurial orientation of family SMEs: the influence of the family character", *International Business Review*, Vol. 28 No. 1, pp. 48-59, available at: https://doi.org/10.1016/j.ibusrev.2018.06.003

*Alessandri, T.M., Cerrato, D. and Eddleston, K.A. (2018), "The mixed gamble of internationalization in family and nonfamily firms: the moderating role of organizational slack", *Global Strategy Journal*, Vol. 8 No. 1, pp. 46-72, available at: https://doi.org/10.1002/gsj.1201

*Alonso, A.D. and O'Brien, S. (2017), "Family firms' management decision to export/not to export: a resource-based view", *Journal of Advances in Management Research*, Vol. 14 No. 3, pp. 270-287, available at: https://doi.org/10.1108/JAMR-06-2016-0053

*Anand, R. (2013), "Internationalisation of the small and medium family firm in Japan", *International Journal of Business and Globalisation*, Vol. 11 No. 2, pp. 117-135, available at: https://doi.org/10.1504/IJBG.2013.055598

Aria, M. and Cuccurullo, C. (2017), "Bibliometrix: an R-tool for comprehensive science mapping analysis", *Journal of Informetrics*, Vol. 11 No. 4, pp. 959-975, available at: https://doi.org/10.1016/j.joi.2017.08.007

Aria, M. and Cuccurullo, C. (2019), "A brief introduction to bibliometrix", available at: https://cran.rstudio.com/web/packages/bibliometrix/vignettes/bibliometrix-vignette.html (accessed 10 August 2019).

*Arregle, J.L., Hitt, M.A. and Mari, I. (2019), "A missing link in family firms' internationalization research: family structures", *Journal of International Business Studies*, Vol. 50 No. 5, pp. 809-825, available at: https://doi.org/10.1057/s41267-019-00213-z

*Arregle, J.L., Duran, P., Hitt, M. and van Essen, P. (2017), "Why is family firms' internationalization unique? A meta-analysis", *Entrepreneurship Theory and Practice*, Vol. 41 No. 5, pp. 801-831, available at: https://doi.org/10.1111/etap.12246

Arregle, J.L., Hitt, M.A., Sirmon, D.G. and Very, P. (2007), "The development of organizational social capital: attributes of family firms", *Journal of Management Studies*, Vol. 44 No. 1, pp. 73-95, available at: https://doi.org/10.1111/j.1467-6486.2007.00665.x

*Arregle, J.L., Naldi, L., Nordqvist, M. and Hitt, M.A. (2012), "Internationalization of family-controlled firms: a study of the effects of external involvement in governance", *Entrepreneurship Theory and Practice*, Vol. 36 No. 6, pp. 1115-1143, available at: https://doi.org/10.1111/j.1540-6520.2012.00541.x

*Avrichir, I., Meneses, R. and dos Santos, A.A. (2016), "Do family-managed and non-family-managed firms internationalize differently?", *Journal of Family Business Management*, Vol. 6 No. 3, pp. 330-349, available at: https://doi.org/10.1108/JFBM-02-2015-0014

Banalieva, E. and Eddleston, K. (2011), "Home-region focus and performance of family firms: the role of family vs non-family leaders", *Journal of International Business Studies*, Vol. 42 No. 8, pp. 1060-1072, available at: https://doi.org/10.1057/jibs.2011.28

*Basly, S. (2007), "The internationalization of family SME: an organizational learning and knowledge development perspective", *Baltic Journal of Management*, Vol. 2 No. 2, pp. 154-180, available at: https://doi.org/10.1108/17465260710750973

Berrone, P., Cruz, C. and Gomez-Mejia, L.R. (2012), "Socioemotional wealth in family firms: theoretical dimensions, assessment approaches, and agenda for future research", *Family Business Review*, Vol. 25 No. 3, pp. 258-279, available at: https://doi.org/10.1177/0894486511435355

Bhattacharyya, A.K. (2019), "Grounding of jet airways: misgovernance of family business", available at: www.business-standard.com/article/opinion/misgovernance-of-family-business-119051200748_1.html?fbclid=IwAR2S9m0AoAEtsptluD-cBwVbIQ95zeLA2x9yLEoVtglJ0Ey80f-nw3KMrUk (accessed 7 September 2019).

*Bhaumik, S.K. and Driffield, N. (2011), "Direction of outward FDI of EMNEs: evidence from the Indian pharmaceutical sector", *Thunderbird International Business Review*, Vol. 53 No. 5, pp. 615-628, available at: https://doi.org/10.1002/tie.20438

*Bhaumik, S.K., Driffield, N. and Pal, S. (2010), "Does ownership structure of emerging-market firms affect their outward FDI? The case of the Indian automotive and pharmaceutical sectors", *Journal of International Business Studies*, Vol. 41 No. 3, pp. 437-450, available at: https://doi.org/10.1057/jibs.2009.52

Boeker, W. and Karichalil, R. (2002), "Entrepreneurial transitions: factors influencing founder departure", *Academy of Management Journal*, Vol. 45 No. 4, pp. 818-825, available at: https://doi.org/10.5465/3069314

*Boellis, A., Mariotti, S., Minichilli, A. and Piscitello, L. (2016), "Family involvement and firms' establishment mode choice in foreign markets", *Journal of International Business Studies*, Vol. 47 No. 8, pp. 929-950, available at: https://doi.org/10.1057/jibs.2016.23

*Braga, V., Correia, A., Braga, A. and Lemos, S. (2017), "The innovation and internationalisation processes of family businesses", *Review of International Business and Strategy*, Vol. 27 No. 2, pp. 231-247, available at: https://doi.org/10.1108/RIBS-01-2017-0005

Braun, V. and Clarke, V. (2006), "Using thematic analysis in psychology", *Qualitative Research in Psychology*, Vol. 3 No. 2, pp. 77-101, doi: 10.1191/1478088706qp063oa.

*Calabro, A., Mussolino, D. and Huse, M. (2009), "The role of board of directors in the internationalisation process of small and medium sized family businesses", *International Journal of Globalisation and Small Business*, Vol. 3 No. 4, pp. 393-411, available at: https://ssrn.com/abstract=1534148

*Calabro, A., Brogi, M. and Torchia, M. (2016), "What does really matter in the internationalization of small and medium-sized family businesses?", *Journal of Small Business Management*, Vol. 54 No. 2, pp. 679-696, available at: https://doi.org/10.1111/jsbm.12165

*Calabro, A., Torchia, M., Pukall, T. and Mussolino, D. (2013), "The influence of ownership structure and board strategic involvement on international sales: the moderating effect of family involvement", *International Business Review*, Vol. 22 No. 3, pp. 509-523, available at: https://doi.org/10.1016/j.ibusrev.2012.07.002

*Calabro, A., Minola, T., Campopiano, G. and Pukall, T. (2016), "Turning innovativeness into domestic and international corporate venturing: the moderating effect of high family ownership and influence", *European J. Of International Management*, Vol. 10 No. 5, pp. 505-533, available at: https://doi.org/10.1504/EJIM.2016.078789

*Calabro, A., Campopiano, G., Basco, R. and Pukall, T. (2017), "Governance structure and internationalization of family-controlled firms: the mediating role of international entrepreneurial orientation", *European Management Journal*, Vol. 35 No. 2, pp. 238-248, available at: https://doi.org/10.1016/j.emj.2016.04.007

*Casillas, J.C. and Acedo, F.J. (2005), "Internationalisation of spanish family SMEs: an analysis of family involvement", *International Journal of Globalisation and Small Business*, Vol. 1 No. 2, pp. 134-151, available at: https://doi.org/10.1504/IJGSB.2005.008010

Cerrato, D. and Piva, M. (2012), "The internationalization of small and medium-sized enterprises: the effect of family management, human capital and foreign ownership", *Journal of Management and Governance*, Vol. 16 No. 4, pp. 617-644, available at: https://doi.org/10.1007/s10997-010-9166-x

*Cesinger, B., Bouncken, R., Fredrich, V. and Kraus, S. (2014), "The alchemy of family enterprises' internationalisation: dexterous movers or prodigal laggards?", *European J. Of International Management*, Vol. 8 No. 6, pp. 671-696, available at: https://doi.org/10.1504/EJIM.2014.064902

*Cesinger, B., Hughes, M., Mensching, H., Bouncken, R., Fredrich, V. and Kraus, S. (2016), "A socioemotional wealth perspective on how collaboration intensity, trust, and international market knowledge affect family firms' multinationality", *Journal of World Business*, Vol. 51 No. 4, pp. 586-599, available at: https://doi.org/10.1016/j.jwb.2016.02.004

*Cheong, K.C., Lee, P.P. and Lee, K.H. (2015), "The internationalisation of family firms: case histories of two chinese overseas family firms", *Business History*, Vol. 57 No. 6, pp. 841-861, doi: 10.1080/00076791.2014.982106.

Chrisman, J.J., Chua, J.H. and Sharma, P. (2005), "Trends and directions in the development of a strategic management theory of the family firm", *Entrepreneurship Theory and Practice*, Vol. 29 No. 5, pp. 555-576, available at: https://doi.org/10.1111/j.1540-6520.2005.00098.x

Chua, J.H., Chrisman, J.J., Steier, L.P. and Rau, S.B. (2012), "Sources of heterogeneity in family firms: an introduction", *Entrepreneurship Theory and Practice*, Vol. 36 No. 6, pp. 1103-1113, available at: https://doi.org/10.1111/j.1540-6520.2012.00540.x

*Claver, E., Rienda, L. and Quer, D. (2007), "The internationalisation process in family firms: choice of market entry strategies", *Journal of General Management*, Vol. 33 No. 1, pp. 1-14, available at: https://doi.org/10.1177/030630700703300101

*Claver, E., Rienda, L. and Quer, D. (2009), "Family firms' international commitment: the influence of family-related factors", *Family Business Review*, Vol. 22 No. 2, pp. 125-135, available at: https://doi.org/10.1177/0894486508330054

*Colli, A., Garcia-Canal, E. and Guillen, M.F. (2013), "Family character and international entrepreneurship: a historical comparison of Italian and Spanish 'new multinationals'", *Business History*, Vol. 55 No. 1, pp. 119-138, available at: https://doi.org/10.1080/00076791.2012.687536

Cook, D.J., Mulrow, C.D. and Haynes, R.B. (1997), "Systematic reviews: synthesis of best evidence for clinical decisions", *Annals of Internal Medicine*, Vol. 126 No. 5, pp. 376-380, doi: 10.7326/0003-4819-126-5-199703010-00006.

Cook, D.J., Greengold, N.L., Ellrodt, A.G. and Weingarten, S.R. (1997), "The relation between systematic reviews and practice guidelines", *Annals of Internal Medicine*, Vol. 127 No. 3, pp. 210-216, doi: 10.7326/0003-4819-127-3-199708010-00006.

*Corsi, C. and Prencipe, A. (2018), "Innovation propensity and international development of small and medium firms: the moderating effects of corporate governance structure", *International Journal of Entrepreneurship and Innovation Management*, Vol. 22 Nos 1/2, pp. 80-102, doi: 10.1504/IJEIM.2018.089717.

Credit Suisse Research Institute (2018), *The CS Family 1000 in 2018*, Credit Suisse Research Institute.

*D'Angelo, A., Majocchi, A. and Buck, T. (2016), "External managers, family ownership and the scope of SME internationalization", *Journal of World Business*, Vol. 51 No. 4, pp. 534-547, available at: https://doi.org/10.1016/j.jwb.2016.01.004

*Davis, P.S. and Harveston, P.D. (2000), "Internationalization and organizational growth: the impact of internet usage and technology involvement among entrepreneur-led family businesses", *Family Business Review*, Vol. 13 No. 2, pp. 107-120, available at: https://doi.org/10.1111/j.1741-6248.2000.00107.x

Debicki, B.J., Kellermanns, F.W., Chrisman, J.J., Pearson, A.W. and Spencer, B.A. (2016), "Development of a socioemotional wealth importance (SEWi) scale for family firm research", *Journal of Family Business Strategy*, Vol. 7 No. 1, pp. 47-57, available at: https://doi.org/10.1016/j.jfbs.2016.01.002

De Massis, A., Frattini, F., Majocchi, A. and Piscitello, L. (2018), "Family firms in the global economy: toward a deeper understanding of internationalization determinants, processes and outcomes", *Global Strategy Journal*, Vol. 8 No. 1, pp. 3-21, available at: https://doi.org/10.1002/gsj.1199

Deloitte Family Business Center (2018), "Next-generation family businesses: exploring business ecosystems", available at: www2.deloitte.com/content/dam/Deloitte/lu/Documents/strategy/lu-next-generation-family-businesses.pdf (accessed 5 September 2019).

Donnelley, R.G. (1964), "The family business", *Harvard Business Review*, Vol. 42, pp. 93-105.

*Dou, J., Jacoby, G., Li, J., Su, Y. and Wu, Z. (2019), "Family involvement and family firm internationalization: the moderating effects of board experience and geographical distance", *Journal of International Financial Markets, Institutions and Money*, Vol. 59, pp. 250-261, available at: https://doi.org/10.1016/j.intfin.2018.12.004

*Eddleston, K.A., Sarathy, R. and Banalieva, E.R. (2018), "When a high-quality niche strategy is not enough to spur family-firm internationalization: the role of external and internal contexts", *Journal of International Business Studies*, Vol. 50 No. 5, pp. 783-808, available at: https://doi.org/10.1057/s41267-018-0199-8

*Evert, R.E., Sears, J.B., Martin, J.A. and Payne, G.T. (2018), "Family ownership and family involvement as antecedents of strategic action: a longitudinal study of initial international entry", *Journal of Business Research*, Vol. 84, pp. 301-311, available at: https://doi.org/10.1016/j.jbusres.2017.07.019

*Fang, H., Kotlar, J., Memili, E., Chrisman, J.J. and Massis, A.D. (2018), "The pursuit of international opportunities in family firms: generational differences and the role of knowledge-based resources", *Global Strategy Journal*, Vol. 8 No. 1, pp. 136-157, available at: https://doi.org/10.1002/gsj.1197

*Fernández, Z. and Nieto, M.J. (2005), "Internationalization strategy of small and medium-sized family businesses: some influential factors", *Family Business Review*, Vol. 18 No. 1, pp. 77-89, available at: https://doi.org/10.1111/j.1741-6248.2005.00031.x

*Fernandez, Z. and Nieto, M.J. (2006), "Impact of ownership on the international involvement of SMEs", *Journal of International Business Studies*, Vol. 37 No. 3, pp. 340-351, available at: https://doi.org/10.1057/palgrave.jibs.8400196

*Gallo, M.A. and Sveen, J. (1991), "Internationalizing the family business: facilitating and restraining factors", *Family Business Review*, Vol. 4 No. 2, pp. 181-190, available at: https://doi.org/10.1111/j.1741-6248.1991.00181.x

*Gallo, M.A. and Pont, C.G. (1996), "Important factors in family business internationalization", *Family Business Review*, Vol. 9 No. 1, pp. 45-59, available at: https://doi.org/10.1111/j.1741-6248.1996.00045.x

*Gomez-Mejia, L.R., Makri, M. and Kintana, M.L. (2010), "Diversification decisions in family-controlled firms", *Journal of Management Studies*, Vol. 47 No. 2, pp. 223-252, available at: https://doi.org/10.1111/j.1467-6486.2009.00889.x

*Graves, C. and Thomas, J. (2004), "Internationalisation of the family business: a longitudinal perspective", *International Journal of Globalisation and Small Business*, Vol. 1 No. 1, pp. 7-27.

*Graves, C. and Thomas, J. (2006), "Internationalization of Australian family businesses: a managerial capabilities perspective", *Family Business Review*, Vol. 19 No. 3, pp. 207-224, available at: https://doi.org/10.1111/j.1741-6248.2006.00066.x

*Graves, C. and Thomas, J. (2008), "Determinants of the internationalization pathways of family firms: an examination of family influence", *Family Business Review*, Vol. 21 No. 2, pp. 151-167, available at: https://doi.org/10.1111/j.1741-6248.2008.00119.x

*Hernandez, V., Nieto, M.J. and Boellis, A. (2018), "The asymmetric effect of institutional distance on international location: family versus nonfamily firms", *Global Strategy Journal*, Vol. 8 No. 1, pp. 22-45, available at: https://doi.org/10.1002/gsj.1203

Hernandez-Linares, R., Sarkar, S. and Cobo, M.J. (2018), "Inspecting the achilles heel: a quantitative analysis of 50 years of family business definitions", *Scientometrics*, Vol. 115 No. 2, pp. 929-951, available at: https://doi.org/10.1007/s11192-018-2702-1

Hiebl, M. and Mayrleitner, B. (2019), "Professionalization of management accounting in family firms: the impact of family members", *Review of Managerial Science*, Vol. 13 No. 5, pp. 1037-1068, available at: https://doi.org/10.1007/s11846-017-0274-8

*Jimenez, A., Majocchi, A. and Piana, B.D. (2019), "Not all family firms are equal: the moderating effect of family involvement on the political risk exposure of the foreign direct investment portfolio. Preliminary evidence from Spanish multinational enterprises", *Thunderbird International Business Review*, Vol. 61 No. 2, pp. 309-323, available at: https://doi.org/10.1002/tie.22032

*Kano, L. and Verbeke, A. (2018), "Family firm internationalization: heritage assets and the impact of bifurcation bias", *Global Strategy Journal*, Vol. 8 No. 1, pp. 158-183, available at: https://doi.org/10.1002/gsj.1186

*Kao, M.S. and Kuo, A. (2017), "The effect of uncertainty on FDI entry mode decisions: the influence of family ownership and involvement in the board of directors", *Journal of Family Business Strategy*, Vol. 8 No. 4, pp. 224-236, available at: https://doi.org/10.1016/j.jfbs.2017.09.003

*Kao, M.S., Kuo, A. and Chang, Y.C. (2013), "How family control influences FDI entry mode choice", *Journal of Management and Organization*, Vol. 19 No. 4, pp. 367-385, doi: 10.1017/jmo.2013.23.

*Kontinen, T. and Ojala, A. (2010a), "Internationalization pathways of family SMEs: psychic distance as a focal point", *Journal of Small Business and Enterprise Development*, Vol. 17 No. 3, pp. 437-454, available at: https://doi.org/10.1108/14626001011068725

Kontinen, T. and Ojala, A. (2010b), "The internationalization of family businesses: a review of extant research", *Journal of Family Business Strategy*, Vol. 1 No. 2, pp. 97-107, available at: https://doi.org/10.1016/j.jfbs.2010.04.001

Kubíček, A. and Machek, O. (2019), "Gender-related factors in family business succession: a systematic literature review", *Review of Managerial Science*, Vol. 13 No. 5, pp. 963-1002, available at: https://doi.org/10.1007/s11846-018-0278-z

*Kuo, A., Kao, M.S., Chang, Y.C. and Chiu, C.F. (2012), "The influence of international experience on entry mode choice: difference between family and non-family firms", *European Management Journal*, Vol. 30 No. 3, pp. 248-263, available at: https://doi.org/10.1016/j.emj.2012.03.008

*Liang, X., Wang, L. and Cui, Z. (2014), "Chinese private firms and internationalization: effects of family involvement in management and family ownership", *Family Business Review*, Vol. 27 No. 2, pp. 126-141, available at: https://doi.org/10.1177/0894486513480885

*Majocchi, A., D'Angelo, A., Forlani, E. and Buck, T. (2018), "Bifurcation bias and exporting: Can foreign work experience be an answer? Insight from European family SMEs", *Journal of World Business*, Vol. 53 No. 2, pp. 237-247, available at: https://doi.org/10.1016/j.jwb.2017.11.005

*Marinova, S. and Marinov, M. (2017), "Inducing the internationalisation of family manufacturing firms from a transition context", *European Business Review*, Vol. 29 No. 2, pp. 181-204, available at: https://doi.org/10.1108/EBR-07-2016-0085

Martin, G.P., Gomez-Mejia, L.R. and Wiseman, R.M. (2013), "Executive stock options as mixed gambles: revisiting the behavioral agency model", *Academy of Management Journal*, Vol. 56 No. 2, pp. 451-472, available at: https://doi.org/10.5465/amj.2010.0967

*Menéndez-Requejo, S. (2005), "Growth and internationalisation of family businesses", *International Journal of Globalisation and Small Business*, Vol. 1 No. 2, pp. 122-133, available at: https://doi.org/10.1504/IJGSB.2005.008009

*Mitter, C., Duller, C., Feldbauer-Durstmuller, B. and Kraus, S. (2014), "Internationalization of family firms: the effect of ownership and governance", *Review of Managerial Science*, Vol. 8 No. 1, pp. 1-28, available at: https://doi.org/10.1007/s11846-012-0093-x

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G. and The Prisma, G. (2009), "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement", *Annals of Internal Medicine*, Vol. 151 No. 4, pp. 264-269, doi: 10.1371/journal.pmed.1000097.

*Monreal-Pérez, J. and Sánchez-Marín, G. (2017), "Does transitioning from family to non-family controlled firm influence internationalization?", *Journal of Small Business and Enterprise Development*, Vol. 24 No. 4, pp. 775-792, available at: https://doi.org/10.1108/JSBED-02-2017-0029

*Okoroafo, S.C. (1999), "Internationalization of family businesses: evidence from northwest Ohio, USA", *Family Business Review*, Vol. 12 No. 2, pp. 147-158, available at: https://doi.org/10.1111/j.1741-6248.1999.00147.x

Parris, D.L. and Peachey, J.W. (2013), "A systematic literature review of servant leadership theory in organizational contexts", *Journal of Business Ethics*, Vol. 113 No. 3, pp. 377-393, available at: https://doi.org/10.1007/s10551-012-1322-6

*Pongelli, C., Calabro, A. and Basco, R. (2019), "Family firms' international make-or-buy decisions: captive offshoring, offshore outsourcing, and the role of home region focus", *Journal of Business Research*, Vol. 103, pp. 596-606. https://doi.org/10.1016/j.jbusres.2018.02.033.

*Pukall, T.J. and Calabro, A. (2014), "The internationalization of family firms: a critical review and integrative model", *Family Business Review*, Vol. 27 No. 2, pp. 103-125, available at: https://doi.org/10.1177/0894486513491423

*Ramón-Llorens, M.C., García-Meca, E. and Duréndez, A. (2017), "Influence of CEO characteristics in family firms internationalization", *International Business Review*, Vol. 26 No. 4, pp. 786-799, available at: https://doi.org/10.1016/j.ibusrev.2017.01.007

*Ratten, V., Dana, L.P. and Ramadani, V. (2017), "Internationalisation of family business groups in transition economies", *International Journal of Entrepreneurship and Small Business*, Vol. 30 No. 4, pp. 509-525, available at: https://doi.org/10.1504/IJESB.2017.082896

*Ratten, V., Ramadani, V., Dana, L.P., Hoy, F. and Ferreira, J. (2017), "Family entrepreneurship and internationalization strategies", *Review of International Business and Strategy*, Vol. 27 No. 2, pp. 150-160, doi: 10.1108/RIBS-01-2017-0007.

*Ray, S., Mondal, A. and Ramachandran, K. (2018), "How does family involvement affect a firm's internationalization? An investigation of Indian family firms", *Global Strategy Journal*, Vol. 8 No. 1, pp. 73-105, available at: https://doi.org/10.1002/gsj.1196

*Rexhepi, G., Ramadani, V., Rahdari, A. and Anggadwita, G. (2017), "Models and strategies of family businesses internationalization: a conceptual framework and future research directions", *Review of International Business and Strategy*, Vol. 27 No. 2, pp. 248-260, available at: https://doi.org/10.1108/RIBS-12-2016-0081

RStudio Team (2020), "RStudio: integrated development environment for R", RStudio, PBC, Boston, MA, available at: www.rstudio.com/

*Sanchez-Bueno, M.J. and Usero, B. (2014), "How may the nature of family firms explain the decisions concerning international diversification?", *Journal of Business Research*, Vol. 67 No. 7, pp. 1311-1320, available at: https://doi.org/10.1016/j.jbusres.2013.09.003

Sánchez-Marín, G., Pemartín, M. and Monreal-Pérez, J. (2020), "The influence of family involvement and generational stage on learning-by-exporting among family firms", *Review of Managerial Science*, Vol. 14 No. 1, pp. 311-334, available at: https://doi.org/10.1007/s11846-019-00350-7

*Sciascia, S., Mazzola, P., Astrachan, J.H. and Pieper, T.M. (2013), "Family involvement in the board of directors: effects on sales internationalization", *Journal of Small Business Management*, Vol. 51 No. 1, pp. 83-99, doi: 10.1111/j.1540-627X.2012.00373.x.

*Shi, H.X., Graves, C. and Barbera, F. (2019), "Intergenerational succession and internationalisation strategy of family SMEs: evidence from China", *Long Range Planning*, Vol. 52 No. 4, available at: https://doi.org/10.1016/j.lrp.2018.05.004

Srinivasan, S., Arora, N. and Goldberg, J.S. (2019), Dabur India: growing professional management from family roots, HBS No. 9-119-089, Harvard Business School Publishing, Boston, MA.

*Swinth, R.L. and Vinton, K.L. (1993), "Do family-owned businesses have a strategic advantage in international joint ventures?", *Family Business Review*, Vol. 6 No. 1, pp. 19-30, available at: https://doi.org/10.1111/j.1741-6248.1993.00019.x

*Thite, M. (2013), "The growing global footprint of Indian multinationals: an interview with KM Birla, Chairman, Aditya Birla group", *Journal of Indian Business Research*, Vol. 5 No. 2, pp. 144-151, available at: https://doi.org/10.1108/17554191311320791

*Tsao, C.W., Wang, M.J., Lu, C.M., Chen, S.J. and Wang, Y.H. (2018), "Internationalization propensity in family-controlled public firms in emerging markets: the effects of family ownership, governance, and top management team heterogeneity", *Journal of Small Business Strategy*, Vol. 28 No. 1, pp. 28-37, available at: https://libjournals.mtsu.edu/index.php/jsbs/article/view/1078

Ward, J. (2006), "The Harilela enterprises: an Indian family business in Hong Kong", KEL No. 249, Kellogg School of Management, Evanston, IL.

*Yeoh, P., L. (2014), "Internationalization and performance outcomes of entrepreneurial family SMEs: the role of outside CEOs, technology sourcing, and innovation", *Thunderbird International Business Review*, Vol. 56 No. 1, pp. 77-96, available at: https://doi.org/10.1002/tie.21597

*Zahra, S.A. (2003), "International expansion of US manufacturing family businesses: the effect of ownership and involvement", *Journal of Business Venturing*, Vol. 18 No. 4, pp. 495-512, available at: https://doi.org/10.1016/S0883-9026(03)00057-0

*Zaefarian, R., Eng, T.Y. and Tasavori, M. (2016), "An exploratory study of international opportunity identification among family firms", *International Business Review*, Vol. 25 No. 1, pp. 333-345, available at: https://doi.org/10.1016/j.ibusrev.2015.06.002

*Zain, M. and Kassim, N.M. (2012), "Strategies of family businesses in a newly globalized developing economy", *Journal of Family Business Management*, Vol. 2 No. 2, pp. 147-165, available at: https://doi.org/10.1108/20436231211261880

Further reading

*Abdellatif, M., Amann, B. and Jaussau, J. (2010), "Family versus nonfamily business: a comparison of international strategies", *Journal of Family Business Strategy*, Vol. 1 No. 2, pp. 108-116, available at: https://doi.org/10.1016/j.jfbs.2010.04.004

*Banno, M. and Trento, S. (2017), "International expansion of family firms: the moderating role of successors and external managers", *International Journal of Globalisation and Small Business*, Vol. 8 No. 4, pp. 292-315, available at: https://doi.org/10.1504/IJGSB.2016.081420

*Baronchelli, G., Bettinellia, C., Boscob, B.D. and Loane, S. (2016), "The impact of family involvement on the investments of Italian small-medium enterprises in psychically distant countries", *International Business Review*, Vol. 25 No. 4, pp. 960-970, available at: https://doi.org/10.1016/j.ibusrev.2016.01.011

*Bassetti, T., Masob, L.D. and Lattanzi, N. (2015), "Family businesses in Eastern European countries: How informal payments affect exports", *Journal of Family Business Strategy*, Vol. 6 No. 4, pp. 219-233, available at: https://doi.org/10.1016/j.jfbs.2015.07.004

*Benito-Hernandez, S., Lopez-Cozar-Navarro, C. and Priede-Bergamini, T. (2014), "Factors determining exportation and internationalization in family businesses: the importance of debt", *South African Journal of Business Management*, Vol. 45 No. 1, pp. 13-25.

*Bobillo, A.M., Rodríguez-Sanz, J.A. and Tejerina-Gaite, F. (2013), "Shareholder activism and internationalization in the family firm", *Journal of Business Economics and Management*, Vol. 14 No. 5, pp. 867-885, available at: http://dx.doi.org/10.3846/16111699.2012.707621

*Boers, B. (2017), "Go east! how family businesses choose markets and entry modes when internationalising", *International Journal of Globalisation and Small Business*, Vol. 8 No. 4, pp. 333-353, available at: https://doi.org/10.1504/IJGSB.2016.081421

*Boyd, B. and Dyhr Ulrich, A.M. (2014), "Market entry strategies into the BRIC countries: a comparison of danish family and non-family businesses", *International Journal of Globalisation and Small Business*, Vol. 6 No. 1, pp. 15-36, doi: 10.1504/IJGSB.2014.063405.

*Carr, C. and Bateman, S. (2009), "International strategy configurations of the world's top family firms: another factor affecting performance", *Management International Review*, Vol. 49 No. 6, pp. 733-758, doi: 10.1007/s 11 575-009-00 18-3.

*Chen, Y.M., Liu, H.H., Ni, Y.T. and Wu, M.F. (2015), "A rational normative model of international expansion: strategic intent perspective, market positions, and founder CEOs/family-successor CEOs", *Journal of Business Research*, Vol. 68 No. 7, pp. 1539-1543, available at: https://doi.org/10.1016/j.jbusres.2015.01.048

*Chung, H.M. and Dahms, S. (2016), "Particularistic ties and internationalisation: evidence from taiwanese family business groups", *European J. Of International Management*, Vol. 10 No. 5, pp. 558-580, available at: https://doi.org/10.1504/EJIM.2016.078792

*Claver, E., Rienda, L. and Quer, D. (2008), "Family firms' risk perception: empirical evidence on the internationalization process", *Journal of Small Business and Enterprise Development*, Vol. 15 No. 3, pp. 457-471, available at: https://doi.org/10.1108/1462600081089228

*Crick, D. and Chaudhry, S. (2013), "An exploratory study of UK based, family-owned, Asian firms' motives for internationalising", *Journal of Small Business and Enterprise Development*, Vol. 20 No. 3, pp. 526-547, available at: https://doi.org/10.1108/JSBED-04-2013-0051

*Crick, D., Bradshaw, R. and Chaudhry, S. (2006), "Successful" internationalising UK family and non-family-owned firms: a comparative study", *Journal of Small Business and Enterprise Development*, Vol. 13 No. 4, pp. 498-512, available at: https://doi.org/10.1108/14626000610705714

*Deng, F.J., Huang, L.Y., Carraher, S. and Duan, J. (2009), "International expansion of family firms: an integrative framework using Taiwanese manufacturers", *Academy of Entrepreneurship Journal*, Vol. 15 No. 1, pp. 25-42.

*Durana, P., Kostovab, T. and Essen, M.V. (2017), "Political ideologies and the internationalization of family-controlled firms", *Journal of World Business*, Vol. 52 No. 4, pp. 474-488, available at: https://doi.org/10.1016/j.jwb.2016.07.001

*Farias, S.L., Nataraajan, R. and Kovacs, E.P. (2009), "Global business partnering among family-owned enterprises", *Journal of Business Research*, Vol. 62 No. 6, pp. 667-672, doi: 10.1016/j.jbusres.2008.02.003.

*Hadryś-Nowak, A. (2018), "Family entrepreneurship orientation in family owned SMEs: a key resource for internationalization?", *Entrepreneurial Business and Economics Review*, Vol. 6 No. 2, pp. 153-169, available at: https://doi.org/10.15678/EBER.2018.060208

*Hernandez-Perlines, F. (2018), "Moderating effect of absorptive capacity on the entrepreneurial orientation of international performance of family businesses", *Journal of Family Business Management*, Vol. 8 No. 1, pp. 58-74, available at: https://doi.org/10.1108/JFBM-10

*Hernández-Perlines, F. and Xu, W. (2018), "Conditional mediation of absorptive capacity and environment in international entrepreneurial orientation of family businesses", *Frontiers in Psychology*, Vol. 9, pp. 1-14, doi: 10.3389/fpsyg.2018.00102.

*Hernández-Perlines, F., Moreno-García, J. and Yañez-Araque, B. (2016), "The mediating role of competitive strategy in international entrepreneurial orientation", *Journal of Business Research*, Vol. 69 No. 11, pp. 5383-5389, available at: https://doi.org/10.1016/j.jbusres.2016.04.142

*Herrera-Echeverri, H., Geleilate, J.G., Gaitan-Riaño, S., Haar, J. and Soto-Echeverry, N. (2016), "Export behavior and board independence in Colombian family firms: the reverse causality relationship", *Journal of Business Research*, Vol. 69 No. 6, pp. 2018-2029, available at: https://doi.org/10.1016/j.jbusres.2015.10.147

*Holt, D.T. (2012), "Article commentary: Strategic decisions within family firms: understanding the controlling family's receptivity to internationalization", *Entrepreneurship Theory and Practice*, Vol. 36 No. 6, pp. 1145-1151, available at: https://doi.org/10.1111%2Fj.1540-6520.2012.00542.x

*Horgos, D. (2013), "Global sourcing: a family-firm's perspective", *Journal of Small Business and Entrepreneurship*, Vol. 26 No. 3, pp. 221-240, available at: http://dx.doi.org/10.1080/08276331.2013.808028

*Huang, K.P., Lo, S.C., Liu, C.M. and Tung, J. (2014), "Internationalisation of family business: the effect of ownership and generation involvement", *The Anthropologist*, Vol. 17 No. 3, pp. 757-767, doi: 10.1080/09720073.2014.11891490.

*Janjuha-Jivraj, S., Martin, L. and Danko, A. (2012), "Internationalization of a 'born-again global': How a family-led crisis enabled the realization of internationalization opportunities", *Journal of Small Business and Entrepreneurship*, Vol. 25 No. 2, pp. 201-215, available at: https://doi.org/10.1080/08276331.2012.10593569

*Jorge, M., Couto, M., Veloso, T. and Franco, M. (2017), "When family businesses go international: management sets the path", *Journal of Business Strategy*, Vol. 38 No. 1, pp. 31-38, available at: https://doi.org/10.1108/JBS-01-2016-0006

*Kontinen, T. and Ojala, A. (2011a), "International opportunity recognition among small and medium-sized family firms", *Journal of Small Business Management*, Vol. 49 No. 3, pp. 490-514.

*Kontinen, T. and Ojala, A. (2011b), "Social capital in relation to the foreign market entry and post-entry operations of family SMEs", *Journal of International Entrepreneurship*, Vol. 9 No. 2, pp. 133-151, available at: https://doi.org/10.1007/s10843-010-0072-8

*Kontinen, T. and Ojala, A. (2012), "Social capital in the international operations of family SMEs", *Journal of Small Business and Enterprise Development*, Vol. 19 No. 1, pp. 39-55, available at: https://doi.org/10.1108/14626001211196398

*Kraus, S., Mensching, H., Calabrò, A., Cheng, C.F. and Filser, M. (2016), "Family firm internationalization: a configurational approach", *Journal of Business Research*, Vol. 69 No. 11, pp. 5473-5478, available at: https://doi.org/10.1016/j.jbusres.2016.04.158

*Laffranchini, G., Hadjimarcou, J.S., Kim, S.H. and Braun, M. (2016), "The internationalization of family-firms: a signal detection approach", *Journal of Family Business Management*, Vol. 6 No. 3, pp. 291-309, available at: https://doi.org/10.1108/JFBM-11-2015-0039

*Larimo, J. (2013), "Small and medium-size enterprise export performance", *International Studies of Management and Organization*, Vol. 43 No. 2, pp. 79-100, doi: 10.2753/IMO0020-8825430204.

*Lehrer, M. and Celo, S. (2017), "Boundary-spanning and boundary-buffering in global markets: a German perspective on the internationalization of family firms", *Review of International Business and Strategy*, Vol. 27 No. 2, pp. 161-179, doi: 10.1108/RIBS-09-2016-0053.

*Liu, C.M. (2014), "Internationalisation of family firm: the role of entrepreneurial orientation, ownership and generational involvement", *Revista de Cercetare şi Intervenţie Socială*, Vol. 47, pp. 180-191.

*Marin, Q., Hernández-Lara, A.B., Campa-Planas, F. and Sánchez-Rebull, M.V. (2017), "Which factors improve the performance of the internationalization process? Focus on family firms", *Applied Economics*, Vol. 49 No. 32, pp. 3181-3194, doi: 10.1080/00036846.2016.1257103.

*Memili, E., Misra, K., Chrisman, J.J. and Welsh, D.H. (2017), "Internationalisation of publicly traded family firms: a transaction cost theory perspective and longitudinal analysis", *International Journal of Management and Enterprise Development*, Vol. 16 Nos 1/2, pp. 80-108, available at: https://doi.org/10.1504/IJMED.2017.082538

*Meneses, R., Coutinho, R. and Carlos Pinho, J. (2014), "The impact of succession on family business internationalisation: the successors' perspective", *Journal of Family Business Management*, Vol. 4 No. 1, pp. 24-45, available at: https://doi.org/10.1108/JFBM-01-2013-0

*Mensching, H., Calabro, A., Eggers, F. and Kraus, S. (2016), "Internationalisation of family and non-family firms: a conjoint experiment among CEOs", *European J. Of International Management*, Vol. 10 No. 5, pp. 581-604, available at: https://doi.org/10.1504/EJIM.2016.078795

*Merino, F., Monreal-Pérez, J. and Sánchez-Marín, G. (2015), "Family SMEs' internationalization: disentangling the influence of familiness on spanish firms' export activity", *Journal of Small Business Management*, Vol. 53 No. 4, pp. 1164-1184, doi: 10.1111/jsbm.12111.

*Minetti, R., Murro, P. and Zhu, S.C. (2015), "Family firms", *Economica*", , Vol. 82 No. 1, pp. 1177-1216, available at: https://doi.org/10.1111/ecca.12156

*Moini, H., Kalouda, F. and Tesar, G. (2010), "Internationalisation of family-owned businesses in the Czech Republic", *International Journal of Entrepreneurial Venturing*, Vol. 2 Nos 3/4, pp. 400-418, available at: https://doi.org/10.1504/IJEV.2010.03712

*Pacheco, L.M. (2017), "Internationalization vs family ownership and management: the case of Portuguese wine firms", *International Journal of Wine Business Research*, Vol. 29 No. 2, pp. 1751-1062, doi: 10.1108/IJWBR-10-2016-0034.

*Pascucci, F. and Bartoloni, S. (2018), "Explaining the internationalisation pathways of family firms: a qualitative research", *International Journal of Business and Globalisation*, Vol. 20 No. 4, pp. 537-556, available at: https://doi.org/10.1504/IJBG.2018.093594

*Patel, V.K. and Pieper, T.M. (2012), "The global family business: challenges and drivers for cross-border growth", *Business Horizons*, Vol. 55 No. 3, pp. 231-239, available at: https://doi.org/10.1016/j.bushor.2012.01.002

*Piva, E., Rossi-Lamastra, C. and Massis, A.D. (2013), "Family firms and internationalization: an exploratory study on high-tech entrepreneurial ventures", *Journal of International Entrepreneurship*, Vol. 11 No. 2, pp. 108-129, available at: https://doi.org/10.1007/s10843-012-0100-y

*Pongelli, C., Caroli, M.G. and Cucculelli, M. (2016), "Family business going abroad: the effect of family ownership on foreign market entry mode decisions", *Small Business Economics*, Vol. 47 No. 3, pp. 787-801, available at: https://doi.org/10.1007/s11187-016-9763-4

*Procher, V.D., Urbig, D. and Volkmann, C. (2013), "Time to BRIC it? Internationalization of European family firms in Europe, North America and the BRIC countries", *Applied Economics Letters*, Vol. 20 No. 16, pp. 1466-1471, doi: 10.1080/13504851.2013.815302.

*Puig, N. and Pérez, P.F. (2009), "A silent revolution: the internationalisation of large Spanish family firms", *Business History*, Vol. 51 No. 3, pp. 462-483, doi: 10.1080/00076790902844088.

*Ratten, V. and Tajeddini, K. (2017), "Innovativeness in family firms: an internationalization approach", *Review of International Business and Strategy*, Vol. 27 No. 2, pp. 217-230, doi: 10.1108/RIBS-12-2016-0085.

*Reuber, A.R. (2016), "An assemblage–theoretic perspective on the internationalization processes of family firms", *Entrepreneurship Theory and Practice*, Vol. 40 No. 6, pp. 1269-1286, available at: https://doi.org/10.1111%2Fetap.12243

*Santulli, R., Torchia, M., Calabro, A. and Gallucci, C. (2019), "Family ownership concentration and firm internationalization: integrating principal-principal and socioemotional wealth perspectives", *Journal of International Entrepreneurship*, Vol. 17 No. 2, pp. 220-248, available at: https://doi.org/10.1007/s10843-019-00245-4

*Scholes, L., Mustafa, M. and Chen, S. (2016), "Internationalization of small family firms: the influence of family from a socioemotional wealth perspective", *Thunderbird International Business Review*, Vol. 58 No. 2, pp. 131-146, available at: https://doi.org/10.1002/tie.21729

*Sciascia, S., Mazzola, P., Astrachan, J.H. and Pieper, T.M. (2012), "The role of family ownership in international entrepreneurship: exploring nonlinear effects", *Small Business Economics*, Vol. 38 No. 1, pp. 15-31, doi: 10.1007/sl 1187-010-9264-9.

*Segaro, E.L., Larimo, J. and Jones, M.V. (2014), "Internationalisation of family small and medium sized enterprises: the role of stewardship orientation, family commitment culture and top management team", *International Business Review*, Vol. 23 No. 2, pp. 381-395, available at: https://doi.org/10.1016/j.ibusrev.2013.06.004

*Selekler-Gökþen, N. and Yýldýrým-Öktem, O. (2008), "The impact of board characteristics on the internationalization of business group affiliates", *Bogazici Journal, Review of Social, Economic and Administrative Studies*, Vol. 22 Nos 1/2, pp. 53-71.

*Singh, H. and Wyrobek, J. (2013), "Do agency concerns inhibit internationalisation of family owned firms?: a comparative study of the USA and Poland", *International Journal of Business Innovation and Research*, Vol. 7 No. 3, pp. 279-297, available at: https://doi.org/10.1504/IJBIR.2013.053632

*Singh, R. and Kota, H.B. (2017), "A resource dependency framework for innovation and internationalization of family businesses: Evidence from India", *Journal of Entrepreneurship in Emerging Economies*, Vol. 9 No. 2, pp. 207-231, available at: https://doi.org/10.1108/JEEE-04-2016-0013

Sirmon, D., Arregle, J., Hitt, M. and Webb, J. (2008), "The role of family influence in firms' strategic responses to threat of imitation", *Entrepreneurship Theory and Practice*, Vol. 32 No. 6, pp. 979-998, available at: https://doi.org/10.1111/j.1540-6520.2008.00267.x

*Stadler, C., Mayer, M.C., Hautz, J. and Matzler, K. (2018), "International and product diversification: which strategy suits family managers?", *Global Strategy Journal*, Vol. 8 No. 1, pp. 184-207, available at: https://doi.org/10.1002/gsj.1190

*Stieg, P., Cesinger, B., Apfelthaler, G., Kraus, S. and Cheng, C.F. (2018), "Antecedents of successful internationalization in family and non-family firms: how knowledge resources and collaboration intensity shape international performance", *Journal of Small Business Strategy*, Vol. 28 No. 1, pp. 14-27, available at: https://libjournals.mtsu.edu/index.php/jsbs/article/view/1123

*Tsang, E.W. (2002), "Learning from overseas venturing experience: the case of Chinese family businesses", *Journal of Business Venturing*, Vol. 17 No. 1, pp. 21-40, available at: https://doi.org/10.1016/S0883-9026(00)00052-5

*Tung, J., Lo, S.C., Chung, T. and Huang, K.P. (2014), "Family business internationalisation: the role of entrepreneurship and generation involvement", *The Anthropologist*, Vol. 17 No. 3, pp. 811-822, doi: 10.1080/09720073.2014.11891495.

*Velez-Ocampo, J., Govindan, K. and Gonzalez-Perez, M.A. (2017), "Internationalization of Mexican family firms: the cases of Xignux and Grupo alfa", *Review of International Business and Strategy*, Vol. 27 No. 2, pp. 180-198, doi: 10.1108/RIBS-10-2016-0058.

*Wach, K. (2015), "Incremental versus rapid internationalisation of firms: results of exploratory investigation from Poland", *Entrepreneurial Business and Economics Review*, Vol. 3 No. 4, pp. 37-48, doi: 10.15678/EBER.2015.030403.

*Wach, K. (2017), "Exploring the role of ownership in international entrepreneurship: how does ownership affect internationalisation of polish firms?", *Entrepreneurial Business and Economics Review*, Vol. 5 No. 4, pp. 205-224, available at: http://doi.org/10.15678/EBER.2017.050410

*Wąsowska, A. (2017), "Internationalisation of family firms: the role of ownership structure and composition of top management team", *Entrepreneurial Business and Economics Review*, Vol. 5 No. 1, pp. 169-185, available at: http://dx.doi.org/10.15678/EBER.2017.050110

*Yamanoi, J. and Asaba, S. (2018), "The impact of family ownership on establishment and ownership modes in foreign direct investment: the moderating role of corruption in host countries", *Global Strategy Journal*, Vol. 8 No. 1, pp. 106-135, available at: https://doi.org/10.1002/gsj.1198

Corresponding author
Apoorva Jain can be contacted at: apoorvaj95@gmail.com

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**

VDDDA Based Multifunction Voltage-Mode MISO Filter

Dhruv Goel[1*], Deva Nand[1]

[1*]*Department of Electronics and Communication Engineering*
Delhi Technological University
New Delhi, India
dg210895@gmail.com[1*]

[1]*Department of Electronics and Communication Engineering*
Delhi Technological University
New Delhi, India
devkamboj07@gmail.com[1]

***Abstract* - This paper proposes a voltage-mode second-order multifunction filter that uses a voltage differencing differential difference amplifier (VDDDA) as an active building block. The proposed filter is sort of MISO multifunction filter having three input nodes and single output node. It comprises of one VDDDA, one resistor and two identical capacitors. In this proposal, the filter offers low-pass (LP), high-pass (HP), band-stop (BS), band-pass (BP), and all-pass (AP) responses for the same circuit configuration by selecting the appropriate input signals. The quality factor and natural frequency both can be electronically controllable and tunable with a dc bias current. Theoretical results have been conformed in SYMICA DE simulations using 0.18µm TSMCS CN018 CMOS process parameters with supply voltages of ±0.9V.**

Index Terms - VDDDA, analog filter, voltage-mode, MISO.

I. INTRODUCTION

Despite the significant development in digital signal processing and its real time applications, one can never replace the analog signal processing due to its particular advantages. Amplification, rectification, modulation, etc., processed on the message signals can only be performed by the analog processing. Various devices that can process the analog electrical signals have been designed and further modifications in circuits are done according to their characteristics. These modified devices could now further be used for applications such as filters, summers, clippers and clampers, amplifiers, oscillators, etc. [1].

The key application of analog signal processing is electronic filter. A filter can be designed by allowing the electrical signal of particular frequency range to pass through it depending upon their type and blocking the unwanted or undesired frequencies, which can be "low pass" , "high pass" , "band pass" , "band reject" and "all pass" filter. "Current–mode", "voltage-mode", "trans-conductance-mode" and "trans-resistance-mode filters" are various available modes of the filter operation [2]. Multifunction filter is the one that can perform two or more types of filtering functions whereas universal filter is the one that can provide all five types of filtering functions. Further according to number of inputs and outputs, filters are classified into four types which are "Single-Input-Single-Output" (SISO), "Single-Input-Multiple-Output" (SIMO), "Multiple-Input-Single-Output" (MISO) or "Multiple-Input-Multiple-Output" (MIMO) type. For the designing of MISO type voltage-mode universal filters in signal processing, designers are giving major importance due to following key features: (i) Realization of all standard five filter functions utilizing same circuit topology on the basis of input ports used, (ii) Minimum use of active and passive components in circuit design, and (iii) Design simplicity and versatility that bring cost reduction to IC manufacturer [3].

Voltage Differencing Differential Difference Amplifier (VDDDA) [4] is one such new versatile 'voltage differencing' active building block (ABB). It is a six-terminal universal block to realize various amplifiers with less transistor count and constitute dynamic properties required for any real-time applications. Device's internal structure consists of two active blocks, one is operational transconductance amplifier (OTA) [5] followed by differential difference amplifier (DDA) [6], [7]. This building block consists of three addition/subtraction input voltage terminals where the transconductance (g_m) of the block is electronically tuneable. In analog signal processing, the motivation behind this work and using VDDDA as an active building block is that it offers various advance features, for example, electronic controllability, avoiding the need of external voltage summing circuit in circuit design [3], quick circuit synthesization with least components count, etc. Especially, in the design of voltage-mode active filter, the voltage summing is required in filter thus; VDDDA is a good choice for designers. Moreover, the output impedance of VDDDA is low which allows the cascadability of devices without the need of voltage buffer. It can provide inverting as well as non-inverting voltage-mode outputs simultaneously from the same circuit topology.

This paper describes a three input and single output multifunction filter. The proposal comprises of single VDDDA, two capacitances and one resistance. All passive components are of floating type. It provides all five filter functions with three input voltage supplies which can be digitally selected. The natural frequency of a filter and its quality factor both can be electronically controllable and tuneable via a dc current source. It is cost efficient to IC manufacturer because of its simple and versatile design.

II. CIRCUIT DESIGN AND PROPOSED FILTER ANALYSIS

A. VDDDA Overview

Voltage Differencing Differential Difference Amplifier (VDDDDA) is a six-terminal universal active block. This building block involves three addition/subtraction input voltage nodes where the transconductance (g_m) of the block is electronically tuneable. The VDDDA circuit symbol and its equivalent circuit is represented in Fig. 1. V_+, V_-, Z, P and N are the five input voltage ports. The Z port is also characterized as the current output port and the W port is the voltage output port. Ideally, the input impedance at V_+, V_-, Z, P and N ports are high and the output impedance at W port is

978-1-7281-9287-1/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 06:20:42 UTC from IEEE Xplore. Restrictions apply.

low. The difference in the input voltages, V_+ and V_- is transmitted as output current I_z via g_m which is tuneable too. The differential input voltages of V_z, V_n and V_p constitutes output voltage V_w.

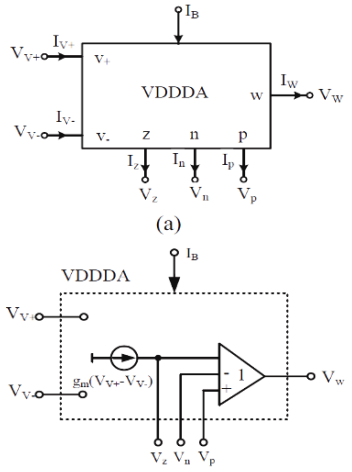


Fig. 1 (a) Symbolic representation of VDDDA, (b) Equivalent representation of VDDDA [4].

The port characteristic equation for an ideal VDDDA can be modelled in form of matrix equation which is as follows:

$$\begin{bmatrix} I_{v+} \\ I_{v-} \\ I_z \\ I_n \\ I_p \\ V_w \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ g_m & -g_m & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 1 & 0 \end{bmatrix} \begin{bmatrix} V_{v+} \\ V_{v-} \\ V_z \\ V_n \\ V_p \\ I_w \end{bmatrix} \quad (1)$$

Following are the equations derived from matrix equation,

$$I_{v+} = I_{v-} = I_n = I_p \quad (2)$$

$$I_z = g_m \left(V_{v+} - V_{v-}\right) \quad (3)$$

$$V_w = V_z - V_n + V_p \quad (4)$$

where in ideal scenario, the currents at ports v_+, v_-, n and p are equal to zero from Eq. (2). In Eq. (3) and (4), g_m and V_w represent the transconductance and the output voltage of ideal VDDDA respectively.

B. Proposed MISO Filter Description

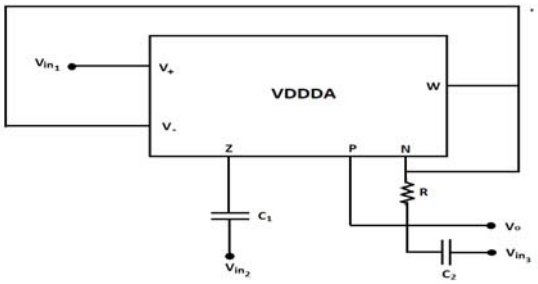


Fig. 2 VDDDA based voltage-mode MISO filter

The proposal of VM MISO filter is presented in Fig. 2. The filter comprises of single VDDDA, two capacitors and one floating resistor. In this proposal, the filter incorporates the least active and passive component count, which is a desirable feature for integration. It exhibits high input impedance for LP terminal. In future study, circuit can further be modified by replacing the resistor with CMOS implemented electronic resistor. In this filter design, there are three input voltage signals, namely V_{in1}, V_{in2}, V_{in3} and V_0 is the only output voltage. Assuming case of an ideal VDDDA block, mathematical analysis of the MISO filter offers the output voltage, which is given as,

$$V_0 = \frac{s^2 V_{in3} + s\frac{(C_1 V_{in2} + C_2 R g_m V_{in3})}{2C_1C_2R} + \frac{g_m V_{in1}}{2C_1C_2R}}{D(s)} \quad (5)$$

where,

$$D(s) = s^2 + s\frac{(C_1 + C_2 R g_m)}{2C_1C_2R} + \frac{g_m}{2C_1C_2R} \quad (6)$$

Eq. (5) concludes that the proposed MISO filter has unity gain and if we want to achieve desired gain from the active filter, designers need to add voltage amplifies to get additional gain. From Eq. (5), we can get all standard five output filter responses by selecting the input voltage terminals (V_{in}) accordingly and selection can be done digitally through below conditions,

- The second order LP response can be achieved if $V_{in1} = V_{in}$ and $V_{in2} = V_{in3} = 0$.
- The second order HP response can be achieved if $-V_{in2} = V_{in3} = V_{in}$ and $V_{in1} = 0$.
- The second order BP response can be achieved if $V_{in2} = V_{in}$ and $V_{in1} = V_{in3} = 0$.
- The second order BS response can be achieved if $V_{in1} = -V_{in2} = V_{in3} = V_{in}$.
- The second order AP response can be achieved if $V_{in1} = \frac{-V_{in2}}{3} = V_{in3} = V_{in}$.

For hardware implementation of HP and BS filter response, there is need of inverting unity gain amplifier and for AP filter implementation, inverting triple gain amplifier is required. From the characteristic equation i.e., Eq. (6), we can find the pole frequency and quality factor given by:

$$\omega_0 = \sqrt{\frac{g_m}{2C_1C_2R}} \quad (7)$$

$$Q_0 = \frac{\sqrt{2C_1C_2Rg_m}}{C_1 + C_2Rg_m} \quad (8)$$

Note from Eqs (7) and (8), the ω_0 and the Q_0 both can be electronically tuned through transconductance, g_m. However, it turns out that the proposed MISO filter requires a matching condition for its components ($C_1 = C_2Rg_m$) to achieve the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 06:20:42 UTC from IEEE Xplore. Restrictions apply.

HP, BS and AP functions. The relative sensitivities of ω_0 and Q_0 for the proposed filter may be found as,

$$S_{g_m}^{\omega_0} = \frac{1}{2}; \quad S_{C_1}^{\omega_0} = S_{C_2}^{\omega_0} = S_R^{\omega_0} = -\frac{1}{2}; \tag{9}$$

$$S_{C_1}^{Q_0} = \frac{1}{2} - \frac{C_1}{C_1 + C_2 R g_m}; \; S_{g_m}^{Q_0} = S_{C_2}^{Q_0} = S_R^{Q_0} = \frac{1}{2} - \frac{C_2 R g_m}{C_1 + C_2 R g_m}$$

It is observed that the magnitude of relative sensitivities is either equal to one or less than one.

III. Analysis of Non-Ideal Behavior

In real-time scenarios, the overall performance and parameters of the MISO filter are altered by the influence of an error in voltage tracking of VDDDA block. In this section, we will study and analyse these non-ideal characteristics of VDDDA which can be represented as following:

$$V_w = \beta_z V_z - \beta_n V_n + \beta_p V_p \tag{10}$$

"where in Eq. (10), β_z, β_n and β_p represent the non-ideal voltage gain (or, voltage tracking error) of VDDDA from Z, N and P ports to W port, respectively and it is less than one in magnitude. For ideal case, is taken as unity gain." [4] The voltage filter response for the non-ideal case is represented by Eq (11):

$$V_0^* = \frac{s^2 V_{in3} + \frac{s\beta_z(C_1 V_{in2} + C_2 R g_m V_{in3})}{C_1 C_2 R(1+\beta_n)} + \frac{g_m \beta_z V_{in1}}{C_1 C_2 R(1+\beta_n)}}{D(s)} \tag{11}$$

where,

$$D^*(s) = s^2 + s\left(\frac{1}{C_2 R} + \frac{C_2 R g_m \beta_z - C_1 \beta_p}{C_1 C_2 R(1+\beta_n)}\right) + \frac{g_m \beta_z}{C_1 C_2 R(1+\beta_n)} \tag{12}$$

Note from Eq. (12) that the characteristic equation parameters (natural frequency and quality factor) are also affected by the voltage tracking error. Following are non-ideal estimations of $\omega_0$and Q_0:

$$\omega_0^* = \sqrt{\frac{g_m \beta_z}{C_1 C_2 R(1+\beta_n)}} \tag{13}$$

$$Q_0^* = \frac{1}{1 + \frac{C_2 R g_m \beta_z}{C_1(1+\beta_n)} - \frac{\beta_p}{(1+\beta_n)}} \sqrt{\frac{C_2 R g_m \beta_z}{C_1(1+\beta_n)}} \tag{14}$$

IV. Simulation Results

The functionality of CMOS implemented multifunction MISO filter is verified through SYMICA DE simulations with model library of 0.18 µm TSMC CMOS parameter file [8] at ±0.9V supply voltages. The CMOS structure of VDDDA is given in Fig. 3 and the transistor's sizes are illustrated in Table 1. Following are the values of the components used in proposed filter: $C_1 = C_2 = 30$ pF, R = 3.3 kΩ and the bias current I_B is set to 50 µA. Figs. (4-6) provide the gain functions of LP, HP and BP functions and Fig. 7 and 8 provide the gain and phase plots of BS and AP functions of the proposed MISO filter respectively. Due to the voltage tracking error or the non-ideal behavior of VDDDA, there is a deviation in the simulated natural frequency and the theoretical one. The simulated ω_0 = 1.106 MHz and the theoretical pole frequency from Eq. (7) is 1.131 MHz with the same value of passive components and bias current. Thus, it gives a deviation of 2.21% in natural frequency due to non-ideal behavior of VDDDA. The quality factor (Q_0) of filter is 0.707. The gain responses of BP function for different I_B are shown in Fig. 9, where $C_1 = C_2 = 30$ pF and I_B is equal to 50 µA, 100 µA & 150 µA and the corresponding values of ω_0 are 1.106 MHz, 1.272 MHz and 1.355 MHz respectively. Therefore, it is inferenced from Eqs. (7) and (8) that both ω_0 and Q_0 can be tuned electronically via g_m by changing I_B. Fig. 10 depicts the transient response for BP function. In this experiment, the proposed filter is fed with the input sinusoidal signal of 180 mV amplitude and frequency of 1 MHz.

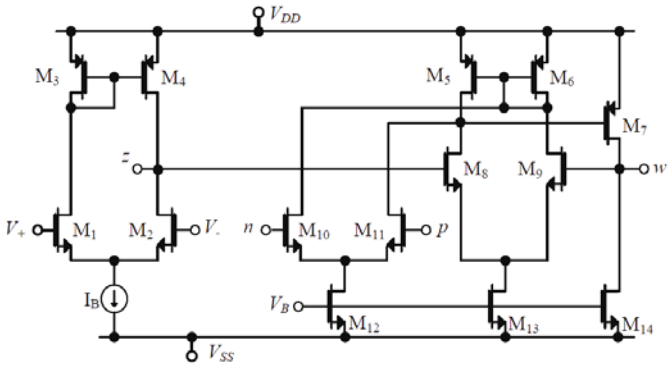


Fig. 3 VDDDA implemented by CMOS technology [4]

TABLE I
Aspect Ratios of Transistors [4]

MOS transistor	W/L (µm/µm)
M_1, M_2	9/1.08
M_3, M_4	3.96/1.08
M_5 - M_7	3.6/0.18
M_8 - M_{11}	0.72/1.08
M_{12} - M_{14}	2.16/1.08

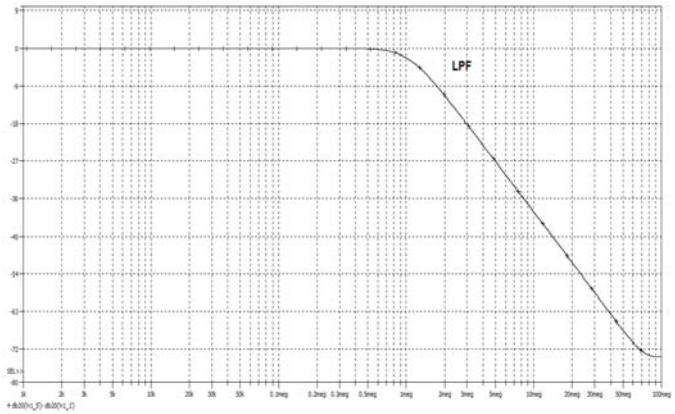


Fig. 4 Gain plot for LP

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 06:20:42 UTC from IEEE Xplore. Restrictions apply.

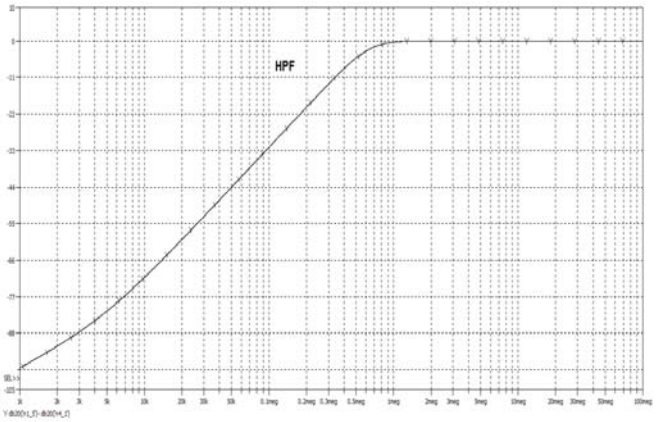


Fig. 5 Gain plot for HP

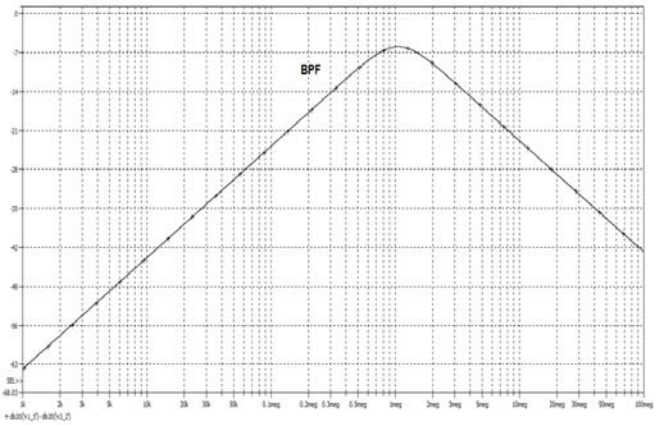


Fig. 6 Gain plot for BP

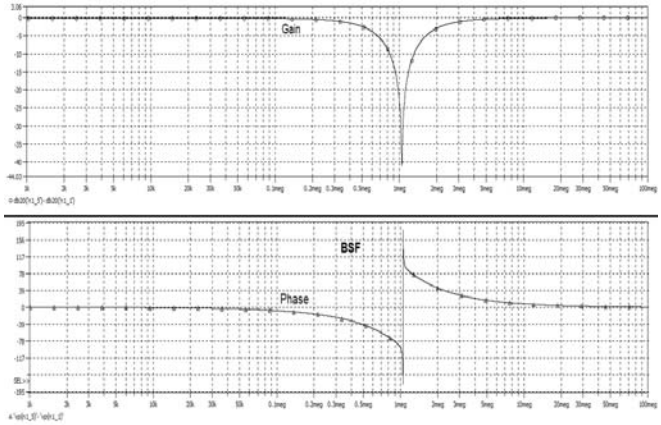


Fig. 7 Gain and phase plot for BS

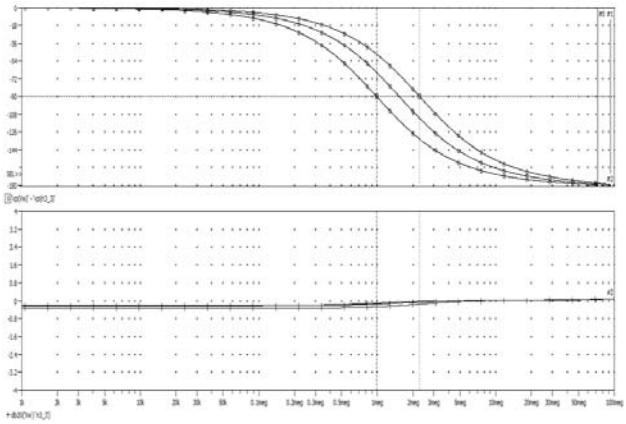


Fig. 8 Phase and gain plot for AP

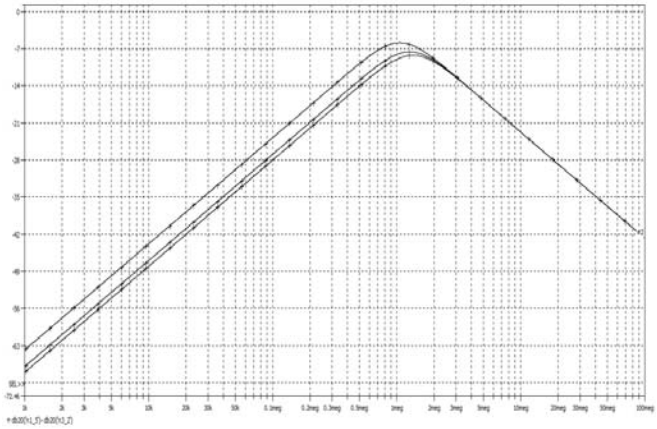


Fig. 9 BP plots for different values of I_B

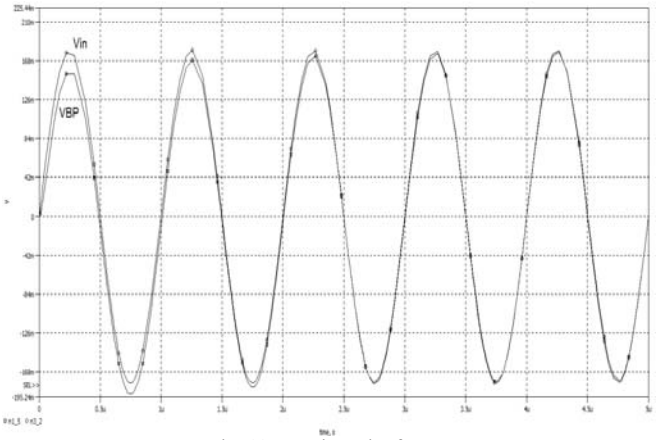


Fig. 10 Transient plot for BP

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 06:20:42 UTC from IEEE Xplore. Restrictions apply.

V. Comparative Study

Many literature surveys have been conducted for voltage-mode multifunction filters using different active building blocks. A comparison has been done between the proposed MISO filter and the other referenced VM MISO filter as shown in Table 2. This table encapsulates the pros and cons of the previously designed filters with the proposed filter. From the table, for a second-order filtering function; we require at least two capacitors for quadratic equation. Filters in [9-11, 13, 15, 17, 18, 21-23] consist of floating passive elements in their circuit. Electronic tuning for the ω_0 and Q_0 can be achieved with filters [11, 13, 16-18, 20, 22-24]. All standard five filter responses can be realized in all surveyed filters. Those circuits that can't provide high input impedance are [10, 12, 13, 15, 17, 18, 20-23]. The filters that can provide low output impedance are [12, 14, 15, 21, and 24]. The mentioned voltage-mode MISO filters in [9, 10, 15, 18-20] require matching condition between passive elements. Also, the circuits in [9-11, 13, 18, and 23] need inverting input signals for filter responses.

TABLE II
Comparison of Different VM MISO Filters

Refs	ABB	No. of ABB	No. of R & C	Grounded passive elements	Electronic Tuning	Matching condition	High input impedance	Low output impedance	Inverting input
9	CCII	3	2+2	No	No	Yes	Yes	No	Yes
10	CFA	1	3+2	No	No	Yes	No	No	Yes
11	OTA & CCII	3	0+2	No	Yes	No	Yes	No	Yes
12	DDCC	2	2+2	Yes	No	No	No	Yes	No
13	OTA	2	0+2	No	Yes	No	No	No	Yes
14	DDCC	3	2+2	Yes	No	No	Yes	Yes	No
15	OTRA	1	4+4	No	No	Yes	No	Yes	No
16	OTA	6	0+2	Yes	Yes	No	Yes	No	No
17	CCTA	1	2+2	No	Yes	No	No	No	No
18	DVCC	1	2+2	No	Yes	Yes	No	No	Yes
19	DVCC	3	4+2	Yes	No	Yes	Yes	No	No
20	VDTA	1	0+2	Yes	Yes	Yes	No	No	No
21	CDBA	2	4+2	No	No	No	No	Yes	No
22	VD-DIBA	1	1+2	No	Yes	No	No	No	No
23	VDDDA	1	2+2	No	Yes	No	No	No	Yes
24	VDDDA	2	0+2	Yes	Yes	No	Yes	Yes	No
This work	VDDDA	1	1+2	No	Yes	Yes	LP	No	Yes

VI. Conclusions

In this proposal, voltage-mode biquad MISO filter has been proposed. The MISO filter proposes various advantages which are: First, the proposed filter can provide all five standard functions: Low-Pass (LP), High-Pass (HP), Band-Pass (BP), Band-Stop (BS) and All-Pass (AP) responses with digitally terminal switching of input ports, Secondly, the natural frequency and the quality factor both can be electronically controllable and tuneable. Thirdly, it consists of single VDDDA, two capacitors and one resistor which means it has least components count. Finally, it has high input impedance for LP response. Simulation results verified the theoretical concepts and validity of synthesis. However, in future with slight modification in the circuit, proposed filter can further be used in the implementation of voltage-mode quadrature oscillator, VDDDA based shadow filter and many other applications.

References

[1] A. S. Sedra and K. C. Smith, “Microelectronic circuits”, Florida: Holt, Rinehart and Winston, 1991.

[2] R. Raut and M. N. S. Swamy, “Modern Analog Filter Analysis and Design: A Practical Approach”, Copyright 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.

[3] M. E. Van Valkenburg, “Analog Filter Design”, Holt, Rinehart & Winston, 1982.

[4] N. Herensar, R. Sotner, B. Metin, J. Koton and K. Vrba, “VDDDA - New ‘voltage differencing’ device for analog signal processing,” Proceeding of the 8th International Conference on Electrical and Electronics Engineering, pp. 17-20, 2013.

[5] R. L. Geiger and E. Sanchez-Sinencio, “Active filter design using operational transconductance amplifiers: A tutorial”, IEEE Circuits Devices Mag., vol. 1, pp. 20–32, 1985.

[6] E. Sackinger and W. Guggenbiihl, “A versatile building block: CMOS differential difference amplifier”, IEEE J. Solid-State Circuits, vol. 22, pp. 287–294, 1987.

[7] W. Chiu, S.-I. Liu, H.-W. Tsao, and J.-J. Chen, “CMOS differential difference current conveyors and their applications”, IEE Proc. Circuits Devices Syst., vol. 143, pp. 91–96, 1996.

[8] TSMC 0.18 μm level-7 SCN018 CMOS technology parameters [online]: ftp://ftp.isi.edu/pub/mosis/ vendors/tsmc-018/t44e lo epi-params.txt.

[9] J. W. Horng. High-input impedance voltage-mode universal biquadratic filter using three plus-type CCIIs. IEEE Trans Circuits Syst II 2001; 48(10):996–7.

[10] J. W. Horng, C. K. Chang, J. M. Chu. Voltage-mode universal biquadratic filter using single current-feedback amplifier. IEICE Trans Fundam 2002; 85:1970–3.

[11] J. W. Horng. High input impedance voltage-mode universal biquadratic filter using two OTAs and one CCII. Int J Electron 2003; 90(3):153–7.

[12] C. M. Chang, H. P. Chen. Universal capacitor-grounded voltage-mode filter with three inputs and a single output. Int J Electron 2003; 90(6):401–6.

[13] J. W. Horng. Voltage-mode universal biquadratic filter using two OTAs and two capacitors. IEICE Trans Fundam 2003; 86A (2):411–3.

[14] J. W. Horng. High-input and low-output impedance voltage-mode universal biquadratic filter using DDCCs. IEEE Trans Circuits Syst II 2007; 54(8):649–52.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 06:20:42 UTC from IEEE Xplore. Restrictions apply.

[15]S. Kılınc, A. Ü. Keskin, U. Cam. Cascadable voltage-mode multifunction biquad employing single OTRA. Frequenz 2007; 61:84–6.
[16]M. Kumngern, M. Somdunyakanok, P. Prommee. High-input impedance voltage-mode multifunction filter with three-Input single-output based on simple CMOS OTAs. Int Symp Commun Inf Technol 2008:426–31.
[17]N. Herencsar, J. Koton, K. Vrba. Single CCTA-based universal biquadratic filters employing minimum components. Int J Comput Electr Eng 2009;1(3):309–12.
[18]W. Tangsrirat, O. Channumsin. Voltage-mode multifunctional biquadratic filter using single DVCC and minimum number of passive elements. Indian J Pure Appl Phys 2011; 49:703–7.
[19]J. W. Horng, C. H. Hsu, C. Y. Tseng. High input impedance voltage-mode universal biquadratic filters with three inputs using three CCs and grounding capacitors. Radioengineering 2012; 21(1):290–6.
[20]J. Satansup, W. Tangsrirat. Single VDTA-based voltage-mode electronically tunable universal filter. In: The 27th International Technical Conference on Circuits Systems, Computers and Communications. 2012.
[21]J. K. Pathak, A. K. Singh, R. Senani. New voltage mode universal filters using only two CDBAs. ISRN Electron 2013 [Article ID 987867].
[22]K. L. Pushkar, D. R. Bhaskar, D. Prasad. A new MISO-type voltage-mode universal biquad using single VD-DIBA. ISRN Electron 2013 [Article ID 478213].
[23]A. Chaichana, S. Sangyaem, W. Jaikla. Multifunction voltage-mode filter using single voltage differencing differential difference amplifier. In: 3rd international conference on mechatronics and mechanical engineering; 2016.
[24]S. Sangyaem, S. Siripongdee, W. Jaikla, F. Khateb. Five-inputs single-output voltage mode universal filter with high input and low output impedance using VDDDAs. Optik 2017; 128:14–25.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 06:20:42 UTC from IEEE Xplore. Restrictions apply.

2020 IEEE International Conference for Innovation in Technology (INOCON)
Bengaluru, India. Nov 6-8, 2020

VDTA based Fractional Order Universal Filter

Jyoti Srivastava
Department of Electrical Engineering
Delhi Technological University
Delhi, India
jyotisrivastava_mt2k18@dtu.ac.in

***Abstract*—In this paper, electronic tunability of current mode multiple input single output (MISO) type modified biquad fractional order universal filter employing single voltage differencing transconductance amplifier (VDTA) and two grounded fractional capacitors is discussed. The tunability of all the filters, namely, fractional high pass filter (FHPF), fractional low pass filter (FLPF), fractional band pass filter (FBPF), fractional band reject filter (FBRF), and fractional all pass filter (FAPF) with respect to α and I_{bias} is shown. The workability of the modified filter is verified using PSPICE simulations.**

Keywords—Operational Transconductance Amplifier (OTA), multi input single output (MISO), Voltage Differencing Transconductance Amplifier (VDTA), Universal filter Differential Voltage Controlled Current Source (DVCCS)

I. INTRODUCTION

Analog signal processing applications frequently employ active filters as important basic building block. Ever since the introduction of the VDTA [1], it has been extensively utilized for realization of filter circuits and sinusoidal oscillators [2–5]. In like manner, use of fractional order circuits are gaining more attention, in research areas of biomedical, control systems, and instrumentation, as it explains dynamics of natural system more closely. Analog filter design using fractional order elements gives a better control of attenuation gradient in the pass/stop band. Several fractional order filter circuits utilizing different active building blocks (ABBs) [4-15] have been presented recently. Many of these circuits have generalized an existing integer order filter circuit into fractional order domain. A careful perusal of the existing literature on fractional order filters reveals that very few circuits of fractional order filters utilizing VDTAs have been presented [4-5].

In this paper, we have generalized the modified single VDTA based three input and a single output current-mode electronically tuneable universal biquadratic filter [2] in fractional order domain and examined the electronic tunability of the various parameters of the filter. The pole frequency (ω_0) of the modified circuit can be electronically tuned by using two parameters α and I_{bias} independently. Also, the stability of the presented circuit has been analysed.

II. THE VOLTAGE DIFFERENCING TRANSCONDUCTANCE AMPLIFIER (VDTA)

The VDTA is an adaptable ABB as shown in Fig. 1.

Essentially, VDTA device comprises of an input voltage subtractor where input voltage difference ($v_p - v_n$) is transferred through the z-terminal in form of current (i_z) by the first transconductance gain (g_{m1}) and the dual output transconductance amplifiers that are responsible for converting the voltage at z-terminal to the currents at x-terminals by transconductance gain (g_{m2}). Also, zc terminal gives the copy of current i_z as i_{zc}. Evaluating the fundamental operation of an ideal VDTA, the following matrix depicts its terminal relationship:

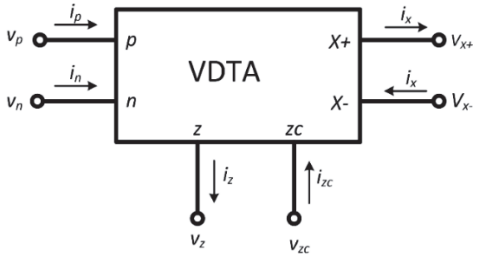


Fig. 1. VDTA structure

$$\begin{bmatrix} i_p \\ i_n \\ i_z \\ i_x \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ g_{m1} & -g_{m1} & 0 & 0 \\ 0 & 0 & 0 & g_{m2} \end{bmatrix} \begin{bmatrix} v_p \\ v_n \\ v_{zc} \\ v_z \end{bmatrix} \quad (1)$$

where g_{m1} and g_{m2} are the electronically controllable transconductances gain of the VDTA using OTAs. Its relation with external bias current (I_{bias}) is given in equation (2).

$$g_m = \frac{I_{bias}}{2V_T} \quad (2)$$

Since the VDTA is yet not available, there are several possibilities of its implementation, implementing it through ICs is commercially more feasible. In this paper VDTA is implemented using three OTA ICs LM13700 at two transconductance gain stages and six input output terminals as shown in Fig 2.

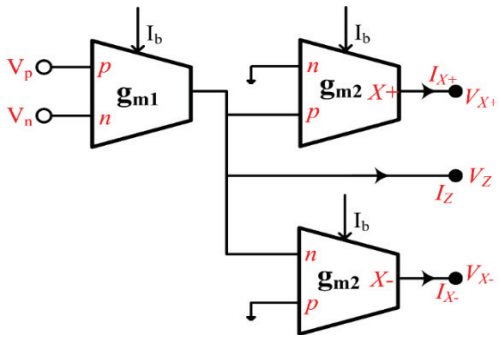


Fig. 2. VDTA using LM13700 IC

III. THE FRACTIONAL ORDER VDTA FILTER

The generalized MISO fractional order filter circuit comprising of single VDTA and two FCs is shown in Fig 3. The use of FCs [16-17] provides the flexibility of tuning through two parameters namely, transconductance gain (g_m) and α making it easier to be tuned and controlled.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:12:32 UTC from IEEE Xplore. Restrictions apply.

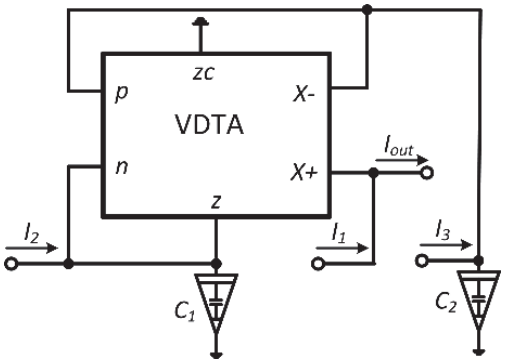


Fig. 3. The modified fractional order current mode filter

Equations (3) and (4) gives the transfer function of the fractional order filter responses.

$$I_{out} = \frac{D(s)I_1 + (g_{m2}/C_1)s^{\alpha} I_2 + (g_{m1}g_{m2}/C_1C_2)I_3}{D(s)} \tag{3}$$

$$D(s) = s^{2\alpha} + (g_{m1}/C_1)s^{\alpha} + (g_{m1}g_{m2}/C_1C_2) \tag{4}$$

Conditions for realization of FLPF, FHPF, FBPF, FBSF, and FAPF are shown in Table 1.

TABLE I. CONDITIONS FOR UNIVERSAL FILTER RESPONSE

Filters	Cases
FLP	$I_3 = I_{in}, I_1 = I_2 = 0$
FHP	$I_1 = -I_2 = -I_3 = I_{in}, g_{m1} = g_{m2}$
FBP	$I_2 = I_{in}, I_1 = I_3 = 0$
FBS	$I_1 = -I_2 = I_{in}, I_3 = 0, g_{m1} = g_{m2}$
FAP	$I_1 = -I_2/2 = I_{in}, I_3 = 0, g_{m1} = g_{m2}$

For any fractional order filter with a transfer function T(s), following important parameters should be determined [11]: ω_m (maxima or minima frequency), ω_h (pole cutoff frequency/half power frequency), and ω_{rp} (right phase frequency). The bandwidth of any filter can also be calculated using ω_h.

The equations to find values of ω_m, ω_h, and ω_{rp} for different filters can be calculated from equations (5)-(17).

A. FLPF

$$\frac{I_{out}}{I_{in}} = \frac{(g_{m1}g_{m2}/C_1C_2)}{D(s)} \tag{5}$$

For ω_m

$$2X^3 + 3\cos\left(\frac{\alpha\pi}{2}\right)X^2 + 2[1 + k\cos(\alpha\pi)]X + k\cos\left(\frac{\alpha\pi}{2}\right) = 0 \tag{6}$$

For ω_h

$$Y^4 + 2\cos\left(\frac{\alpha\pi}{2}\right)Y^3 + (1 + 2k\cos(\alpha\pi)Y^2 + 2k\cos\left(\frac{\alpha\pi}{2}\right)Y - k^2 = 0 \tag{7}$$

For ω_{rp}

$$\omega_{rp} = \left(\frac{-a\cos\left(\frac{\alpha\pi}{2}\right) - \sqrt{a^2\cos^2\left(\frac{\alpha\pi}{2}\right) - 4ab\cos(\alpha\pi)}}{2\cos(\alpha\pi)}\right)^{\frac{1}{\alpha}} \tag{8}$$

B. FHPF

$$\frac{I_{out}}{I_{in}} = \frac{s^{2\alpha}}{D(s)} \tag{9}$$

For ω_m, ω_h and ω_{rp}

$$\begin{aligned} &\omega_{mFLPF} \cdot \omega_{mFHPF} = \omega_{hFLPF} \cdot \omega_{hFHPF} \\ &= \omega_{rpFLPF}.\omega_{rpFHPF} = (ab)^{\frac{1}{\alpha}} \end{aligned} \tag{10}$$

C. FBPF

$$\frac{I_{out}}{I_{in}} = \frac{(g_{m2}/C_1)s^{\alpha}}{D(s)} \tag{11}$$

For ω_m

$$\left(X^2 - k\right)\left(X^2 + \cos\left(\frac{\alpha\pi}{2}\right)X + k\right) \tag{12}$$

For ω_h

$$\begin{aligned} &Y^4 + 2\cos\left(\frac{\alpha\pi}{2}\right)Y^3 - Y^2 + 2k\cos(\alpha\pi)Y^2 \\ &-\left(8k\cos^2\left(\frac{\alpha\pi}{2}\right) + 8\sqrt{k}\cos\left(\frac{\alpha\pi}{2}\right)\right)Y^2 + 2k\cos(\alpha\pi)Y \\ &+ k^2 = 0 \end{aligned} \tag{13}$$

For ω_{rp}

$$\omega_{rp} = \left(\frac{-a \pm \sqrt{a^2 - 4ab\cos^2\left(\frac{\alpha\pi}{2}\right)}}{2\cos\left(\frac{\alpha\pi}{2}\right)}\right)^{\frac{1}{\alpha}} \tag{14}$$

D. FBSF

$$\frac{I_{out}}{I_{in}} = \frac{s^{2\alpha} + (g_{m1}g_{m2}/C_1C_2)}{D(s)} \tag{15}$$

$$\left(X^2 - k\right)\left(\begin{aligned} &X^4\cos\left(\frac{\alpha\pi}{2}\right) + X^3 + \left[\begin{aligned} &4k\cos\left(\frac{\alpha\pi}{2}\right) \\ &-2k\cos(\alpha\pi)\cdot\cos\left(\frac{\alpha\pi}{2}\right)\end{aligned}\right]X^2 \\ &+kX + k^2\cos\left(\frac{\alpha\pi}{2}\right)\end{aligned}\right) \tag{16}$$

$$Y^4 - 2\cos\left(\frac{\alpha\pi}{2}\right)Y^3 - (1 - 2k\cos(\alpha\pi))Y^2 - 2k\cos\left(\frac{\alpha\pi}{2}\right)Y + k^2 = 0 \tag{17}$$

where $X = \frac{\omega_m^{\alpha}}{a}, Y = \frac{\omega_h^{\alpha}}{a}, k = \frac{b}{a}, a = \frac{g_{m1}}{C_1}$ and $b = \frac{g_{m2}}{C_2}$

E. FAPF

$$\frac{I_{out}}{I_{in}} = \frac{s^{2\alpha} - (g_{m1}/C_1)s^{\alpha} + (g_{m1}g_{m2}/C_1C_2)}{D(s)} \tag{18}$$

By putting the value of $\alpha = 1$, all the important frequency parameters for different filters namely, FLPF, FHPF, FBPF, and FBRF are shown in Table II. For the values of α other than 1, these important parameters of different filter can be obtained using equation (5)-(17).

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:12:32 UTC from IEEE Xplore. Restrictions apply.

TABLE II. FILTER PARAMETERS WHEN A=1

Types of filters	ω_m	ω_h	ω_{rp}
FLPF	$a\sqrt{(k-1)}$	$\left(a\sqrt{\left(k\pm\frac{\sqrt{8k^2-4k+1}}{2}-\frac{1}{2}\right)}\right)$	$\sqrt{ab}$
FHPF	$\frac{b}{\sqrt{(k-1)}}$	$\frac{ab}{\left(a\sqrt{\left(k\pm\frac{\sqrt{8k^2-4k+1}}{2}-\frac{1}{2}\right)}\right)}$	$\sqrt{ab}$
FBPF	$a\sqrt{k}$	$a\left(\frac{\sqrt{4k+1}}{2}\pm\frac{1}{2}\right)$	$0,\infty$
FBRF	$a\sqrt{k}$	$a\left(\frac{\sqrt{4k+1}}{2}\pm\frac{1}{2}\right)$	$\sqrt{ab}$

IV. STABILITY ANALYSIS

For any fractional order system, the stability can be obtained by mapping s-plane into W-planes [18]. The stability graph is plotted by keeping the coefficients of $s^{2\alpha}$ positive in equation (4) where value of α lies between 0 and 1. Table III shows the different conditions for stability.

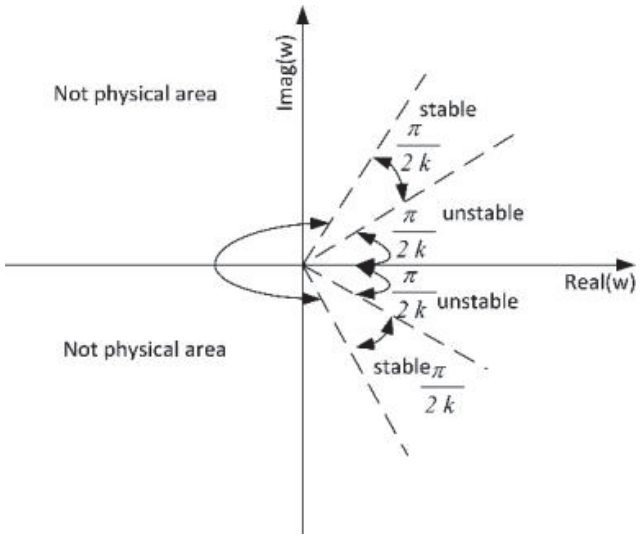


Fig. 4. W-plane (region of stability)

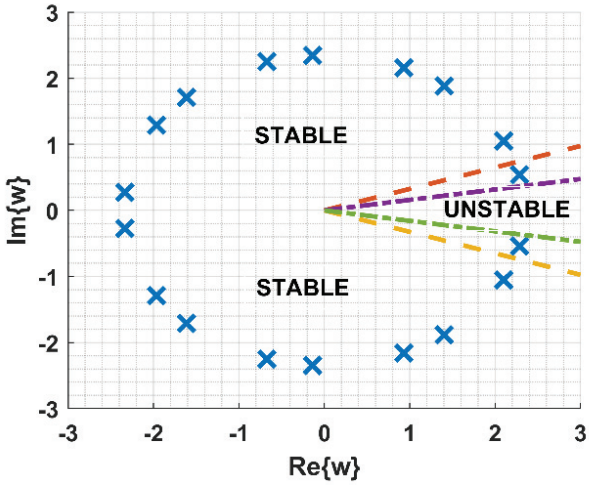


Fig. 5. Stability plot of FLPF C=0.382μF and α=0.9

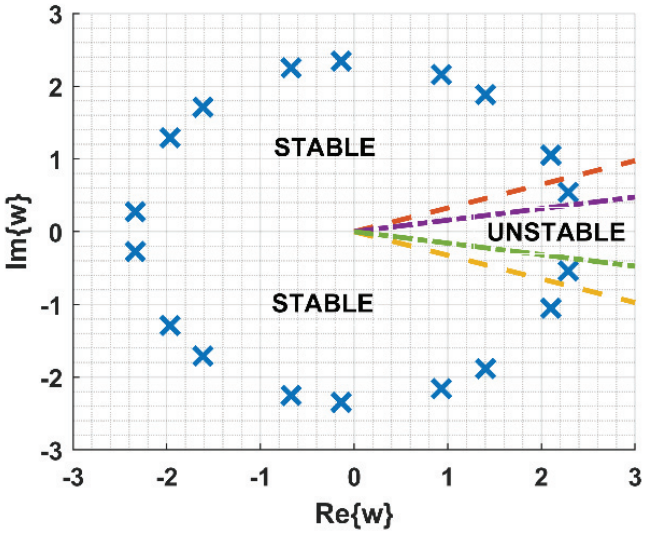


Fig. 6. Stability plot of FHPF C=0.382μF and α=0.9

TABLE III. DIFFERENT CASES OF STABILITY, ROOTS, ω_0 AND Q

Cases	Relations	Condition for stability and roots	ω_0 , Q
1	$a \geq 4b$	$\alpha < 2$, $r_{1,2} = \frac{-a \pm \sqrt{(a)^2 - 4(ab)}}{2} = g_{1,2}e^{j\pi}$	$\omega_{0_{1,2}} = g_{1,2}^{1/\alpha}$, $Q = \frac{-1}{2\cos(\pi/\alpha)}$
2	$a < 4b$	$\alpha < \frac{2\delta}{\pi}$, $\delta = \cos^{-1}\left(\frac{-a}{2\sqrt{ab}}\right) > \frac{\pi}{2}$, $r_{1,2} = \sqrt{ab}e^{\pm j\delta}$	$\omega_0 = \left(\sqrt{ab}\right)^{1/\alpha}$, $Q = \frac{-1}{2\cos(\delta/\alpha)}$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:12:32 UTC from IEEE Xplore. Restrictions apply.

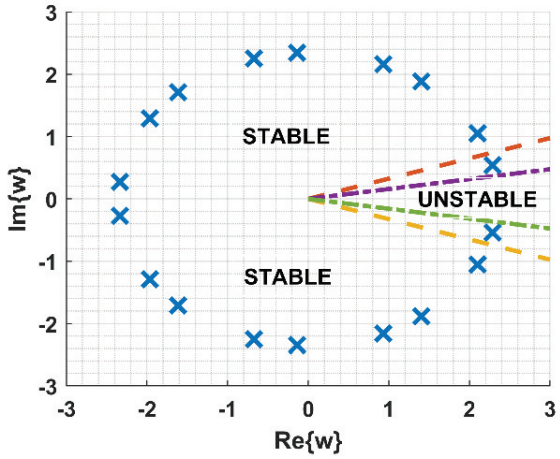


Fig. 7. Stability plot of FBPF C=0.382µF and α=0.9

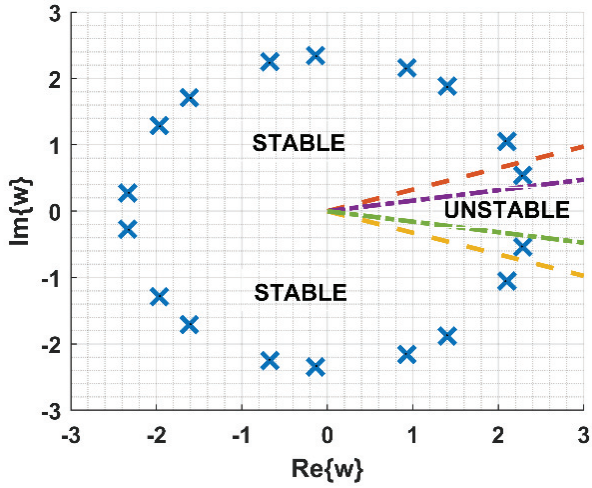


Fig. 8. Stability plot of FBSF C=0.382µF and α=0.9

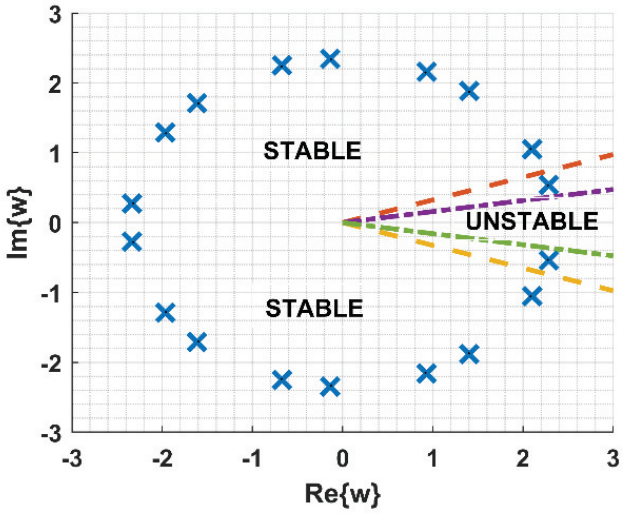


Fig. 9. Stability plot of FAPF C=0.382µF and α=0.9

V. Simulation Results

The workability of the FO filters has been verified through PSPICE simulation using the macro model of OTA IC LM 13700.

A. Tunability with α:

It can be achieved by replacing each ordinary capacitors with equal valued FCs ($C_1 = C_2 = 0.382\ \mu F/(rad/sec)^{(1-\alpha)}$) designed using Oustaloup method [16] for α=0.6, α=0.7, α=0.8, α=0.9 in the frequency range of 0.1 Hz to 100 MHz. All the filters designed in this section have values $I_{bias1} = I_{bias2} = 0.1079mA$ (R_{bias}=265 KΩ).

The different frequency parameters for filters for α=0.6 to α=1.0 have been listed in Table (IV)-(V) and their simulation results are shown in Fig (10)-(14). The simulation values and theoretical values differ in a very small range of error (less than 5%).

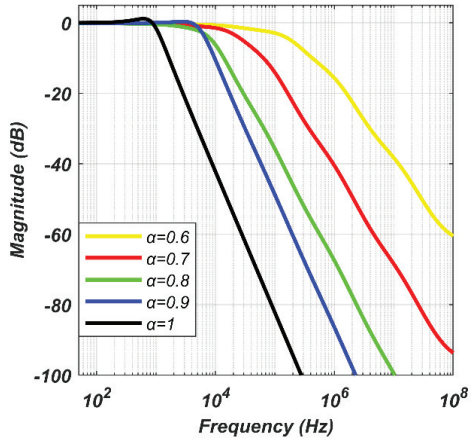


Fig. 10. Tunability of FLP with respect to α

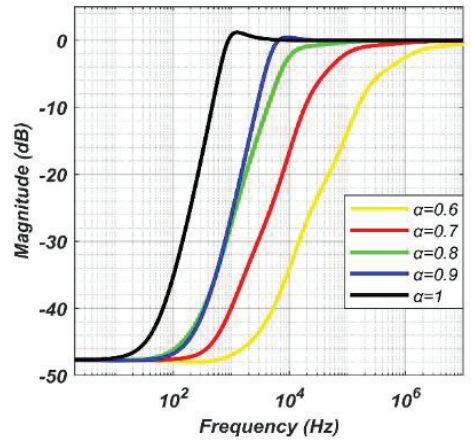


Fig. 11. Tunability of FHP with respect to α

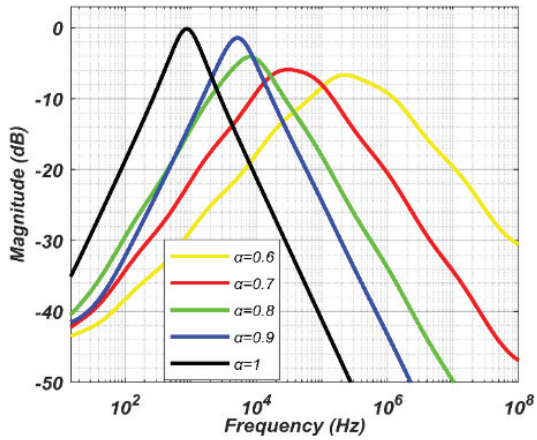


Fig. 12. Tunability of FBP with respect to α

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:12:32 UTC from IEEE Xplore. Restrictions apply.

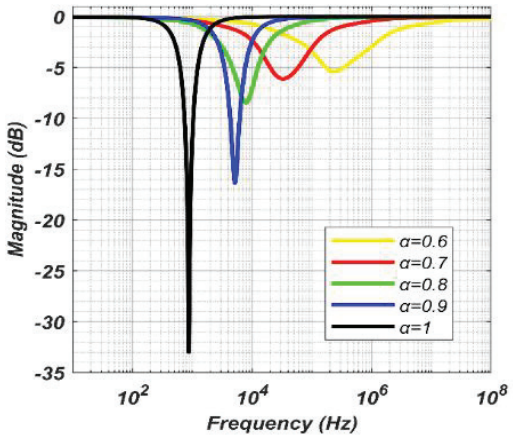


Fig. 13. Tunability of FBS with respect to α

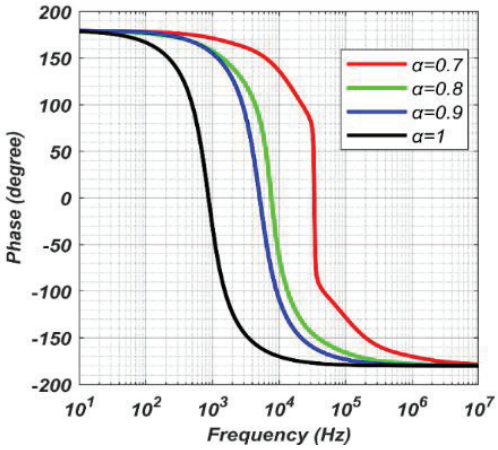


Fig. 14. Tunability of FAP with respect to α

A. Tunability with I_{bias}

Tunability of I_{bias} is obtained by varying the external bias current of OTAs, hereby keeping the value of $I_{bias1}$and I_{bias2} same for all the conditions $I_{bias1} = I_{bias2} = 0.1079\ mA, 0.40857\ mA, 1.1\ mA, and\ 1.9\ mA$ with the values of the bias resistors R_{bias} as 265 KΩ, 70 KΩ, 25 KΩ and 15 KΩ respectively. All the important frequency parameters for filters have been listed in Table (VI)-(VII) and their simulation results is shown in Fig (15)-(19). Again, the error between theoretical and simulated values are very less and acceptable (less than 5%).

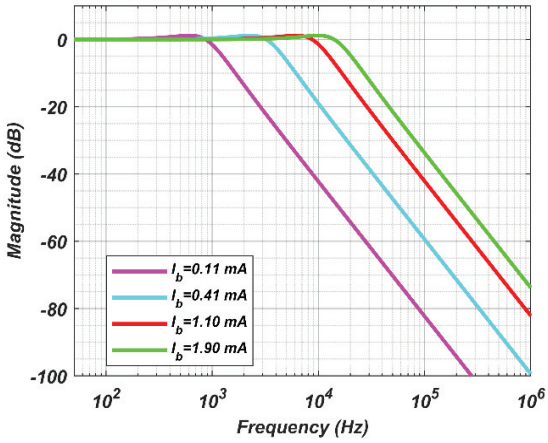


Fig. 15. Tunability of FLP with respect to I_{bias}

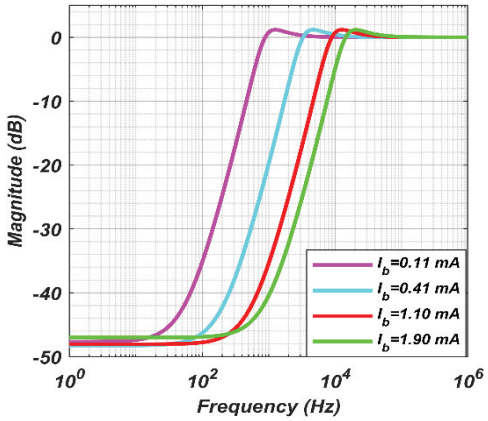


Fig. 16. *Tunability of FHP with respect to I_{bias}*

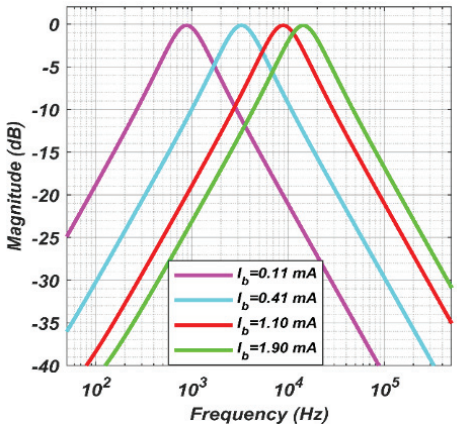


Fig. 17. *Tunability of FBP with respect to I_{bias}*

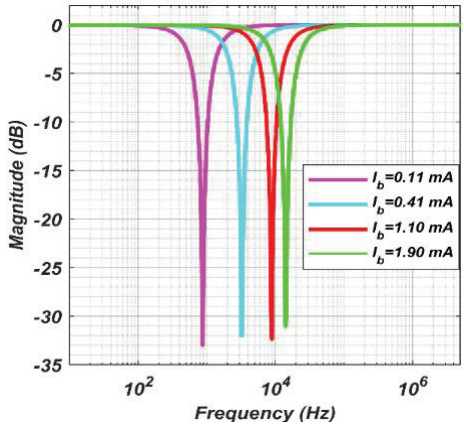


Fig. 18. *Tunability of FBS with respect to I_{bias}*

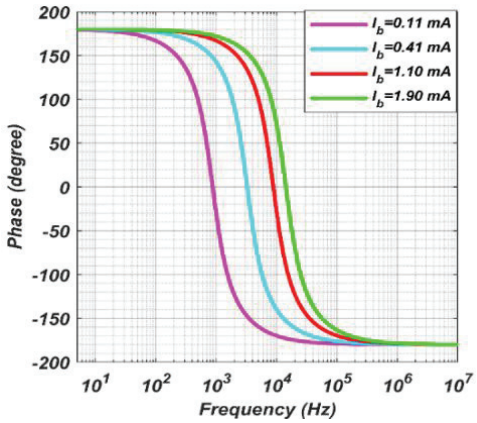


Fig. 19. *Tunability of FLP with respect to Ibias*

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:12:32 UTC from IEEE Xplore. Restrictions apply.

TABLE IV. CUT-OFF FREQUENCIES (F_H) (IN KHZ) OF FRACTIONAL FILTERS FOR DIFFERENT A

	1	**0.9**	**0.8**	**0.7**	**0.6**
FLPF (Simulated)	1.11	6.1054	6.3855	19.020	104.557
FLPF (Theoretical)	1.0999	2.426	6.110	19.405	94.190
FHPF (Simulated)	0.89	4.4173	9.0124	67.9611	824.383
FHPF (Theoretical)	0.88	2.0835	9.018	61.271	758.168

TABLE V. MAX/MIN FREQUENCIES (F_M) (IN KHZ) OF FRACTIONAL FILTERS FOR DIFFERENT A

	1	**0.9**	**0.8**	**0.7**	**0.6**
FBPF (Simulated)	0.871	5.2481	8.1283	30.903	223.872
FBPF (Theoretical)	0.865	2.426	7.423	34.482	267.230
FBRF (Simulated)	0.871	5.1286	7.9433	33.113	234.423
FBRF (Theoretical)	0.865	2.426	7.423	34.482	267.230

TABLE VI. CUT-OFF FREQUENCIES (F_H) (IN KHZ) OF FRACTIONAL FILTERS FOR DIFFERENT I_{BIAS}

	265 k	**70 k**	**25 k**	**15 k**
FLPF (Simulated)	1.1103	4.1389	11.103	17.826
FLPF (Theoretical)	1.0996	4.1639	11.659	19.431
FHPF (Simulated)	0.6898	2.58767	7.001	11.262
FHPF (Theoretical)	0.6796	2.57315	7.2059	12.0009

TABLE VII. MAX/MIN FREQUENCIES (F_M) (IN KHZ) OF FRACTIONAL FILTERS FOR DIFFERENT I_{BIAS}

	265 k	**70 k**	**25 k**	**15 k**
FBPF (Simulated)	0.871	3.2359	8.9125	14.453
FBPF (Theoretical)	0.864	3.2735	9.1659	15.276
FBRF (Simulated)	0.871	3.2359	8.9125	14.453
FBRF (Theoretical)	0.864	3.2735	9.1659	15.276

TABLE VIII. SLOPE OF ATTENUATION IN PASSBAND/STOPBAND (DB/DECADE)

	1	**0.9**	**0.8**	**0.7**	**0.6**
FLPF	-40.01	-37.624	-31.47	-27.187	-23.778
FHPF	40.436	36.833	30.20	26.156	23.267

VI. CONCLUSION

This paper has presented a modified configuration of VDTA based fractional order universal filter and its tunability with respect to value of α and g_m. Without changing the circuit configuration, the presented circuit can realize all the five standard biquadratic fractional filters. Electronic control of its parameter ω_0 is shown through α and I_{bias}. The condition for the stability of all the modified filters are also realized in this paper which gives us the absolute idea of the location of the poles.

REFERENCES

[1] D. Biolek, R. Senani, V. Biolkova, and Z. Kolka, "Active elements for analog signal processing: classification, review, and new proposals," Radioengineering, 17, no. 4, pp. 15–32, 2008.

[2] J. Satansup, and W. Tangsrirat, "Compact VDTA-based current-mode electronically tunable universal filters using grounded capacitors," Microelectronics Journal, 45, no..6, pp. 613-618, 2014.

[3] J. Satansup, and W. Tangsrirat, "Single VDTA-based current-mode electronically tunable multifunction filter," In 4th International Science, Social Science, Engineering and Energy Conference, 2012, pp. 1-8.

[4] P. Rani, and R. Pandey, "Voltage differencing transconductance amplifier based fractional order multiple input single output universal filter," Solid State Electronics Letters 1, no. 2, pp. 110-118, 2019.

[5] D. Prasad, M. Kumar, and M. W. Akram, "Current mode fractional order filters using VDTAs with Grounded capacitors," International Journal of Electronics and Telecommunications, 65, no. 1, pp. 11-17, 2019.

[6] A. S. Ali, A. G. Radwan, and A. M. Soliman, "Fractional order Butterworth filter: active and passive realizations," IEEE journal on emerging and selected topics in circuits and systems, 3, no. 3, pp. 346-354, 2013

[7] T. J. Freeborn, B. Maundy, and A. Elwakil, "Fractional-step Tow-Thomas biquad filters," Nonlinear Theory and Its Applications, IEICE, 3, no. 3, pp. 357-374, 2012.

[8] J. Koton, J. Jerabek, N. Herencsar, and D. Kubanek, "Current conveyors in current-mode circuits approximating fractional-order low-pass filter," In 2017 European Conference on Circuit Theory and Design (ECCTD), 2017, pp. 1-4.

[9] A. Soltan, A. G. Radwan, and A. M. Soliman, "CCII based fractional filters of different orders," Journal of advanced research, 5, no. 2, pp. 157-164, 2014.

[10] L. Langhammer, R. Sotner, J. Dvorak, J. Jerabek, and J. Polak, "Fully-differential tunable fractional-order filter with current followers and current amplifiers," In 2017 27th International Conference Radioelektronika (RADIOELEKTRONIKA), 2017, pp. 1-6.

[11] A. G. Radwan, A. S. Elwakil, and A. M. Soliman, "On the generalization of second-order filters to the fractional-order domain." Journal of Circuits, Systems, and Computers, 18, no. 2, pp. 361-386, 2009.

[12] R. Verma, N. Pandey, and R. Pandey, "Electronically tunable fractional order filter," Arabian Journal for Science and Engineering, 42, no. 8, pp. 3409-3422, 2017.

[13] E. Kaskouta, T. Kamilaris, R. Sotner, J. Jerabek, and C. Psychalinos, "Single-input multiple-output and multiple-input single-output fractional-order filter designs," In 2018 41st International Conference on Telecommunications and Signal Processing (TSP), 2018, pp. 1-4.

[14] J. Jerabek, R. Sotner, D. Kubánek, J. Dvorak, L. Langhammer, N. Herencsar, and K. Vrba, "Fractional-order low-pass filter with electronically adjustable parameters," In 2016 39th International Conference on Telecommunications and Signal Processing (TSP), 2016, pp. 569-574.

[15] J. Dvorak, Z. Polesakova, J. Jerabek, L. Langhammer, A. Kartci, and J. Koton. "Non-Integer-Order Low-Pass Filter with Electronically Controllable Parameters," In 2018 IEEE International Symposium on Circuits and Systems (ISCAS), 2018, pp. 1-5.

[16] A. Oustaloup, F. Levron, B. Mathieu, and F. M. Nanot, "Frequency-band complex noninteger differentiator: characterization and synthesis," IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications 47, no. 1, pp. 25-39, 2000.

[17] J. Valsa, P. Dvorak, and M. Friedl, "Network Model of the CPE," Radioengineering, 20, no. 3, pp. 619–626, 2011.

[18] A. G. Radwan, A. M. Soliman, A. S. Elwakil, and A. Sedeek, "On the stability of linear systems with fractional-order elements," Chaos, Solitons & Fractals, 40, no. 5, pp. 2317-2328, 2009.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 20,2021 at 07:12:32 UTC from IEEE Xplore. Restrictions apply.