SCHOLARLY PUBLICATIONS

A CURRENT AWARENESS BULLETIN
OF RESEARCH OUTPUT

@DTU

(130th Edition)

OCTOBER 2023

BY: CENTRAL LIBRARY

DELHI TECHNOLOGICAL UNIVERSITY
(FORMERLY *DELHI COLLEGE OF ENGINEERING)*
GOVT. OF N.C.T. OF DELHI
SHAHBAD DAULATPUR, MAIN BAWANA ROAD
DELHI 110042

PREFACE

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

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

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

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

Central Library

CONTENTS

1. A NOVEL SIMILARITY MEASURE AND SCORE FUNCTION OF PYTHAGOREAN FUZZY SETS AND THEIR APPLICATION IN ASSIGNMENT PROBLEM, ***6.Vineet KUMAR***, ***3.Anjana GUPTA*** and ***3.Harish Chander TANEJA***, Applied Mathematics, DTU

2. A roadmap from classical cryptography to post-quantum resistant cryptography for 5G-enabled IoT: Challenges, opportunities and solutions, ***6.Diksha Chawla*** and ***3.Pawan Singh Mehra***, CSE, DTU

3. A single ended, single port configuration based 9 T SRAM cell for stability enhancement, ***8.Vansh Singhal***, ***6.Bhawna Rawat***, ***3.Poornima Mittal*** and ***8.Brijesh Kumar***, Electronics, DTU

4. A Stable Deep Learning Model for Recognizing Human Actions in Noisy Videos, ***8.Ishwar Chand***, ***3.Sanjay Kumar***, ***8.Jatin Popli*** and ***8.Jitesh Kumar***, CSE, DTU

5. A Study of Recent Advancements in Deep Learning for Natural Language Processing, ***8.Aditya Raj***, ***3.Rajni Jindal***, ***8.Aditya Kumar Singh*** & ***8.Ajay Pal***, CSE, DTU

6. A Variable-DC-Offset Modulation Scheme for Improved Performance of a Differential-Mode Inverter, Kartik Tank, ***8.1.Shantanu Gupta***, Man Mohan Garg, Sudip K. Mazumder and Moien Mohamadi, Electrical, DTU

7. A versatile 4 K insert for characterization of the superconducting joints, Ajit Nandawadekar, ***3.Mukhtiar Singh*** and Soumen Kar, Electrical, DTU

8. An approach to preserve functions with exponential growth by using modified Lupa¸s-Kantrovich operators, ***6.Neha Kajla*** and ***3.Naokant Deo***, Applied Mathematics, DTU

9. An improved teaching learning based optimization method to enrich the flight control of a helicopter system, ***3.ABHISHEK CHAUDHARY*** and ***3.BHARAT BHUSHAN***, Electrical, DTU

10. An In-Depth Analysis of Variation in Characteristic Performance of OLED with Respect to Position of Charge Generation Layer, ***6.Sugandha Yadav***, ***3.Poornima Mittal*** and Shubham Negi, Electronics, DTU

11. Analysing Effectiveness of Different Physiological Biomarkers in Detecting Stress, ***8.Anshuman Raj Chauhan***, ***8.Akhil*** and ***3.Sanjay Kumar***, CSE, DTU

12. Artifcial intelligence in cerebral stroke images classifcation and segmentation: A comprehensive study, Gourav Kumar Sharma, Santosh Kumar, ***3.Virender Ranga*** and Mahendra Kumar Murmu, IT, DTU

13. Attenuation of hotspot phenomenon in PV systems, ***6.Swatilekha Ghosh***, ***3.Santosh Kumar Singh*** and ***3.Vinod Kumar Yadav***, Environment and Electrical, DTU

14. Availability analysis of vertical milling centre using Markov approach and Monte Carlo simulations, ***3.Kumar Girish***, ***8.Mishra Ravi*** & ***8.Khan Tayyab***, Mechanical, DTU

15. Cracking the Figurative Code: A Survey of Metaphor Detection Techniques, Vrinda Kohli, ***6.Himanshu Nandanwar*** and ***3.Rahul Katarya***, CSE, DTU

16. DAC-BiNet: Twitter crime detection using deep attention convolutional bi-directional aquila optimal network, ***6.Monika*** and ***3.Aruna Bhat***, CSE, DTU

17. Design of 8 -bit low power SAR ADC in 45nm for biomedical implants, ***3.Mohit Tyagi***, ***3.Poornima Mittal*** and Parvin Kumar, Electronics, DTU

18. Detection of Lung Disorders From X-Ray Images Using Hybrid Deep Learning, ***8.Aneesha Sharma***, ***8.Anshika Yadav*** and ***3.Manoj Kumar***, CSE, DTU

19. Devanagari Character Recognition using MLP-Mixer and CNN Extracted Features, ***8.Abhay Gupta***, ***8.Abhishek Jha***, ***8.Abhishek Sharma*** & ***3.Anukriti Kaushal***, CSE, DTU

20. Development and Tribological Analysis of Bio-Based Lubricant Using Two Non-Edible Oils: Castor Oil and Neem Oil, ***8.Ayan Bindra***, ***8.Dipali Yadav***, ***8.Manan Anand*** and ***3.R.C. Singh***, Mechanical, DTU

21. Development of Conducting Biopolymer-Based Biosensor for Heavy-Metal Ion Detection, ***6.Shreya Dutta***, Ravi Butola, Bhupendra Pratap Singh and ***3.Prakash Chandra***, Biotechnology, DTU

22. Ditch incorporated organic thin film transistor based organic all-p inverter: a novel approach, Sakshi Gupta, ***3.Poornima Mittal*** and Pradeep Juneja, Electronics, DTU

23. DREAM: Deep Learning-based Recognition of Emotions from Multiple Affective Modalities using consumer-grade body sensors and video cameras, ***6.Aditi Sharma*** and Akshi Kumar, DTU

24. Edge-based computation of super-resolution Superlet spectrograms for real-time estimation of heart rate using an IoMT based reference signal less PPG sensor, Pankaj, Ashish Kumar, ***3.Manjeet Kumar*** and Rama Komaragiri, Electronics, DTU

25. Effect of sensitizer on the luminescence of thermally stable Eu3+-activated metasilicate phosphor for solar cell applications, ***6.Deepali*** and ***3.M. Jayasimhadri***, Applied Physics, DTU

26. Efficient image encryption using the Tinkerbell map in conjunction with linear feedback shift registers, ***6.Puneet Kumar Pal***, ***3.Dhirendra Kumar*** and ***8.Varun Agarwal,*** Applied Mathematics, DTU

27. Examining deep learning's capability to spot code smells: a systematic literature review, ***3.Ruchika Malhotra***, ***6.Bhawna Jain*** and Marouane Kessentini, CSE, DTU

28. Experimental Evaluation of CNG Substitution Ratio on Exhaust Gas Emissions of Diesel/CNG Dual Fuel Combustion, ***6.Neeraj KUMAR***, ***6.Bharat Bhushan ARORA*** and ***3.Sagar MAJI***, Mechanical, DTU

29. Fault Location Algorithm for Multi-terminal Radial Medium Voltage DC Microgrid, ***8.1.Vaibhav Nougain***, Sukumar Mishra, Soumya Shubhra Nag and Aleksandra Lekic, Electrical, DTU

30. Financial burden of seeking diabetesmellitus care in India: Evidence from a Nationally Representative Sample Survey, ***6.Mehak Nanda*** and Rajesh Sharma, USME, DTU

31. Finishing and Wear analysis of 3D-Printed Workpiece Through Centrifugal Force-Assisted Abrasive Flow Machining, ***6.A. Bhardwaj*** and P. Ali, Mechanical, DTU

32. Fusion of Transformer Model and Skeleton Detection Model for Abnormal Human Activity Detection with Transfer Learning, ***8.Akhilesh Kumar Verma***, ***8.Abhishek Soren***, ***8.Ankit Kumar Shrivastav*** and ***3.Sanjay Kumar***, CSE, DTU

33. Highly sensitive twin core photonic crystal fber for hazardous cancer cell detection in THz frequency regime, ***6.Vishal Chaudhary*** and ***3.Sonal Singh***, Electronics, DTU

34. Hindi Personality Net: Personality Detection in Hindi Conversational Data using Deep Learning with Static Embedding, ***8.AKSHI KUMAR***, ***8.DIPIKA JAIN*** and ***3.ROHIT BENIWAL***, CSE, DTU

35. Identifying discernible indications of psychological well-being using ML: explainable AI in reddit social media interactions, ***8.Pahalage Dona Thushari***, Nitisha Aggarwal, Vajratiya Vajrobol, Geetika Jain Saxena, Sanjeev Singh and Amit Pundir, Software Engineering, DTU

36. Insights into Structural, Electronic and Thermoelectric Properties of ZnTMN2 (TM= Zr and Hf): A First-Principles Study, ***6.Rajesh Kumar***, ***6.Ramesh Kumar***, ***6.Sangeeta***, ***6.Kulwinder Kumar*** and ***3.Mukhtiyar Singh***, Applied Physics, DTU

37. Integrated shannon entropy and COPRAS optimal model-based recommendation framework, ***6.Neha Punetha*** and ***3.Goonjan Jain***, Applied mathematics, DTU

38. Low Power Pre-charge Voltage Level and Low Swing Logic Based 8T SRAM Cell for High Speed CMOS Circuits, ***3.Ram Murti Rawat*** and ***6.V. Kumar***, CSE, DTU

39. Mechanical and surface characterization of Er2O3/La2O3/CeO2 doped carbide coating developed using high velocity oxy fuel (HVOF), ***6.Mohit Vishnoi***, ***3.Qasim Murtaza*** and ***3.Paras Kumar***, Mechanical, DTU

40. Modeling the Influence of Ground-Granulated Blast Furnace, Slag on Hydration of Cement, ***6.Andualem Yadeta***, ***3.Pradeep Goyal*** and ***3.Raju Sarkar***, Civil, DTU

41. Modelling and Manufacturing of Abrasive Water Jet Machining of Reinforced AI Alloy Cast Hybrid Nanocomposites with Mechanical and Metallurgical Behavior through Artificial Neural Network, Kingshuk Das Bakshi, Thilagham K T, R. Yamini, Vijaya Bhaskar Reddy, Anjali Sandeep Gaikwad & ***7.Anshika Gupta***, Mechanical, DTU

42. Multifunctional Sm3+ induced bismuth lithium tungstate phosphor for flexible anti-counterfeiting and encryption, ***6.Indrajeet Maurya***, ***6.Mukesh K. Sahu***, ***6.Vikas***, A. K. Vishwakarma and ***3.M. Jayasimhadri***, Applied Physics, DTU

43. Multi-response Optimization of Friction Stir Welded Reinforced Joints of Dissimilar Aluminum Alloys, ***6.Sumit Jain*** and ***3.R. S. Mishra***, Mechanical, DTU

44. New insights into the therapeutic approaches for the treatment of tauopathies, Himanshi Singh, ***3.Asmita Das***, Mohammad Moshahid Khan and Tayebeh Pourmotabbed, Biotechnology, DTU

45. Novel electrode for thermal additive centrifugal force-assisted abrasive fow machining, ***6.Anant Bhardwaj***, ***3.Krovvidi Srinivas*** and ***3.Rajiv Chaudhary***, Mechanical, DTU

46. Observational Constraints and Cosmological Implications of Scalar-Tensor f (R, T), Gravity, Amine Bouali, ***7.Himanshu Chaudhary***, Tiberiu Harko, Francisco S.N. Lobo, Taoufik Ouali and Miguel A. S. Pinto, Applied Mathematics, DTU

47. Poor air quality as an important predictor of climate change in Delhi, ***7.GAURAV YADAV***, ***3.GEETA SINGH*** and S. D. ATTRI, Environmental, DTU

48. Potato Peel Waste as an Economic Feedstock for PHA Production by Bacillus circulans, ***6.Sonika Kag***, ***3.Pravir Kumar*** and ***3.Rashmi Kataria***, Biotechnology, DTU

49. Predicting Agriculture Yields Based on Machine Learning Using Regression and Deep Learning, PRIYANKA SHARMA, PANKAJ DADHEECH, ***7.1.NAGENDER ANEJA*** AND SANDHYA ANEJA, CSE, DTU

50. Prediction of Energy Dissipation over Stepped Spillway with Baffles Using Machine Learning Techniques, ***7.Saurabh Pujari***, ***6.Vijay Kaushik*** and ***3.S. Anbu Kumar***, Civil

51. Pronunciation Scoring With Goodness of Pronunciation and Dynamic Time Warping, ARPIT BAJGOTI, RISHIK GUPTA, NISHTHA JATANA, GEETIKA DHAND, CHARU GUPTA, PANKAJ DADHEECH, UMAR YAHYA, AND ***7.1.NAGENDER ANEJA***, CSE, DTU

52. Recent developments in phycosynthesis of zinc oxide nanoparticles for biomedicine and environmental applications, ***6.Shrutika Chaudhary*** and ***3.Navneeta Bharadvaja***, Biotechnology, DTU

53. RING Domain Mutation Hinders the E3 Ligase Activity of RNF8 and Affects UBE2N Binding, ***6.Mehar Sahu*** and ***3.Pravir Kumar***, Biotechnology, DTU

54. Seismic Data Reconstruction Based on Double Sparsity Dictionary Learning With Structure Oriented Filtering, Lakshmi Kuruguntla, Vineela Chandra Dodda, ***3.Anup Kumar Mandpura***, Sunil Chinnadurai and Karthikeyan Elumalai, Electrical, DTU

55. Speaker Independent Bangla Isolated Speech Recognition Using Deep Neural Network, Dipayan Bhadra, ***6.Mehrab Hossain*** and Fatema Alam, DTU

56. Surface topography assessment using chemical assisted ball end magnetorheological finishing, ***6.Anand Sharma*** and ***3.Mahendra Singh Niranjan***, Mechanical, DTU

57. SwinAnomaly: Real-Time Video Anomaly Detection Using Video Swin Transformer and SORT, ARPIT BAJGOTI, RISHIK GUPTA, PRASANALAKSHMI BALAJI, ***6.1.RINKY DWIVEDI***, MEENA SIWACH AND ***7.1.DEEPAK GUPTA***, CSE, DTU

58. Synthesis, Structural and Fluorescence Investigations of Novel Li2Ba5W3O15:Sm3+ Phosphors for Photonic Device Applications, ***6.Anu*** and ***3.A. S. Rao***, Applied Physics, DTU

59. Thermodynamic Properties of Conical Quantum Dot Modulated by External Fields and Rashba Spin–Orbit Interaction, ***8.Yash Gupta***, ***8.Pranay Khosla***, ***8.Sakshi Arora***, ***8.Priyanka*** and ***3.Rinku Sharma***, Applied Physics, DTU

60. Unsupervised domain alignment of fingerprint denoising models using pseudo annotations, Indu Joshi, ***8.Tushar Prakash***, ***3.Rohit Kumar***, Antitza Dantcheva Sumantra Dutta Roy, Prem Kumar Kalra, Mathematics, DTU

61. Unveiling the Potential of Ti3C2Tx MXene for Gas Sensing: Recent Developments and Future Perspectives, Nitesh K. Chourasia, Ankita Rawat, Ritesh Kumar Chourasia, Hemant Singh, ***6.Ramesh Kumar Kulriya***, ***3.Vinod Singh*** and Pawan Kumar Kulriya, Applied Physics, DTU

62. Use of Poly Vinyl Chloride (PVC) Tape to Increase the Strength of Bamboo, ***8.Rishabh Rastogi***, ***8.Yogesh***, ***8.Nikhil Kumar Pankaj*** and ***3.B. R. G. Robert***, Civil, DTU

63. Using discussion thread context in Sentiment analysis for improving Cyberbullying detection, ***3.Manoj Kumar***, ***8.Himanshu***, ***8.Vishal Choudhary*** and ***8.Yash Nishal***, CSE, DTU

64. Visual-Motion-Interaction Guided Pedestrian Intention Prediction Framework, ***6.Neha Sharma***, ***3.Chhavi Dhiman*** and ***3.S. Indu***, Electronics, DTU

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

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

Vineet KUMAR, PhD Candidate (corresponding author)
E-mail: vineet_dtuphd002191@dtu.ac.in
Department of Applied Mathematics
Delhi Technological University, Delhi, India

Professor Anjana GUPTA, PhD
E-mail: anjanagupta@dce.ac.in
Department of Applied Mathematics
Delhi Technological University, Delhi, India

Professor Harish Chander TANEJA, PhD
E-mail: hctaneja@dce.ac.in
Department of Applied Mathematics
Delhi Technological University, Delhi, India

A NOVEL SIMILARITY MEASURE AND SCORE FUNCTION OF PYTHAGOREAN FUZZY SETS AND THEIR APPLICATION IN ASSIGNMENT PROBLEM

Abstract. *In real-life problems, things are imprecise because of imprecision/inaccuracy, and the exact value of the measured quantities is impossible to get. Sometimes, due to time pressure/ incomplete knowledge, it is difficult for the decision-makers to provide their opinion. To describe the imprecision, the information in terms of the fuzzy is provided to allow the decision-makers to express their inputs freely. There is a valuable role of the fuzzy set (FS) and intuitionistic fuzzy set (IFS) to describe uncertainty under the uncertain situations. In the literature, various models are available for assignment problems under fuzzy sets and intuitionistic fuzzy sets. The Pythagorean fuzzy set (PFS) has a larger domain space than the intuitionistic fuzzy set to describe the membership grade. To handle the uncertainty in practical applications of assignment problems (AP), we have proposed a method to solve the Pythagorean fuzzy assignment problem (PFAP) using the proposed similarity measure and a score function. Numerical examples are given to explain the methodology.*

Keywords*: Pythagorean Fuzzy Set, Similarity Measure, Score Function, Assignment Problem*

JEL Classification: C02, C44, C45, C60, C61, C62

DOI: 10.24818/18423264/57.3.23.19

1. Introduction

An assignment problem is a linear programming problem (LPP) that deals with allocation and scheduling. The problem of assignment arises because available resources have varying degrees of efficiency for performing different activities. In classical assignment problems, it is assumed that the decision-maker is sure about the precise value of the cost of the assignment problem but, practically, these factors are imprecise.

Zadeh (1965) introduced the fuzzy set (FS) to deal with uncertainty in real-life problems. Gurukumaresan et al. (2020) used the centroid method for the solution of the fuzzy assignment problem. Tsai et al. (1999) worked on the multiobjective fuzzy deployment of manpower. Chanas et al. (1984) took the demand and supply as the fuzzy numbers in the transportation problem and used parametric programming for solving the problem. Verma & Merigó (2019) worked on generalised similarity measures for Pythagorean fuzzy sets and their applications to multiple attribute decision-making. Kumar and Gupta (2011) solved the fuzzy assignment problems and fuzzy travelling salesman problems with different membership functions. Yuen and Ting (2012) performed the textbook selection using the fuzzy PROMETHEE II method. Thakre et al. (2018) worked on the placement of staff in LIC using the fuzzy assignment problems.

To consider the vague and imprecise information in the practical problem, the different extensions of the fuzzy set have been introduced by some authors. The intuitionistic fuzzy set (IFS) proposed by Atanassov (1984) is an extension of the fuzzy set. He considered the membership and nonmembership of the element. Roseline and Amirtharaj (2015) solved the intuitionistic fuzzy assignment problem by using the ranking of intuitionistic fuzzy numbers (IFN). Boran et al. (2012) used the TOPSIS method of the intuitionistic fuzzy set for solving the renewable energy problem. Mukherjee and Basu (2012) solved the assignment problem under IFS by using similarity measure and score function. Kumar and Bajaj [2014] introduced the problem of an interval-valued intuitionistic fuzzy assignment problem and solved it with similarity measure and score function.

Yager (2013) introduced a Pythagorean fuzzy set (PFS). Yager overcomes the situation when membership degree (τ) + nonmembership degree (ζ) >1 in IFS. PFS is an extension of IFS with the condition that the square sum of the membership degree and the nonmembership degree is less than or equal to 1 (τ_A(u)^2+ζ_A(u)^2 $\leq$ 1). The concept of Pythagorean fuzzy sets (PFS) gives the larger preference domain for decision makers (DM). DMs can define their support and against the degree of membership as $\tau(x)$ = 4/5, $\zeta(x)$= 2/5. In this case, 4/5+2/5>1 is not valid in IFS but squaring (4/5)^2 + (2/5)^2 <1 implies the Pythagorean fuzzy set is more suitable than the intuitionistic fuzzy set. Paul Augustine Ejegwa (2019) worked on the Pythagorean fuzzy set and its application in career placement using max-min composition. Fei and Deng [18] solved the problem of the Pythagorean fuzzy multi-criteria problem. Shahzadi et al. (2018)

proposed the solution of the decision-making approach under the Pythagorean fuzzy Yager weighted operators. Peng and Yang (2015) defined some results for Pythagorean fuzzy sets.

Over the period, score function and similarity measure of Pythagorean were introduced by many authors. Agheli et al. (2022) defined the new similarity measure for Pythagorean fuzzy sets and application on multiple-criterion decision- making. Zhang and Xu (2014) worked on TOPSIS for multi-criteria decision- making with PFS. Peng & Yang (2016) defined the score function and distance measure for the interval-valued Pythagorean fuzzy number (IVPFN) to analyse the problem. After that Garg [25] proposed the score function for PFN and IVPFN to overcome some limitations of the score function defined by Peng & Yang (2016).

In this work, we have developed a methodology to solve the assignment problem with Pythagorean fuzzy values. The score function defined by Garg (2017) has some limitations. To overcome these limitations, we have proposed a new score function. Additionally, we have defined the new similarity measure to validate our result. So far, there is no literature regarding Pythagorean fuzzy assignment problems using similarity measure and score function.

The paper is organised as follows: Some basic knowledge of FS, IFS, PFS, and arithmetic operations on Pythagorean fuzzy numbers are discussed in section 2. In section 3, we have proposed a novel similarity measure and score function. Also, the limitations of previously defined score functions have been pointed out. The methodology to solve PFAP using similarity measure and score function is given in section 4. Illustrative examples are also given in this section. Section 5, presents the comparative study and concluding remarks.

2. Preliminaries

In this section, we have discussed some basic definitions and arithmetic operations that are required for our work.

Definition 2.1 (1965) A fuzzy set (FS) $\tilde{A}$ is defined on universal set U as

$$\tilde{A} = \{\langle u, \tau_A(u) | u \in U\rangle\},$$

characterized by the membership function

$$\tau_A(u): U \to [0,1].$$

Here $\tau_A(u)$ is the membership degree of the element u to the set $\tilde{A}$.

Definition 2.2 (1984) An intuitionistic fuzzy set $\tilde{A}$ on U is defined as a set of ordered pair given by

$$\tilde{A} = \{\langle u, \tau_{\tilde{A}}(u), \zeta_{\tilde{A}}(u)\rangle | u \in U\},$$

where $\tau_{\tilde{A}}(u), \zeta_{\tilde{A}}(u): U \to [0,1]$ are the degree of membership and degree of non-membership of the element $u \in U$, with the condition $(\tau_{\tilde{A}}(u)) + (\zeta_{\tilde{A}}(u)) \leq 1$, the degree of indeterminacy is given by $\xi_{\tilde{A}}(u) = 1 - \tau_{\tilde{A}}(u) - \zeta_{\tilde{A}}(u)$.

Definition 2.3 (2013) A Pythagorean fuzzy set $\tilde{A}$ on U is defined as given by

$$\tilde{A} = \{< u, \tau_A(u), \zeta_A(u) > | u \in U\},$$

where $\tau_A(u), \zeta_A(u): U \to [0,1]$ are the degree of membership and degree of non-membership of the element $u \in U$, with the condition $(\tau_A(u))^2 + (\zeta_A(u))^2 \leq 1$, the degree of indeterminacy is given by $\xi_A(u) = \sqrt{1 - (\tau_A^2 + \zeta_A^2)}$.

The domain of a Pythagorean fuzzy set is larger than intuitionistic fuzzy sets. While working in the space of PFS, one may have much more choice of assigning value to member and nonmembership from [0, 1].

Definition 2.4 (2015) The addition, multiplication, and scalar multiplication on two PFNs $\tilde{A}_1 =< \tau_{A_1}(u), \zeta_{A_1}(u) >$ and $\tilde{A}_2 =< \tau_{A_2}(u), \zeta_{A_2}(u) >$ are defined as follows:

(i) $\tilde{A}_1 \oplus \tilde{A}_2 = \left\langle \sqrt{\tau_{A_1}^2 + \tau_{A_2}^2 - \tau_{A_1}^2 \tau_{A_2}^2}, \zeta_{A_1}\zeta_{A_2} \right\rangle,$

(ii) $\tilde{A}_1 \otimes \tilde{\mathrm{A}}_2 = \left\langle \tau_{A_1}\tau_{A_2}, \sqrt{\zeta_{A_1}^2 + \zeta_{A_2}^2 - \zeta_{A_1}^2\zeta_{A_2}^2} \right\rangle,$

(III) $k\tilde{A}_1 = \left\langle \sqrt{1 - (1 - \tau_{A_1}^2)^k}, \zeta_{A_1}^k \right\rangle, k > 0.$

3. Similarity Measure and Score Function of Pythagorean Fuzzy Set

In this section, we have defined the novel similarity measure and score function of Pythagorean fuzzy sets.

Definition 3.1: Suppose $\tilde{A}$ and $\tilde{B}$ be two PFSs. The similarity measure SM: $\tilde{A} \times \tilde{B} \to [0, 1]$ is defined as follows

$$\mathrm{S}(\tilde{A}, \tilde{B}) = \frac{\sum_{j=1}^{m} \tau_A^2(u_j).\tau_B^2(u_j) + \zeta_A^2(u_j).\zeta_B^2(u_j)}{\sum_{j=1}^{m} \left[\left(\tau_A^4(u_j) \vee \tau_B^4(u_j)\right) + \left(\zeta_A^4(u_j) \vee \zeta_B^4(u_j)\right)\right]}$$

Theorem 3.1: Similarity measure (SM) between two PFS $\tilde{A}$ and $\tilde{B}$, then the following are true.

(S1) $0 \leq S(\tilde{A}, \tilde{B}) \leq 1$

(S2) $S(\tilde{A}, \tilde{B}) = 1$ iff A=B

(S3) $S(\tilde{A}, \tilde{B}) = S(\tilde{B}, \tilde{A})$

(S4) $S(\tilde{A},\tilde{C}) \le S(\tilde{A},\tilde{B})$ and $S(\tilde{A},\tilde{C}) \le S(\tilde{B},\tilde{C})$ for all $\tilde{A}$, $\tilde{B}$, $\tilde{C}$ such that $\tilde{A} \subseteq \tilde{B} \subseteq \tilde{C}$.

Proof: (S1) Since for all u_j, 1 ≤ j ≤ m, we have $\tau^2_{\tilde{A}}(u_j).\tau^2_{\tilde{B}}(u_j) \le \tau^4_{\tilde{A}}(u_j) \vee \tau^4_{\tilde{B}}(u_j)$ and $\zeta^2_{\tilde{A}}(u_j).\zeta^2_{\tilde{B}}(u_j) \le \zeta^4_{\tilde{A}}(u_j) \vee \zeta^4_{\tilde{B}}(u_j)$. Therefore, for each u_j, we have

$$\left[\tau^2_{\tilde{A}}(u_j).\tau^2_{\tilde{B}}(u_j) + \zeta^2_{\tilde{A}}(u_j).\zeta^2_{\tilde{B}}(u_j)\right] \le \left[\left\{\tau^4_{\tilde{A}}(u_j) \vee \tau^4_{\tilde{B}}(u_j)\right\} + \left\{\zeta^4_{\tilde{A}}(u_j) \vee \zeta^4_{\tilde{B}}(u_j)\right\}\right]$$

Therefore, for all u_j, 1 ≤ j ≤ m, we have

$$\sum_{j=1}^{m}\left[\tau^2_{\tilde{A}}(u_j).\tau^2_{\tilde{B}}(u_j) + \zeta^2_{\tilde{A}}(u_j).\zeta^2_{\tilde{B}}(u_j)\right] \le$$

$$\sum_{j=1}^{m}\left[\left\{\tau^4_{\tilde{A}}(u_j) \vee \tau^4_{\tilde{B}}(u_j)\right\} + \left\{\zeta^4_{\tilde{A}}(u_j) \vee \zeta^4_{\tilde{B}}(u_j)\right\}\right]$$

$$0 \le S^s(\tilde{A},\tilde{B}) \le 1.$$

(S2). Suppose $S(\tilde{A},\tilde{B}) = 1$, $\dfrac{\sum_{j=1}^{m}\left[\tau^2_{\tilde{A}}(u_j).\tau^2_{\tilde{B}}(u_j) + \zeta^2_{\tilde{A}}(u_j).\zeta^2_{\tilde{B}}(u_j)\right]}{\sum_{j=1}^{m}\left[\left\{\tau^4_{\tilde{A}}(u_j) \vee \tau^4_{\tilde{B}}(u_j)\right\} + \left\{\zeta^4_{\tilde{A}}(u_j) \vee \zeta^4_{\tilde{B}}(u_j)\right\}\right]} = 1$

$$\sum_{j=1}^{m}\left[\tau^2_{\tilde{A}}(u_j).\tau^2_{\tilde{B}}(u_j) + \zeta^2_{\tilde{A}}(u_j).\zeta^2_{\tilde{B}}(u_j)\right] = \sum_{j=1}^{m}\left[\left\{\tau^4_{\tilde{A}}(u_j) \vee \tau^4_{\tilde{B}}(u_j)\right\} + \left\{\zeta^4_{\tilde{A}}(u_j) \vee \zeta^4_{\tilde{B}}(u_j)\right\}\right].$$

Now, we claim that $\tau^2_{\tilde{A}}(u_j).\tau^2_{\tilde{B}}(u_j) = \tau^4_{\tilde{A}}(u_j) \vee \tau^4_{\tilde{B}}(u_j)$ and $\zeta^2_{\tilde{A}}(u_j).\zeta^2_{\tilde{B}}(u_j) = \zeta^4_{\tilde{A}}(u_j) \vee \zeta^4_{\tilde{B}}(u_j)$.

Suppose $\tau^2_{\tilde{A}}(u_j).\tau^2_{\tilde{B}}(u_j) \ne \tau^4_{\tilde{A}}(u_j) \vee \tau^4_{\tilde{B}}(u_j)$, since $\tau^2_{\tilde{A}}(u_j).\tau^2_{\tilde{B}}(u_j) \le \tau^4_{\tilde{A}}(u_j) \vee \tau^4_{\tilde{B}}(u_j)$, there exists $k_1 > 0$ such that $\tau^2_{\tilde{A}}(u_j).\tau^2_{\tilde{B}}(u_j) + k_1 = \tau^4_{\tilde{A}}(u_j) \vee \tau^4_{\tilde{B}}(u_j)$.

Similarly, there exists $k_2 > 0$ such that $\zeta^2_{\tilde{A}}(u_j).\zeta^2_{\tilde{B}}(u_j) + k_2 = \zeta^4_{\tilde{A}}(u_j) \vee \zeta^4_{\tilde{B}}(u_j)$. By hypothesis it follows that $k_1 + k_2 = 0$. This implies $k_1 = -(k_2)$, which is not possible. This implies that

$\tau^2_{\tilde{A}}(u_j).\tau^2_{\tilde{B}}(u_j) = \tau^4_{\tilde{A}}(u_j) \vee \tau^4_{\tilde{B}}(u_j)$ and $\zeta^2_{\tilde{A}}(u_j).\zeta^2_{\tilde{B}}(u_j) = \zeta^4_{\tilde{A}}(u_j) \vee \zeta^4_{\tilde{B}}(u_j)$.

This implies that $\tau^2_{\tilde{A}}(u_j) = \tau^2_{\tilde{B}}(u_j)$ and $\zeta^2_{\tilde{A}}(u_j) = \zeta^2_{\tilde{B}}(u_j)$.

Hence $\tilde{A} = \tilde{B}$.

(S3) $S(\tilde{A}, \tilde{B}) = S(\tilde{B}, \tilde{A})$ is trivial.

(S4) For three PFSs $\tilde{A}, \tilde{B}$ and $\tilde{C}$ in U. The similarity measures between $\tilde{A}, \tilde{B}$ and $\tilde{A}, \tilde{C}$ are given as

$$S(\tilde{A}, \tilde{B}) = \frac{\sum_{j=1}^{m}\left[\tau^2_{\tilde{A}}(u_j).\tau^2_{\tilde{B}}(u_j) + \zeta^2_{\tilde{A}}(u_j).\zeta^2_{\tilde{B}}(u_j)\right]}{\sum_{j=1}^{m}\left[\left\{\tau^4_{\tilde{A}}(u_j) \vee \tau^4_{\tilde{B}}(u_j)\right\} + \left\{\zeta^4_{\tilde{A}}(u_j) \vee \zeta^4_{\tilde{B}}(u_j)\right\}\right]}$$

$$S(\tilde{A}, \tilde{C}) = \frac{\sum_{j=1}^{m}\left[\tau^2_{\tilde{A}}(u_j).\tau^2_{\tilde{C}}(u_j) + \zeta^2_{\tilde{A}}(u_j).\zeta^2_{\tilde{C}}(u_j)\right]}{\sum_{j=1}^{m}\left[\left\{\tau^4_{\tilde{A}}(u_j) \vee \tau^4_{\tilde{C}}(u_j)\right\} + \left\{\zeta^4_{\tilde{A}}(u_j) \vee \zeta^4_{\tilde{C}}(u_j)\right\}\right]}.$$

Let $\tilde{A} \subseteq \tilde{B} \subseteq \tilde{C}$ for all $u_j \in \tilde{U}$, we have $\tau^2_{\tilde{A}}(u_j) \le \tau^2_{\tilde{B}}(u_j) \le \tau^2_{\tilde{C}}(u_j)$, $\zeta^2_{\tilde{A}}(u_j) \ge \zeta^2_{\tilde{B}}(u_j) \ge \zeta^2_{\tilde{C}}(u_j)$. This implies $\tau^4_{\tilde{A}}(u_j) \le \tau^4_{\tilde{B}}(u_j) \le \tau^4_{\tilde{C}}(u_j)$, $\zeta^4_{\tilde{A}}(u_j) \ge \zeta^4_{\tilde{B}}(u_j) \ge \zeta^4_{\tilde{C}}(u_j)$.

We claim that for all $u_j \in \tilde{U}$, we have

$$\frac{\tau^2_{\tilde{A}}(u_j).\tau^2_{\tilde{B}}(u_j)}{\tau^4_{\tilde{B}}(u_j) + \zeta^4_{\tilde{A}}(u_j)} \le \frac{\tau^2_{\tilde{A}}(u_j).\tau^2_{\tilde{C}}(u_j)}{\tau^4_{\tilde{C}}(u_j) + \zeta^4_{\tilde{A}}(u_j)}$$

Similarly, we have

$$\frac{\zeta^2_{\tilde{A}}(u_j).\zeta^2_{\tilde{B}}(u_j)}{\tau^4_{\tilde{B}}(u_j) + \zeta^4_{\tilde{A}}(u_j)} \le \frac{\zeta^2_{\tilde{A}}(u_j).\zeta^2_{\tilde{C}}(u_j)}{\tau^4_{\tilde{C}}(u_j) + \zeta^4_{\tilde{A}}(u_j)}$$

by adding all above equations, we have

$$\frac{\sum_{j=1}^{m}\left[\tau_{\tilde{A}}^{2}(u_j).\tau_{\tilde{C}}^{2}(u_j)+\zeta_{\tilde{A}}^{2}(u_j).\zeta_{\tilde{C}}^{2}(u_j)\right]}{\sum_{j=1}^{m}\left[\left\{\tau_{\tilde{A}}^{4}(u_j)\vee\tau_{\tilde{C}}^{4}(u_j)\right\}+\left\{\zeta_{\tilde{A}}^{4}(u_j)\vee\zeta_{\tilde{C}}^{4}(u_j)\right\}\right]}\leq \frac{\sum_{j=1}^{m}\left[\tau_{\tilde{A}}^{2}(u_j).\tau_{\tilde{B}}^{2}(u_j)+\zeta_{\tilde{A}}^{2}(u_j).\zeta_{\tilde{B}}^{2}(u_j)\right]}{\sum_{j=1}^{m}\left[\left\{\tau_{\tilde{A}}^{4}(u_j)\vee\tau_{\tilde{B}}^{4}(u_j)\right\}+\left\{\zeta_{\tilde{A}}^{4}(u_j)\vee\zeta_{\tilde{B}}^{4}(u_j)\right\}\right]}.$$

Therefore $S(\tilde{A},\tilde{C})\leq S^{s}(\tilde{A},\tilde{B})$. Similarly $S^{s}(\tilde{A},\tilde{C})\leq S(\tilde{B},\tilde{C})$.

Next, we discuss the limitations of the previously defined score function and to overcome the limitations a new score function propose in this section.

Peng & Yang [7] defined score function and accuracy function for interval-valued Pythagorean fuzzy number (IVPFN):

Consider IVPFN $\widetilde{W}=\langle[\alpha,\beta],[\gamma,\delta]\rangle$, the score function $E_1(\widetilde{W})$ and accuracy function $F_1(\widetilde{W})$ are defined as follows

$$E_1(\widetilde{W})=\frac{\alpha^2+\beta^2-\gamma^2-\delta^2}{2}$$

$$F_1(\widetilde{W})=\frac{\alpha^2+\beta^2+\gamma^2+\delta^2}{2}$$

Suppose that if we consider two IVPFN $\widetilde{W}_1=\langle[0.1,0.2],[0.4,0.5]\rangle$ and $\widetilde{W}_2=\langle[0.1,0.2],[0.5,0.5]\rangle$, the score and accuracy value calculated by Peng & Yang (2016) are $E_1(\widetilde{W}_1)$=-0.1800, $E_1(\widetilde{W}_2)$=-0.1800 and $F_1(\widetilde{W}_1)$=0.2300, $F_1(\widetilde{W}_2)$=0.2300.

According to Peng & Yang (2016) $\widetilde{W}_1\sim\widetilde{W}_2$, but we have seen that $\widetilde{W}_1\neq\widetilde{W}_2$. To overcome this limitation Garg (2017) defined the new score function $E_2(\widetilde{W})$ and defined as

$$E_2(\widetilde{W})=\frac{(\alpha^2-\gamma^2)(1+\sqrt{1-\alpha^2-\gamma^2})+(\beta^2-\delta^2)(1+\sqrt{1-\beta^2-\delta^2})}{2}\in[-1,1]$$

By this score function, the score value of above example are $E_2(\widetilde{W}_1)=-0.3368$ and $E_2(\widetilde{W}_2)=-0.3233$. Here $E_2(\widetilde{W}_2)>E_2(\widetilde{W}_1)$, hence $\widetilde{W}_2>\widetilde{W}_1$.

For IVPFN $\widetilde{W}=\langle[0.81,0.87],[0.11,0.25]\rangle$, the score value $E_2(\widetilde{W})$=1.0022, this is invalid because score value is greater than 1 i.e. $E_2(\widetilde{W})=1.0022\notin[-1,1]$.

Proposed Score Function: We have seen from the above example that the score function defined by Garg (2017) is not giving the appropriate result. To improve this, we have proposed a novel score function as follows.

Definition 3.2 Let $\widetilde{W} = \langle [\alpha, \beta], [\gamma, \delta] \rangle$ be an interval valued Pythagorean fuzzy number (IVPFN). The score function for IVPFN is

$E(\widetilde{W}) = \frac{(\sqrt{3+\alpha-3\gamma})+(\sqrt{3+\beta-3\delta})}{4} \in [0,1]$.

In this if $\alpha = \beta = \tau$ and $\gamma = \delta = \zeta$, then the score function of Interval-valued Pythagorean fuzzy set will become score function of Pythagorean fuzzy set. So, the proposed score function of the Pythagorean fuzzy set is as follows:

$E(\widetilde{W}) = \frac{\sqrt{3+\tau-3\zeta}}{2} \in [0,1]$.

For any two IVPFN/PFN $\widetilde{W}_1$ and $\widetilde{W}_2$,

1. if $E(\widetilde{W}_1) > E(\widetilde{W}_2)$, then $\widetilde{W}_1 > \widetilde{W}_2$
2. if $E(\widetilde{W}_1) < E(\widetilde{W}_2)$, then $\widetilde{W}_1 < \widetilde{W}_2$
3. if $E(\widetilde{W}_1) = E(\widetilde{W}_2)$, then $\widetilde{W}_1 = \widetilde{W}_2$.

4. Application of the Pythagorean Fuzzy Assignment Problem

In this section, we introduce the assignment problem with Pythagorean fuzzy number (PFN) and give two methodologies to solve such problems. One is based on similarity measure and the other is based on a score function.

Pythagorean Fuzzy Assignment Problem (PFAP)

Min $\tilde{Y} = \sum_i^n \sum_j^n \tilde{c}_{ij}^{PFN} x_{ij}$

Subject to

$\sum_j^n x_{ij} = 1, i = 1,2,\ldots n$

$\sum_i^n x_{ij} = 1 j = 1,2,\ldots n$

$x_{ij} \in \{0,1\}$

In assignment problems, the cost is usually deterministic in nature. But in real-life problem, this is very difficult to judge the precise value of the cost. In this unstable condition, we calculate the preference value. Based on preference value, we get the preference for the j^{th} work to the i^{th} person in the form of a composite relative degree of similarity with an ideal solution, Thus we replace c_{ij} by composite relative degree.

4.1. Methodology for Pythagorean Fuzzy Assignment Problem using similarity measure

Step 1 First consider the Pythagorean fuzzy assignment problem decision matrix $G=\{(L_{ij})\}_{m\times n}$

$(L_{ij}) = < \tau_{ij}(x), \zeta_{ij}(x) >$, *i=1, 2....m, j= 1,.....n* are Pythagorean fuzzy numbers.

Step 2 Examine whether the problem is balanced or not. If it is not balanced, then add dummy variables so that the problem is converted into a balanced assignment problem.

Step 3 Calculate the similarity measure of each cost value from Pythagorean positive ideal solution (PPIS) $L^{+} = \langle 1,0 \rangle$ and Pythagorean negative ideal solution (PNIS) $L^{-} = \langle 0,1 \rangle$

$$S\ (L, L^{+}) = \frac{\sum_{j=1}^{m} \tau_{L}^{2}(x_j).\tau_{L^{+}}^{2}(x_j) + \zeta_{L}^{2}(x_j).\zeta_{L^{+}}^{2}(x_j)}{\sum_{j=1}^{m} \left[\left(\tau_{L}^{4}(x_j) \vee \tau_{L^{+}}^{4}(x_j)\right) + \left(\zeta_{L}^{4}(x_j) \vee \zeta_{L^{+}}^{4}(x_j)\right)\right]}$$

$$S\ (L, L^{-}) = \frac{\sum_{j=1}^{m} \tau_{L}^{2}(x_j).\tau_{L^{-}}^{2}(x_j) + \zeta_{L}^{2}(x_j).\zeta_{L^{-}}^{2}(x_j)}{\sum_{j=1}^{m} \left[\left(\tau_{L}^{4}(x_j) \vee \tau_{L^{-}}^{4}(x_j)\right) + \left(\zeta_{L}^{4}(x_j) \vee \zeta_{L^{-}}^{4}(x_j)\right)\right]}$$

Relative similarity matrix calculated column-wise

$$Q = \frac{S(L,L^{+})}{S(L,L^{+}) + S(L,L^{-})}$$

Similarly, relative similarity matrix calculated row-wise

$$R = \frac{S(L,L^{+})}{S(L,L^{+}) + S(L,L^{-})}.$$

Step 4 The composite matrix $[T]_{n \times n}$ is evaluated as $T = Q \times R = q_{ij} \times r_{ij}$, the resultant matrix T represents the preference that j th job is chosen by i th person.

4.2 Methodology for Pythagorean fuzzy assignment problem using the score function

Step 1 Write PFAP in tabular form

Step 2 Convert the Pythagorean fuzzy assignment problem into a crisp assignment problem by using the score function.

Step 3 Examine whether the problem is balanced or not. If it is not balanced, then add dummy variables so that the problem is converted into a balanced assignment problem.

Step 4 The higher cell value of the matrix will indicate the preference of j th job to the i th person

Illustrative Examples:

Here, we have solved the assignment problem by using similarity measure and score function.

Example 4.1: A manufacturing company decides to make six subassemblies through six contractors. One contractor has to receive only one subassembly. The cost of each subassembly is determined by the bids submitted by each contractor and is shown in Table 1 in Pythagorean fuzzy number. The problem is how to assign subassemblies to contractors to get the optimal assignment.

Table 1. Assignment problem based on PFN

S, Q	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6
S_1	(0.7,0.6)	(0.7,0.7)	(0.8,0.5)	(0.7,0.6)	(0.6,0.7)	(0.6,0.5)
S_2	(0.63,0.67)	(0.9,0.5)	(0.8,0.53)	(0.8,0.3)	(0.9,0.2)	(0.45,0.59)
S_3	(0.83,0.4)	(0.5,0.7)	(0.6,0.7)	(0.5,0.7)	(0.20,0.81)	(0.5,0.8)
S_4	(0.63,0.55)	(0.71,0.63)	(0.66,0.35)	(0.9,0.3)	(0.4,0.8)	(0.73,0.4)
S_5	(0.7,0.5)	(0.65,0.35)	(0.32,0.7)	(0.8,0.5)	(0.4,0.9)	(0.85,0.18)
S_6	(0.45,0.75)	(0.83,0.3)	(0.35,0.7)	(0.55,0.8)	(0.5,0.6)	(0.3,0.8)

Solution:

Table 2. S (L, L $^+$) column-wise

S, Q	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6
S_1	0.43	0.39	0.60	0.43	0.29	0.33
S_2	0.33	0.76	0.59	0.63	0.8	0.18
S_3	0.67	0.19	0.29	0.20	0.02	0.17
S_4	0.36	0.43	0.42	0.80	0.11	0.51
S_5	0.46	0.41	0.08	0.60	0.09	0.72
S_6	0.15	0.68	0.09	0.21	0.22	0.06

Table 3. S (L, L $^-$) column-wise

S, Q	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6
S_1	0.29	0.39	0.17	0.29	0.43	0.22
S_2	0.38	0.23	0.19	0.05	0.02	0.33
S_3	0.10	0.46	0.43	0.46	0.65	0.60
S_4	0.26	0.31	0.10	0.05	0.62	0.12
S_5	0.20	0.10	0.48	0.17	0.15	0.02
S_6	0.54	0.06	0.48	0.58	0.33	0.63

Table 4. Relative similarity matrix R (column-wise)

S, Q	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6
S_1	0.59	0.5	0.77	0.59	0.40	0.6
S_2	0.46	0.76	0.75	0.92	0.97	0.35

S, Q	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6
S_3	0.87	0.29	0.4	0.3	0.02	0.22
S_4	0.58	0.58	0.80	0.94	0.15	0.80
S_5	0.69	0.8	0.14	0.77	0.37	0.97
S_6	0.21	0.91	0.15	0.26	0.4	0.08

Similarly

Table 5. Relative similarity matrix S (row-wise)

SQ	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6
S_1	0.34	0.25	0.59	0.34	0.16	0.36
S_2	0.21	0.57	0.56	0.84	0.94	0.12
S_3	0.75	0.08	0.16	0.09	0.00	0.04
S_4	0.33	0.33	0.64	0.88	0.02	0.64
S_5	0.47	0.64	0.01	0.59	0.13	0.94
S_6	0.04	0.82	0.02	0.06	0.16	0.0

Now compute the composite matrix T=R× S=$r_{ij} \times s_{ij}$. This matrix T represents the preference the S^{th} subassembly to C^{th} contractor

Table 6. the preference the S^{th} subassembly to C^{th} contractor

S, Q	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6
S_1	0.11	0.06	0.34	0.11	0.02	0.12
S_2	0.04	0.32	0.31	0.70	0.88	0.01
S_3	0.56	0.0	0.02	0.0	0.0	0.0
S_4	0.10	0.10	0.4	077	0.0	0.4
S_5	0.22	0.40	0.0	0.34	0.01	0.88
S_6	0.0	0.67	0.0	0.0	0.02	0.0

The optimal assignment policy is: subassembly 1→contractor 3, subassembly 2→contractor 5, subassembly 3→contractor 1, subassembly 4→contractor 4, subassembly 5→contractor 6, subassembly 6→contractor 2.

Next, we consider the same Example 4.1 as given above and apply the proposed score function for the Pythagorean fuzzy number (Definition 3.2). We get the Table 7 corresponding to Table 1.

Solution for PFAP:

Table 7. Value calculated by score function

S, Q	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6
S_1	0.689	0.632	0.758	0.689	0.612	0.725
S_2	0.636	0.775	0.743	0.851	0.908	0.648
S_3	0.811	0.592	0.612	0.592	0.439	0.524
S_4	0.704	0.675	0.808	0.866	0.500	0.795
S_5	0.742	0.806	0.552	0.758	0.418	0.910
S_6	0.548	0.856	0.559	0.536	0.652	0.474

The optimal assignment policy is: subassembly 1→contractor 3, subassembly 2→contractor 5, subassembly 3→contractor 1, subassembly 4→contractor 4, subassembly 5→contractor 6, subassembly 6→contractor 2.

5. Comparative analysis and conclusions

The examples mentioned above have vividly demonstrated the proposed similarity measure and score function as a potential tool for solving the assignment problem in Pythagorean fuzzy sets. From the analysis, it is observed that the results obtained by the implementation of the developed similarity measure and score function are more accurate and reliable. Compared to the existing methodologies in the literature, the novel score function proposed in the present paper has the following advantages:

(i) The proposed method has a simple presentation such that it can significantly avoid the information loss that may have previously occurred in the score function defined by Peng & Yang (2016) and Garg's (2017). It is envisioned that there exist certain values where Peng & Yang (2016) and Garg (2017) score function failed to give valid results.

(ii) We have also observed that Example 4.1 cannot be solved by using the score function given by Garg (2017) as the score values of the cell (2,2) representing $\mathcal{S}_2\mathcal{Q}_2$ (0.9, 0.5).

(iii) The diversity and fuzziness of the decision maker's assessment information can be well reflected and modelled using the proposed similarity measure and score function.

(iv) The result offered by using the novel similarity measure and score function is consistent with the result obtained in the existing work, Mukherjee and Basu (2012), and Kumar and Bajaj (2014). Therefore, the proposed method becomes more flexible and convenient for solving the Pythagorean fuzzy assignment problem.

In this paper, we have proposed a methodology to solve the Pythagorean fuzzy assignment problem. We have solved the problem using the similarity measure and the score function to test the optimality of the problem. It is anticipated that the proposed methodology is capable of managing the uncertainty persisting within the intricate assignment problem. The working of proposed technique has been illustrated via examples to test the validity. We further provide a comparison with the existing methods in the literature. From the comparative study and analysis, it can be concluded that the proposed method overcomes the limitations present in the existing work. Table 8 provides a comparative analysis of the proposed score function. Additionally, it would be engrossing to explore the application of the developed approach to picture fuzzy sets, spherical fuzzy sets and interval-valued picture fuzzy sets, etc., also to deal with other linear programming problems.

Table 8. Comparative analysis of the present work

Problem	Score Function by Garg (2017)	Score Function by Peng & Yang (2016)	Proposed Score Function & Similarity Measure
PFAP	Failed	This is not valid when score values & accuracy are the same	Solution exists

REFERENCES

[1] **Agheli, B., Adabitabar Firozja, M., Garg, H. (2022),** ***Similarity measure for Pythagorean fuzzy sets and application on multiple criteria decision making.*** *Journal of Statistics and Management Systems*, 25(4), 749-769;

[2] **Atanassov, K. (1984),** ***Intuitionistic fuzzy sets****. Fuzzy Sets and Systems,* 20 (1), 87-96;

[3] **Boran, F. E., Boran, K., Menlik, T. (2012),** ***The evaluation of renewable energy technologies for electricity generation in Turkey using intuitionistic fuzzy TOPSIS*****.** *Energy Sources, Part B: Economics, Planning, and Policy*, 7(1), 81-90;

[4] **Chanas, S., Kołodziejczyk, W., Machaj, A. (1984),** ***A fuzzy approach to the transportation problem*****.** *Fuzzy sets and Systems*, 13(3), 211-221;

[5] **Ejegwa, P. A. (2019),** ***Pythagorean fuzzy set and its application in career placements based on academic performance using max–min–max composition.*** *Complex & Intelligent Systems*, 5(2), 165-175;

[6] **Fei, L., Deng, Y. (2020),** ***Multi-criteria decision making in Pythagorean fuzzy environment*****.** *Applied Intelligence,* 50(2), 537-561;

[7] **Garg, H. (2017),** ***A new improved score function of an interval-valued Pythagorean fuzzy set based TOPSIS method****. International Journal for Uncertainty Quantification,* 7(5), 463-474;

[8] **Gurukumaresan, D., Duraisamy, C., Srinivasan, R., Vijayan, V. (2020),** ***Optimal solution of fuzzy assignment problem with centroid methods*****.** *Materials Today: Proceedings*, 37, 553-555;

[9] **Kumar, A., Gupta, A. (2011),** ***Methods for solving fuzzy assignment problems and fuzzy travelling salesman problems with different membership functions****. Fuzzy Information and Engineering,* 3(1), 3-21;

[10] **Kumar, G., & Bajaj, R. K. (2014).** ***On solution of interval valued intuitionistic fuzzy assignment problem using similarity measure and score function*****.** *International Journal of Mathematical and Computational Sciences, 8*(4), 715-720.

[11] **Mukherjee, S., Basu, K. (2012),** ***Solution of a class of intuitionistic fuzzy assignment problem by using similarity measures.*** *Knowledge-Based Systems,* 27, 170-179;

[12] **Peng, X., Yang, Y. (2015),** ***Some results for Pythagorean fuzzy sets***. *International Journal of Intelligent Systems,* 30(11), 1133-1160;

[13] **Peng, X., Yang, Y. (2016),** ***Fundamental properties of interval- valued Pythagorean fuzzy aggregation operators.*** *International Journal of Intelligent Systems,* 31(5), 444-487;

[14] **Roseline, S., Amirtharaj, H. (2015),** ***Methods to find the solution for the intuitionistic fuzzy assignment problem with ranking of intuitionistic fuzzy numbers.*** *International Journal of Innovative Research in Science, Engineering and Technology,* 4(7), 10008–10014;

[15] **Shahzadi, G., Akram, M., Al-Kenani, A. N. (2020),** ***Decision-making approach under Pythagorean fuzzy Yager weighted operators.*** *Mathematics,* 8(1), 70;

[16] **Thakre, T. A., Chaudhari, O. K., Dhawade, N. R. (2018),** ***Placement of staff in LIC using fuzzy assignment problem.*** *International Journal of Mathematics Trends and Technology (IJMTT),* 53(4), 259-266;

[17] **Tsai, C. H., Wei, C. C., Cheng, C. L. (1999),** ***Multi-objective fuzzy deployment of manpower.*** *International Journal of the Computer, the Internet and Management,* 7(2), 1-7;

[18] **Verma, R., Merigó, J. M. (2019),** ***On generalized similarity measures for Pythagorean fuzzy sets and their applications to multiple attribute decision- making.*** *International Journal of Intelligent Systems,* 34(10), 2556-2583;

[19] **Yager, R. R. (2013),** ***Pythagorean membership grades in multi-criteria decision making.*** *IEEE Transactions on Fuzzy Systems,* 22(4), 958-965;

[20] **Yuen, K. K. F., Ting, T. O. (2012),** ***Textbook selection using fuzzy PROMETHEE II method.*** *International Journal of Future Computer and Communication,* 1(1), 76;

[21] **Zadeh, L. A. (1965),** ***Fuzzy sets.*** *Information and Control,* 8(3), 338-353;

[22] **Zhang, X., Xu, Z. (2014),** ***Extension of TOPSIS to multiple criteria decision making with Pythagorean fuzzy sets.*** *International Journal of Intelligent Systems,* 29(12), 1061-1078.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/374172386

A roadmap from classical cryptography to post-quantum resistant cryptography for 5G-enabled IoT: Challenges, opportunities and solutions

Article *in* Internet of Things · September 2023

DOI: 10.1016/j.iot.2023.100950

CITATIONS
0

READS
65

2 authors:

Diksha Chawla
Delhi Technological University
4 PUBLICATIONS **5** CITATIONS
SEE PROFILE

Pawan Singh Mehra
Delhi Technological University
50 PUBLICATIONS **616** CITATIONS
SEE PROFILE

All content following this page was uploaded by Diksha Chawla on 05 October 2023.

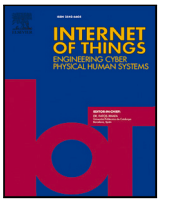
The user has requested enhancement of the downloaded file.

Internet of Things 24 (2023) 100950

Contents lists available at ScienceDirect

Internet of Things

journal homepage: www.elsevier.com/locate/iot



Review article

A roadmap from classical cryptography to post-quantum resistant cryptography for 5G-enabled IoT: Challenges, opportunities and solutions

Diksha Chawla, Pawan Singh Mehra *

Department of Computer Science and Engineering, Delhi Technological University, Delhi 110042, India

ARTICLE INFO

Keywords:
Internet of Things
Wireless sensor networks
Quantum cryptography
Post quantum cryptography
Security
Review

ABSTRACT

5th Generation(5G) -enabled Internet of Things (IoT) connects billions of devices for high-speed data transfer. The data collected fuel many automation applications such as smart cities, industrial automation 5.0, autonomous vehicles and healthcare sectors. However, the critical challenges concerning such an environment are Open Web Security and Mutual Authentication. The recent security protocols of IoT are based on hard mathematical cryptographic structures. Nevertheless, dependency on the security of classical cryptographic primitives is at risk of being broken by Quantum Computing. Recent Quantum attacks on IoT security have motivated us to survey Quantum-resistant solutions comprehensively. Therefore in this paper, we analysed how the technology shifts from classical cryptographic techniques towards quantum-enabled solutions to achieve end-to-end security. The main objective of this survey is to provide a systematic roadmap in the area of Quantum secured 5G-enabled IoT communication by covering current research on 5G-enabled IoT with its key enabling technologies, threats imposed on 5G-enabled IoT applications, state-of-the-art quantum-based solutions and initiatives. A detailed overview of the quantum computing preliminaries, Post Quantum Cryptography (PQC) schemes, Quantum Cryptography (QC), and Quantum Key Distribution (QKD) with their significant advantage in securing IoT-enabled communication over classical scenarios are also presented. Particularly, we emphasize Post-Quantum resistant techniques that must be used soon to secure 5G-enabled IoT communications. The comparative analysis of quantum-resistant schemes with classical cryptographic schemes in terms of key size, data size, time complexity and impact of quantum computers on these schemes are discussed. This paper identifies the technical challenges in Quantum-based schemes and provides future research directions to enable Post-Quantum resistant cryptography for secure communication in 5G-enabled IoT.

1. Introduction

5G wireless technology is transforming and revolutionizing every aspect of the future communication system. 5G technology is revolutionizing the global world with high-quality and quantity of data rates. As depicted in Fig. 1 IoT using 5G infrastructure supports massive ultra-reliable Machine to Machine interactions, Device to Everything (D2E), Vehicle to Infrastructure (V2I) and Vehicle Pedestrian (V2P) to access real-time data. The 5G facilitates users with many benefits such as high-speed data transfer and low latency which enable users to gather critical information anytime and anywhere.

* Corresponding author.
E-mail address: pawansinghmehra@gmail.com (P.S. Mehra).

https://doi.org/10.1016/j.iot.2023.100950
Received 22 April 2023; Received in revised form 6 September 2023; Accepted 20 September 2023
Available online 26 September 2023
2542-6605/© 2023 Elsevier B.V. All rights reserved.

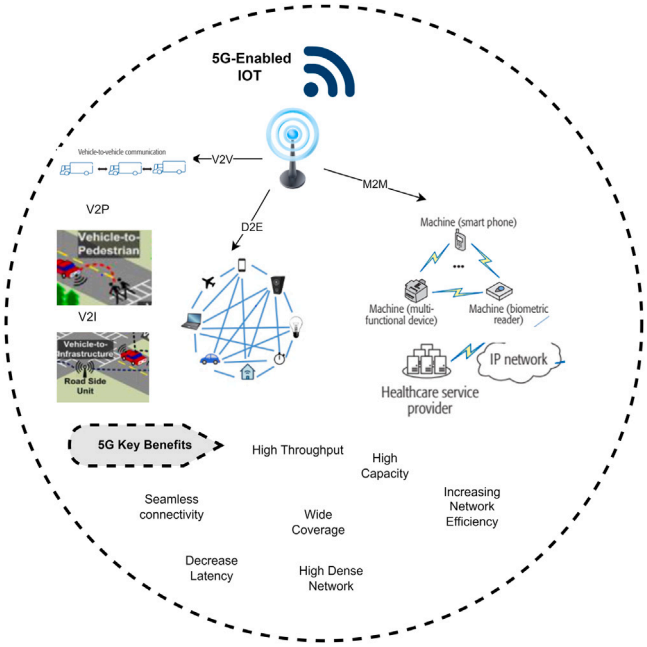


Fig. 1. IoT, by enabling 5G benefits, supports massive ultra-reliable Machine-to-Machine (M2M), Device-to-Everything (D2E) and Vehicle-to-Everything (V2E) interactions.

5G will open up newer possibilities in enabling intelligent devices to fulfil the enormous demand for futuristic IoT frameworks. Moreover, the 5G-IoT communication network will be faster, with more connecting devices supported in heterogeneous network environments [1]. According to Gartner [2], almost 57% of organizations are working on 5G to drive Internet of Things (IoT) Communication. 5G will aid IoT applications such as intelligent agriculture, intelligent environments, securities and emergencies. It is possible because of the improved latency and bandwidth. These applications will ride on the top of the 5G network, requiring more security to lock down the new devices and connections.

In the IoT framework, there are many devices connected. These devices are vulnerable to attacks such as man-in–middle, impersonation, replay, gateway node bypassing and Distributed Denial of Service (DDoS) attacks. Such devices need mutual authentication to verify their counterparts. The discussions regarding application-specific mutual authentication protocols have dominated research in recent years [3,4]. Most widely accepted classical cryptographic techniques, such as RSA, rely on popular asymmetric cryptosystems [5]. The security of RSA is based on the prime factorization method. Once Quantum computing algorithms [4] are available, they threaten the security of most classical algorithms based on symmetric and asymmetric cryptography. The Quantum computing-based Shor's algorithm [6] shows concerns about the security of the prime factorization-based cryptographic algorithm.

In prior research, various authentication techniques [7] have been proposed for safe communication. Later, it was found that all those techniques were vulnerable to security attacks launched by Quantum computers [8]. Quantum computers have exponential power in computation because of Quantum bits (Qubits). In our work, we analysed that hackers are storing a lot of data they can decrypt later by using quantum computers. Therefore, the novelty of our work is to provide the roadmap from classical to Quantum resistant schemes by understanding 5G-enabled IoT security concerns and limitations of classical cryptosystems. It is considered the most urgent requirement to prepare for Quantum-resistant algorithms that can also withstand classical computer attacks. Therefore, our work includes an in-depth discussion of Post Quantum Cryptography (PQC) and Quantum Cryptography (QC) techniques. This article reviews the recent progress of Quantum-based solutions for secure communication in 5G-enabled IoT. The challenges in securing IoT communication, Quantum opportunities and solutions are also considered in our work.

Although several other research proposals exist in the literature covering the Quantum safe communication for IoT [9,10]. However, as per the author's knowledge, the independent systematic review that provides an in-depth discussion on Post-Quantum resistant techniques considering the high-speed connectivity of 5G in the IoT framework is not yet proposed. Therefore, our work pertains to recent research on Quantum resistant techniques to overcome security issues concerning IoT communication [11]. Quantum-resistant solutions include both Post Quantum Cryptography (PQC) and Quantum Cryptography (QC). One of the advantages of using Quantum-based schemes is the Quantum parallelism that provides a significant speedup over hard mathematical calculations. Quantum parallelism is achieved by preparing qubits in a superposition of states. These qubits are used for Quantum Cryptography (QC). However, Post Quantum Cryptography (PQC) prepare us for the era of Quantum computing by revising existing hard mathematical-based algorithms. Due to the limitations of classical cryptographic primitives, it is reasonable to expect Post Quantum Cryptography (PQC) and Quantum Cryptography (QC) techniques for secure data sharing in the 5G-enabled IoT network [12].

1.1. Motivation

Other surveys available in the literature on Post-Quantum computing provide valuable information; however, they do not consider PQC and QC for 5G-enabled IoT communication scenarios. They are focused on a particular type of cryptosystem. In contrast, this article considers thorough revisions on the most relevant aspects influencing Quantum-resistance cryptography-based solutions for IoT developments. This paper aims to investigate Quantum-resistant schemes to secure 5G-enabled IoT applications from Quantum attacks. It covers the recent threats to 5G-enabled IoT applications, classical countermeasures, challenges, and Quantum-resistant solutions. There is a need to identify open challenges and future research directions to secure user privacy.

1.2. Contributions

To the best of our knowledge, our survey is the first that includes an in-depth discussion of Quantum resistant schemes, 5G-enabled IoT security and privacy concerns, Quantum-based solutions for 5G-enabled IoT security, reviews the existing work published up to 2023, serving as a guide for other researchers to understand and apply the existing protocols. The article aims to provide novel contributions, which have not been found altogether in previous literature as follows:

1. Given the continuous developments of 5G-enabled IoT communication systems and their heavy deployment, it is important to examine the technical foundations of 5G concerning IoT requirements, vision and recent trends in 5G-enabled IoT. We discussed how the adoption and utilization of Waveforms, Numerology, frame structure, Millimeter-wave and Massive Multiple-Input and Multiple-Output would be in service in fulfilling the massive connectivity and heterogeneity of the IoT landscape more effectively.
2. A comprehensive analysis to bring about the advantages of 5G for futuristic IoT applications covering requirements and objectives.
3. A detailed discussion of the vulnerabilities and security considerations in 5G-enabled IoT systems by dividing them into five major application areas smart homes, smart farming, e-healthcare and industrial IoT.
4. Review of major developments and research in key management protocols and authentication schemes for IoT communication along with their security issues and challenges that need to be solved in a post-Quantum scenario.
5. Comprehensive discussions on the capabilities of attackers are also deliberated.
6. After detailing the impact of quantum computing on classical cryptographic primitives for IoT, the main focus of our survey is to provide a systematic review of Post-Quantum resistant authentication schemes for 5G-enabled IoT.
7. To identify and discuss the Quantum Key Distribution (QKD) protocols based on the principle of uncertainty and entanglement and how they will be considered for ensuring mutual authentication.
8. A detailed study of the comparison of classical and quantum-resistant authentication schemes, main challenges and future research directions for developing a secure 5G-enabled IoT framework.

1.3. Outline of the paper

The rest of our work is organized as follows. Section 2 reviews the most recent Quantum-resistant schemes and initiatives. Section 3discusses the vision of a 5G-enabled IoT communication scenario. Sections 4 and 5 outline the technical advancement 5G brings in the IoT communication network and research trends. Attacks on IoT applications with classical countermeasures are reviewed in Section 6. Section 7 discusses 5G-enabled IoT application's challenges, vulnerabilities and possible countermeasures. Section 8 discusses the classical cryptographic schemes for secure IoT communication. Section 9 provides an in-detail discussion on authentication and key management. Quantum mechanics fundamentals and the impact of Quantum computing on classical cryptographic primitives are described in Section 10. Section 11 describes Quantum-resistant authentication schemes. The advantages of Quantum resistant schemes are described in Section 12. Quantum-Key Distribution (QKD) protocols and their associated technical challenges are described in Section 13. The security analysis of Quantum-resistant schemes and the comparative analysis of PQC schemes with classical cryptographic primitives are explained in Sections Section 14 and Section 15. The conclusion and future research directions in Quantum Computing and information for securing 5G-enabled IoT communication are discussed in Section 16. The symbols and abbreviations with their descriptions used throughout our work are listed in Table 1.

1.4. Research methodology

Table 2 represents the paper selection criterion. It contains the selection review process, search criterion, databases, and inclusion and exclusion criterion. The paper selection process consists of the following two-level screening and security criteria:

1. **Title and Abstract level Screening:** Initially, we selected papers from 1989 to 2023 and two essential papers 2021 [10], and 2022 [13] based on PQC and QC are considered. In this level of screening, on publication title and abstract, we used inclusion/exclusion criterion. Both authors independently analysed the search results and resolved disagreements through discussions.
2. **Full-Text Screening:** At this level, we analysed the complete paper and based on that we applied the inclusion and exclusion criteria. However, if we found two papers that reviewed similar content then we considered the one which has provided in-depth detail.

Table 1
List of abbreviations/symbols.

Abbreviation	Description	Abbreviation	Description
IoT	Internet of Things	ECC	Elliptic Curve Cryptography
WSN	Wireless Sensor Networks	H2M	Human–To–Machine
OWS	Open Web Security	5G NR	5G New Radio
MA	Mutual Authentication	GWN	Gateway node
QKD	Quantum Key Distribution	AES	Advanced Encryption Standard
M2M	Machine to Machine	DES	Data Encryption Standard
D2E	Device to Everything	OTP	One Time Password
PQC	Post-Quantum Cryptography	QFT	Quantum Fourier Transform
QC	Quantum Cryptography	QW	Quantum Walk
V2V	Vehicle-to-Vehicle communication	X	Position Operator
V2X	vehicle-to-anything	H	Plank's Constant
AKA	Authentication and key agreement	CAQW	Controlled alternate Quantum Walk
DoS	Denial of Service	AQDC	Authenticated Quantum Direct Communication(AQDC)
DDoS	Distributed Denial of Service	$\langle v\|\|v\rangle$	Bra-Ket Vector
MIMA	Man in the Middle attack	AKE	Authenticated Key Exchange
MAC	Media Access Control	FFT	Fast Fourier transform
QBER	Quantum bit error rate	QK-GRID	Quantum Key GRID
QWHF	Quantum Walk Hash Function	DI-QKD	Device-Independent QKD
CAQW	Controlled Alternate Quantum Walks	QRO	Quantum Random Oracle
AKD	Authenticated Key Distribution	BB84	Charles Bennett and Gilles Brassard
EPR	Einstein–Podolsky –Rosen paradox	QKD	Quantum Key Distribution

3. **Security Criteria:** We performed a security analysis intending to extract papers with sufficient information on 5G-enabled IoT security issues and current IoT security solutions, including key management, mutual authentication and cryptographic techniques. The main focus of selected papers concerning IoT security has been using Quantum solutions (Quantum Cryptography or Post Quantum Cryptography) which are applied in different IoT environments such as smart homes, smart cities and smart healthcare. We critically examined the selected security issues and current solutions to develop new directions based on:

   (a) **Quantum-security context:** The selected papers must provide sufficient information as to how Quantum-resistant solutions are being exploited to address the security issues in the broad domain of IoT.
   (b) **Quantum authentication:** The papers based on Quantum authentication exploiting key management issues and providing solutions.

Based on above listed Quantum-enabled security solutions IoT security issues are investigated. Enough IoT security concerns were reported by the study which is important to understand the limitations of classical cryptography and Quantum potential for secure IoT communication.

2. Literature review

In 2019, Bhatt et al. [14] discussed the importance of IoT communication. They discussed various attacks on IoT frameworks such as Mirai botnet, Wi-Fi baby monitor, TRENDnet webcam hack and Stuxnet, due to which more security is required. In their work, they

Table 2
Review methodology, inclusion, exclusion criterion, databases.

Property	Category
Publications	Journal articles Transaction papers Conference, symposium and workshop papers Master thesis of the reputed institute.
Year	1989–2023
	First-level screening using title and abstract followed by Full-text.
Search Strings	Security issues in IoT Mutual authentication and Key distribution 5G impact on IoT Quantum Computing impact on IoT Quantum cryptography Quantum cryptography- based protocols Post Quantum cryptography Quantum tools
Automated search in digital libraries	https://scholar.google.com/ https://www.springer.com/in https://ieeexplore.ieee.org/Xplore/home.jsp https://www.scopus.com/ https://www.sciencedirect.com/ https://www.webofscience.com/wos/author/search
Classification of papers	Security concerns in IoT Classical security countermeasures 5G for smart automation IoT applications and architecture Quantum key distribution Post Quantum Cryptography schemes Mutual authentication
Inclusion Criterion	Focus on IoT security, Quantum cryptography and Post Quantum Cryptography
Exclusion Criterion	Similar content papers found in conferences Weak presentation papers short papers

analysed the drawbacks of classical cryptography and exhibited that QC-based QKD protocols provide secure key exchange methods. The main focus of the authors is to discuss QC-based secure key distribution. However, they do not address Quantum entanglement and Quantum teleportation for secure communication between IoT-enabled devices. In 2019, Cao et al. [15] in their work analysed resource-constrained Narrowband Internet of Things (NB-IoT) devices. It includes dynamic topology changes, bandwidth limitations, and the complex environment of the network. In future IoT systems, devices cannot connect to 5G networks if network congestion happens. The fast progress of Quantum computing-related technologies can handle these difficult problems efficiently. Therefore, the authors proposed a technology based on lattice encryption. As produced by the 3GPP, 5G is the latest generation telecommunication standard. The security of 5G communication relied on cryptography. However, the scheme is limited to lattice-based homomorphic encryption technology for NB-IoT devices.

In 2020, Mitchell et al. [11] analysed how Quantum computing affects the security adopted by 5G in the real world. Classical cryptosystems such as RSA and ECC relied on key distribution mechanisms. However, the channel used to communicate the key is insecure. The author mainly focused on changes required by symmetric and asymmetric encryption schemes for 5G mobile security. But fails to address Quantum-based schemes for futuristic mobile communication. Althobaiti et al. [16] addressed the problem of IoT applications, such as smart cities, intelligent systems, smart homes, smart agriculture, healthcare, drones etc. As these applications contain critical user information, they are under cyber-attacks. Quantum computers make the current cryptographic algorithm unsafe. Hence, they proposed a PQC-based lattice homomorphic encryption algorithm. However, their work lacks security analysis concerning major attacks on IoT communication such as DoS attacks, MIMA and secret key attacks.

In 2021, Ralegankar et al. [17] provide a survey on QC for UAV communication. They discussed the different roles of UAV communication in health care and intelligent deliveries in logistics. UAV communications are rapidly growing due to fast, timely, cost-effective deliveries. However, these communication systems are highly susceptible to various security attacks such as data modification, DoS, snooping, dispatch system, MIMA, and Wi-Fi attacks. Therefore, the authors discussed the role of asymmetric cryptosystems and blockchain-based solutions to secure UAV communication systems. Their scheme is limited to UAV communication and fails to provide an in-depth description of Post-Quantum resistant solutions for smart IoT applications. Soon, with the advancements in IoT billions of devices are connected through the internet. IoT communication security and privacy become the paramount concern. Therefore, Malina et al. [10] reviewed the security issues concerning intelligent infrastructure and provided PQC-based solutions. However, the authors fail to address the QC-based solutions for secure IoT communication.

Table 3
Other researchers work on QC and PQC.

Literature Papers	Basics of Quantum computing	Classical Crypto graphic Primitives	Post Quantum Crypto	Quan-tum Crypto	Basics of Quantum Mechanics	Quantum Crypto graphy Protocols	Quantum IoT	5G enabled IoT	Future directions	MA	Security Consi-derations
1	–	✓(∗)	✓(∗)	–	–	–	–	–	–	–	–
2	–	✓(∗∗)	–	✓(∗)	✓(∗)	✓(∗)	✓(∗∗)	–	✓(∗∗)	✓(∗)	✓(∗∗)
3	✓(∗)	−−	–	✓(∗)	✓(∗)	✓(∗)	✓(∗∗)	–	−−	–	–
4	–	✓(∗)	✓(∗∗)	–	–	–	✓(∗)	✓(∗∗)	–	✓(∗∗)	✓(∗)
5	✓(∗)	✓(∗)	✓(∗)	–	–	–	–	–	–	✓(∗∗)	✓(∗)
6	–	–	–	✓(∗)	–	✓(∗)	–	–	✓(∗)	✓(∗)	✓(∗∗)
7	✓(∗)	✓(∗∗)	✓(∗∗)	–	–	–	✓(∗∗)	–	✓(∗∗)		✓(∗∗)
8	✓(∗)	–	–	✓(∗)	✓(∗)	–	✓(∗∗)	–	–	✓(∗∗)	✓(∗)
9	✓(∗)	–	–	✓(∗∗)	✓(∗)	-	✓(∗∗)	–	–	✓(∗)	✓(∗∗)
10	✓(∗)	–	✓(∗∗)	✓(∗)	✓(∗)	✓(∗)	–	–	✓(∗)	✓(∗)	✓(∗)
Our Work	✓(∗)	✓(∗∗)	✓(∗∗)	✓(∗)	✓(∗)	✓(∗)	✓(∗)	✓(∗)	✓(∗)	✓(∗)	✓(∗)

1: Malina et al. [22]; 2: Lohachab et al. [23]; 3: Bhatt et al. [24]; 4: Cao et al. [25]; 5: Mitchell et al. [26]; 6: Arul et al. [12]; 7: Althobaiti et al. [16]; 8: Ralegankar et al. [27]; 9: kumar et al. [28]; 10: Gill et al. [13]; 11: Our Survey[For comprehensive overview(*); For Detailed Description(**)].

In 2022, Gill et al. [13] discussed Quantum mechanics as the basic foundation of Quantum computing. Therefore, they discussed quantum computing properties such as entanglement and superposition. In their work, the challenges of implementing quantum-enabled communication due to fault-tolerant and reliable quantum computations are also discussed. They also highlighted that due to the fragile nature of Quantum states, the Quantum bits need to be operated at very low temperatures, and fabrication should be accurate. However, the authors do not address the Quantum security for IoT communication. Xu et al. [18] proposed a Quantum key agreement protocol for secure key distribution. Their protocol is based on Greenberger-Horne–Zeilinger (GHZ) states. It allows only legitimate parties to exchange secret keys. Their proposed scheme lacks implementation details. Li et al. [19] proposed the authentication protocol for multiparty secure communication. Their Quantum multiparty authentication protocol is based on GHZ states. Their scheme lacks in addressing MIMA, DoS and traffic analysis attacks. In 2023, Primaatmaja et al. [20] reviewed the device-independent Quantum key distribution. However, their scheme lacks in addressing QC and PQC-related techniques. In the same year, Chawla et al. [21] discussed the impact of Quantum computing on classical cryptographic schemes. However, the authors do not address the in-detail description of Quantum based solutions.

Given these limitations in the existing schemes, this study offers a roadmap that considers systematic and cognitive approaches for secure 5G-enabled IoT. Therefore, our work discussed everything from 5G-enabled IoT communication benefits, challenges and future perspectives to Post- Quantum resistant solutions for 5G-enabled IoT communication. Table 3 represents the comparison of previous surveys with our survey.

3. A vision of 5G enabled IoT

IoT is an encouraging innovation to understand the perspective of associated living. The futuristic smart innovation of IoT supports many applications such as smart homes, smart health care sector, smart agriculture and smart industries [29]. As shown in Fig. 2, the connected living framework represents a smart association of e-health care, industries, home and agriculture. These devices connected to the internet support a wide range of applications which can save people time and money.

The 3rd Generation Partnership Project (3GPP) [30] proposed 5G standards for Authentication and Key Agreement (5G-AKA). Such standards were proposed for future Narrow-Band Internet of Things (NB-IoT) which has evolved from traditional Long Time Evolution (LTE) Technology. Consequently, NB-IoT has been used in various application areas like smart agriculture, parking, asset tracking, remote meter reading, etc.[31]. 5G connectivity promised to deliver multi-Gbps high data speeds, low latency, reliability and tremendous network connectivity. The main issues associated with 5G radio access technology need to be addressed, such as massive growth in connected devices and data traffic. 5G services can be categorized based on the connected services offered [32]:

(a) **Enhanced Mobile Broadband (eMBB):** 5G mobile technology can usher in new fascinating experiences such as by providing XR as the of Augmented Reality (AR) and Virtual reality (VR), with seamless connectivity of 5G and use cloud services.
(b) **Mission-Critical Communication:** Such technology can provide services like vehicle-to-vehicle communication (V2V) and factory automation, providing high reliability and low latency.
(c) **Massive Machine Type Communication (mMTC):** This massive variety of interconnected devices. 5G enable a wide range of connectivity and coverage. It also connects low-cost battery-powered sensors, actuators, trackers, and wearables.
(d) **Ultra-Reliable low latency communication (URLLC) :** This is system-centric, focusing on reliability and latency. The education sector could create virtual reality (VR)- supported distance learning opportunities. 5G could allow doctors to squint into a patient's body using ultra-high-resolution imaging without cutting them in medical applications.

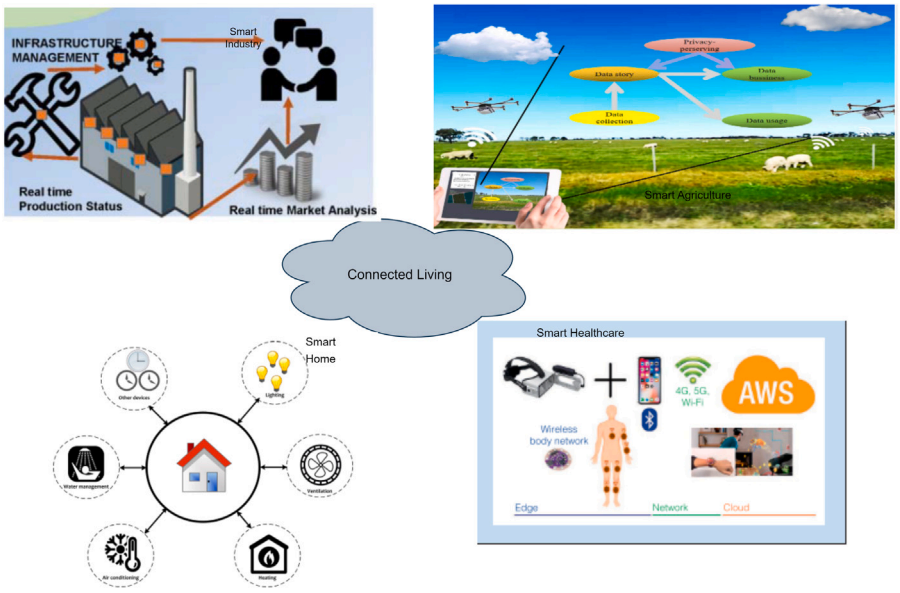


Fig. 2. A vision of automated, connected living.

Therefore, current generations will be enhanced by transitioning to the next generation: 5G. Fig. 3 represents specific 5G benefits that are required to be able to serve various smart devices and applications. It also represents the future high-performance targets as per International Mobile Telecommunications (IMT-2020) that 5G aims to achieve:

5G networks provide benefits to future IoT connectivity in comparison to 4G networks as discussed below [33]:

- Compared to the 4G/LTE network, a 100 times faster data rate is supported by 5G to support futuristic IoT applications.
- Currently, 4G supports a 100–150 Mb/s data rate; however, 5G provides approximately 10–32 Gb/s.
- Previously in 4G, high latency issues arose in cloud computing; however, 5G associated with Fog computing (FC) reduces end-to-end (E2E) latency.
- The accessibility of D2D communication is increased due to 5G, as it enhances the battery power life by almost ten times.
- 4G supports low bandwidth, approximately 10 MHz; however, 5G supports higher bandwidth, approximately 60 GHz.

4. Technical foundation of 5G

5G enables huge growth in the data generated by connected devices. The future generation 5G technology, with its diversified requirements and characteristics, supports a wide range of applications [34]. Therefore, the 3rd Generation Partnership Project (3GPP) realize the need to evolve and standardize components for a new radio access technology (RAT). It then serves diverse use cases. These use cases operate in higher frequencies up to 100 GHz. The evolutions, integral parts of 5G, are Long Term Evolution (LTE) and New Radio (NR). The three essential requirements for the physical design of NR are:

4.1. Waveforms, numerology and frame structure

NR waveform and numerology should be created by considering various link types. To support D2D communication, it provides uplink (UL), downlink (DL), sidelink and backhaul. It also provides support for vehicle-to-everything (V2X) communication. The efficient time and frequency utilization for waveform, numerology and frame structure are necessary to deploy Time Division Duplex (TDD) and frequency division duplex (FDD). As per 3GPP, 5G Orthogonal Frequency division multiplexing (OFDM) technology [1] has been introduced to support high spectral efficiency. This technology is proposed to meet high data rate requirements. Therefore, spectral efficiency is required to meet high data requirements. High spectral efficiency is required in future smart applications such as vehicular communication.

4.2. Millimeter-wave

Millimeter-wave (mm-Wave) provide multi-gigabit communication services. It includes services such as high-definition television (HDTV) and ultra-high-definition video (UHDV) [35]. The huge bandwidth band from approximately 30 to 300 GHz has been supported by mm-Wave. Moreover, due to the continuous demand for mobile traffic, there is often a contradiction between spectrum

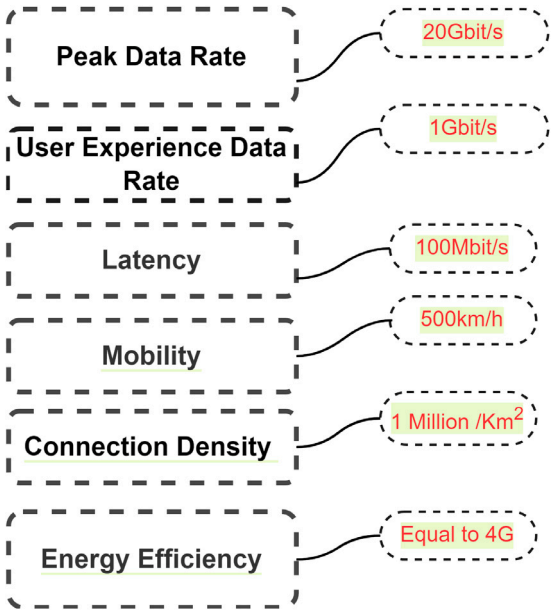


Fig. 3. Performance Targets for 5G IoT as per IMT-2020.

Table 4
mm-wave propagation characteristics and applications in different frequency bands.

Frequency band(GHz)	Path loss exponent		Rain attenuation		Applications
	Line of Sight (LOC)	Non-Line of Sight (NLOC)	5 mm/h (db)	25 Mm/h (db)	
28	1.8–1.9	4.5–4.6	0.18	0.9	In-Band backhaul
38	1.9–2.0	2.7–3.8	0.26	1.4	Access and backhaul
60	2.23	4.19	0.44	2	HD video; access, backhaul and D2D; uplink channel access
73	2	2.45–2.69	0.6	2.4	Multimedia

shortage and capacity requirements. The tremendous spectrum provided to the fifth generation to fulfil the data demand of users of mm-Wave bands. In addition, 5G-enabled IoT devices support massive numbers of connected devices, including autonomous vehicles and connected-car technology, smart city sensors for transport and security, medical devices and AR/VR applications. Table 4 shows the difference between channel characterization of Line of Sight (LOC) and Non–Line of Sight (NLOC) [36].

4.3. Massive multiple-input and multiple-output (MIMO)

It is utilized by 5G mm-Wave and provides high throughput and frequency reused over a small distance, allowing efficient spectrum use. It increases cellular capacity and coverage by using many antennas.5G is specifically designed to support massive MIMO [31], using up to 256 antenna elements in the base station, limiting fitting antennas inside a mobile device. Therefore, MIMO enables intelligent beamforming, beam tracking and beam steering in spectrum bands under 6 GHz. It also enables a 5G mm-wave network to deliver high capacity and efficiency.

5. Recent trends 5G-enabled IoT

In recent years, the increased spectrum of 5G has boosted the connection of many IoT devices [26]. Covering various manufacturers, industries, and consumer applications into a single model requires extensive network and standardization. However, 12 leading companies had already laid the foundation for 5G to boost automation and networking. The major players in the 5G IoT industry and their contributions towards implementing 5G-enabled IoT for the betterment of society are shown in Table 5.

Table 5
Major industries in the 5G IoT industry.

Research industries	Contributions
Samsung	Research on 5G technology was started by Samsung in 2011. Consider the company now one of the 5G domain leaders. Compared to current 4G network [1]. 5G provides data transmission several hundred times faster. Samsung provides an extensive contribution to the IoT open cloud platform. It lets users control home appliances like ACs, refrigerators, washing machines, etc. Some of the key developments of Samsung towards the 5G era: They developed a breakthrough in the 5G-ready antenna. and power amplifier technologies Samsung's Galaxy Watches with biometric sensors, military, heavy manufacturing, mining, etc., are future 5G applications that provide numerous IT solutions.
Huawei	Huawei realized recent advancements in the digital transformation of ICT network infrastructure. Therefore, investing a lot of amount into research focus on 5G wireless networks. The company is also patenting Key technologies. Some of the list activities of Huawei in the 5G domain: Provided integration of NB-IoT chip such as smart energy monitoring via NB-IoT devices. Provided intelligent Dual-Link features and ultra-high-speed broadband.
Qualcomm	Qualcomm R&D includes mm-wave antenna technology. The company also leads the overall 5G chipmaker and leads the 5G spectrum. Qualcomm is also moving its steps in designing and standardizing the new 5G, NR unified air interface. Key contributions towards 5G-enabled IoT: 5G new radio standards outdoor and multi-carrier aggregation boost signals into the Gbps range. They also contribute to cloud analytics virtualized core network functions.
Nokia	Recently Nokia realized the need to provide robust network coverage with reduced cost. It also delivered: End-to-End network slicing functionality The fastest 5G speeds were achieved using the over-the-air (OTA) network.
NEC corporation	Their target is to do modifications in the industry according to the changing needs of organizations: They worked on the automation of the construction company. They devised a facial recognition demo system.
Cisco Systems	some developments made by the company: The company introduced a 5G security architecture. Support 5G services, infrastructure and automation.
LG	LG builds products that utilize the 5G networks rather than only deploying 5G networks. Some of the research work actively done by LG are: For an application such as in-vehicle infotainment specially designed for connected cars, LG announced an agreement with chipmaker Qualcomm.
Ericsson	As the company is proficient in the domain of 5G, many other big famous companies joined with Ericsson. Some of the research activities done by the company are listed as follows: Ericsson did the first 5G demonstration. The company is involved in almost all areas to make 5G a global standard for next-generation wireless technology.
ZTE Corporation	ZTE and its series solution propose the Pre5G concept. Key contributions by the company are: 5G new Radio air interface protocol proposed by ZTE. The company worked on applications requiring a 60 MHz spectrum, such as gigabit Ethernet without fibre connectivity, cloud XR, Autonomous driving and remote surgery.
Verizon	For 5G deployment to the world, it provides smart policies. In the 5G race, some of the contributions made by Verizon are: In four cities in late 2018, the company launched its 5G broadband internet.
Orange	The company contributes not only to smartphones but also to other content such as: Automated refrigerators, cars and augmented reality.

6. 5G-enabled IoT architecture: Security attacks and countermeasures

The deployment of IoT applications is difficult due to heterogeneous environments and resource-limited devices. IoT applications need to address major concerns such as security and privacy at the network and device level [2]. Some of the applications and their corresponding threats are discussed as follows:

6.1. Smart homes

A smart home embodies a technologically upgraded living environment, aiming to provide inhabitants with better standards of life. In futuristic 5G-enabled IoT communication network it is expected to cover billions of heterogeneous devices interacting with each other [37]. In an intelligent home environment, all connected devices increase malicious attacks [38]. The types of possible attacks and their possible countermeasures proposed by various authors are shown in Table 6. The Smart home attacks can be classified under two main categories:

Passive Attacks-It includes an eavesdropping attack without the consent of the communicating parties. The unauthorized interception of ongoing communication and traffic analysis are passive attacks. Hence, deduce useful information from the adversary by monitoring data traffic patterns. Such attacks do not modify data but learn useful information.

Table 6
Security attacks and possible countermeasures in smart homes.

Security attacks	Countermeasures	Refs.
Intrusion detection	Primary and Secondary access points in a home using different sensors	[39]
Replay attack	Symmetric key cryptography	[40]
User Anonymity	Three-factor authentication scheme: XOR and hash function	[41]
Distributed Denial of Service (DDoS) attack	Blockchain-based smart home framework	[42]

Table 7
Smart agriculture possible attacks and countermeasures.

Security attacks	Countermeasures	Refs.
Replay	Feature-based biometric, Timestamp, pairing-based cryptography, hash functions.	[42]
Masquerade	Physiological-based biometric, hashing functions, ECC, and pairing- cryptography.	[44,45]
Tracing	Random numbers in commitments.	[46]
Man-in-the-middle	Homomorphic encryption and hash functions: data aggregation schemes	[47]

Active attacks- Such attacks modify user data. It includes masquerading attacks when an adversary pretends to be an authorized entity to gain special privileges. It also includes an attack such as replay in which an unauthorized person retransmits an originally captured message to produce an unauthorized effect.

6.2. Smart farming

The latest progress in IoT technologies powered by 5G has transformed the strategies employed in smart agriculture. For example, IoT privacy-preserving data aggregation is essential nowadays to protect farmers' privacy and information availability [43]. Additionally, the agriculture sector, due to heterogeneous, internet-connected devices enabled by 5G IoT, has been exposed to potential cyber-attacks categorized as:

- **Data attacks**: include false data injection, insider and cloud data leakage and misinformation attacks.
- **Networking and equipment attacks**:side-channel, botnet, malware injection, DoS and radio frequency jamming.
- **Supply chain Attacks**: includes third-party attacks and data fabrication attacks.

Many authors proposed different types of countermeasures for the attacks smart agriculture sector, as represented in Table 7

6.3. E-healthcare

It is expected that the 5G network will overcome the limitations of previous networks by offering ultra-low latency, high density, high bandwidth, high reliability, and high energy efficiency [26]. Due to advancements in technology, it supports E-healthcare applications [48]. Further, down the line, wearable devices will facilitate a conversation with doctors based on their data to alert those in the wider population if any health anomalies are detected [49]. 5G-enabled IoT supports real-time monitoring systems, necessary alerts and connectivity. Such devices help healthcare professionals be informed and ready during emergencies such as COVID-19, which can ultimately save many lives. Due to wireless connectivity, these systems are susceptible to security risks posed by unauthorized individuals. Descriptions of certain risks affecting e-healthcare systems include [45]:

- **Denial of Service-** The attacker tries to access secret patient data without authentication or permission.
- **Fingerprint and Timing-based Snooping** - In this scenario, the intruder tries to access information during data transient between sensors and from the sensor to a private user location. An intruder can disrupt a patient's health conditions by accessing such information.
- **Router Attack**- As routing allows data to be delivered remotely and encourages network versatility, e-healthcare systems require data to be delivered securely and protect user identity. Extensive focus has been directed towards research in this domain, leading to the development of multiple defence strategies against diverse healthcare sector threats. Several of these countermeasures are outlined in the Table 8.

6.4. Industrial IoT (IIoT)

Industry 4.0, IoT will offer unceasing connectivity and standard communication protocol solutions to existing industrial systems [58]. Therefore, it provides promising transformation to existing IIoT problems. IIoT is applied in automobile manufacturing, engineering machinery, refrigeration equipment, metal smelting, etc. The IIoT creates a strong requirement to safeguard important industrial applications from cybersecurity attacks. Some of the possible attacks in IIoT are [55]: Masquerade, Virus, Trojan Horse, Worms, Cinderella, and Fragmentation. Table 9 shows some existing countermeasures for security threats in the IIoT framework.

Table 8
E-Health care possible attacks and countermeasures.

Attacks	Countermeasures	Refs.
Cross-Site Scripting (XSS) attack	Flirting method to prevent XSS attack.	[50,51]
	Machine learning-based XSS detection system.	[50,52]
SQL Injection Attack	Filtering all user input, data sanitization and ignoring creating SQL queries with user data.	[50,53]
	Tokenizing-comparing model	[54]
DDoS attack	Healthcare environment - Traceback technique	[55]
Path-based DoS Attack	Anti-Reply protection and packet authentication	[56]
Data access attack	Game-theoretic model	[57]

Table 9
Attacks on IIoT and possible countermeasures.

Cyber threats	Countermeasures	Refs.
Phishing attacks	Early analysis of phishing attacks and PHONEY for auto-detection.	[59]
	Intelligence Web Application Firewall (IWAF).	[60]
	Botnets Detection	[61]
Ransomware attacks	Futuristic firewalls - ameliorate traffic filtering capabilities.	[62]
	proposed application-specific machine learning algorithms.	[63]
	Intrusion detection system	[64]
Jamming DoS attacks	Data diversion to alternative routes	[65]
Collision attacks FHSS	The frequency-hopping spread spectrum (FHSS) is utilized to alleviate the interference.	[66]
Data transfer attacks	Data encryption techniques Datagram transport compressed protocol.	[67]
Supply Chain Attacks	Artificial intelligence (AI) and real-time intelligence, self-adapting supply chain systems using machine Learning (ML) for predictive cyber risk analytics.	[68] [69]

Table 10
Attacks on smart transportation and countermeasures.

Security attacks	Countermeasures	Refs.
Sybil attack	Cryptographic solutions	[70]
Falsified entities attack	Authentication	[71]
Replication attack	Key Management	[72]
Jamming attack	Anti-Jamming Techniques	[70]
Wormhole attack	Restriction on Packet transmission distance	[73]
Jamming attack	Detection of malicious component	[70]

6.5. Smart transportation

6.5.1. Vehicle telematics

The field of Automotive and Transportation (A and T) already encompasses vehicle diagnostics, location tracking, and telematics applications. These applications support non-real-time data. However, real-time information about the situation and performance of the vehicle could easily be collected using 5G. Due to its enhanced speed and low latency, future applications can collect driver behaviour and validate the delivery of more state-of-the-art services.

6.5.2. Vehicle infotainment

It is one of the futuristic applications, which boosts using 5G. Some examples of the services offered by vehicle infotainment are in-car retail and marketing, AR/VR-based navigation systems and entertainment services. It also enables the vehicle to synchronize with the user's autopay service, which can be used for external payment-related services such as fuel payment. In the case of autonomous vehicles such as V2X communication, it is complex to ensure targeted functionality. Nonetheless, as a result of diverse connectivity, these applications could also be susceptible to the security risks listed in Table 10 as:

7. 5G-enabled IoT: System vulnerabilities

IoT technology offers users convenience and efficiency to achieve a better quality of life. It has become indispensable in developing intelligent homes, smart industries, and smart agriculture. However, associating objects with the internet raises future security concerns. The further discussion consists of various security attacks on 5G-enabled IoT and possible traditional countermeasures.

7.1. Security attacks

The future Internet of Things (IoT) idea is that the Internet's global, dynamic living structure would be available, sensed, and interconnected. WSN play the foremost role in IoT communication. The sensor node transmits messages to the user device via the gateway node (GWN). The data transmission is done between these communicating entities through an insecure network. The messages exchanged using insecure communication channels can be intercepted, modified, or re-routed by an attacker/adversary [74]. The Security Consideration(SC) are discussed to analyse the security of authentication schemes [24]:

7.1.1. Safe against impersonation (SC1)

Considering a scenario, adversaries share user information with the gateway. The adversary tries to ensure GWN that the message is coming from an authenticated user by forging user credentials. This attack is also caused by flooding the system by spoofing. However, it can be detected by acknowledging each request and maintaining its sequence number [25].

7.1.2. Gateway node bypassing (SC2)

This attack considers a scenario when an attacker directly accesses data shared by sensor nodes without participating in the gateway node. Authentication of gateway nodes is required to maintain data integrity [7].

7.1.3. Privacy sniffing (SC3)

In 5G-enabled IoT communication, it is considered a potential attack in this an attacker to launch further attacks. The eavesdropper exchanged messages between communicating parties, sensor and sniffed email traffic, router configuration, and chat sessions [22].

7.1.4. Replay attack (SC4)

In this type, the attacker seizes the data shared between communicating entities. The attacker then repeats the message to extract secret information. Additionally, an attacker can deceitfully delay or retransmit it to confound the receiving entity evaluating. To prevent such an attack, in the 5G IoT scenario, time passwords, nonce, time-stamp, and Media Access Control (MAC) are to be created for mutual authentication.

7.1.5. Gateway forgery (SC5)

An attacker tries to acquire transmitted messages from GWN. The authenticated gateway node is required to prevent such an attack [7].

7.1.6. Location spoofing (SC6)

Such an attack occurs in the communication environment when devices are communicating with each other. An attacker may broadcast a request message with the wrong location information to disrupt communication. Therefore, device validation is required for messages' authenticity [24].

7.1.7. Man-in middle (SC7)

In this attacker intercepts network data to gain unauthorized access. The attacker expropriates the transmitted messages, making users believe they are talking directly but infecting the conversation.

7.1.8. Denial of service (SC8)

DoS can be explained by considering a scenario where an attacker floods the target with traffic to capture and completely deplete memory resources. Additionally, it disallows legitimate users' access to resources due to this enormous traffic [25].

7.1.9. Privileged-insider (SC9)

An attacker may be a privileged insider or trusted authority who can exploit stored information to unauthorized parties. In addition, session key computation and password guessing attacks [22].

7.1.10. Traffic analysis (SC10)

This attacker listens to network communication to determine application behaviour patterns, routing structure, location of key nodes, and base station location [22].

7.1.11. Mutual authentication (SC11)

The entities must register themselves before any communication process starts. The authentication ensures that two authenticated parties communicate with each other, reducing the probability of non-repudiation and user impersonation attacks. Authentication between the communicating entities must provide safe communication.

8. Classical security countermeasures

Table 11 shows several classical security solutions explored by researchers, with detailed descriptions provided in the following subsections:

Table 11
Techniques adapted by existing authors in the present IoT scenario and their weaknesses.

Author and year	Security threats	Techniques	Advantages	Weaknesses
Lu et al. [78] 2019	-Impersonation -Privacy Sensor node sniffing	Elliptic Curve Cryptography (ECC)	Less memory and computation cost	Data integrity and authentication
Wu et al. [79] 2016	-Gateway node bypassing attack -Replay attack	- Hash Function - XOR computations	- Resilient against the secure key agreement, authentication, credential updation, protection and user anonymity.	Network attacks are possible.
Althobaiti et al. [30] 2013	Attacks: - Eavesdropping -Insider -Gateway forgery attack	-Biometric encryption - Hash Function	- Suitable for both - Homogeneous and heterogeneous environments.	Mutual authentication and confidentiality of data.
Gope et al. [80] 2018	-Location spoofing	-Lightweight RFID-based authentication.	- Easy tracking of real-time objects. - Low cost	Impersonation attack. Limited computing and storage capabilities of RFID Tags.
Deng et al. [81] 2014	-Traffic analysis	- Anti-traffic analysis strategy - Differential Fractal Propagation (DFP) Algorithm -Multi-parent routing scheme -Random walk	-Low overhead and energy consumption.	Time correlation attacks.

8.1. Public key authentication

Asymmetric-cryptosystem relies on the mathematical formulation of problems based on the prime number and discrete logarithm. In this process, secret messages are exchanged between sender and receiver through the internet [75]. The only people who can get meaningful information from the message are the authenticated users with private keys. Moreover, authentication protocols rely heavily on finding the private key from its paired public key. The computing capabilities of private and public key pairs should be kept computationally efficient, especially for deploying wireless IoT [76]. These schemes require further analysis to provide lightweight and computationally logical privacy preservation.

8.2. Private key authentication

In authentication, protocols based on symmetric cryptography, communicating parties share a common secret key for communication. In this method, both entities, i.e. sender and receiver, have the private key for secretly communicating. For instance, the AES, DES and their variants Double DES and Triple DES provide secure communication between sender and receiver [77]. Avoine et al. [27] proposed a three-party secure key interchange protocol based on the private key concept. Their protocol considers a realistic IoT deployment that involves numerous end devices and servers.

8.3. Hashing

The hashing technique takes variable-length messages as input and produces fixed-length code by applying mathematical operations. Due to the massive D2D, H2M and M2M connectivity, user passwords are prone to security attacks. Therefore, hash-based schemes are required to protect passwords. The key agreement's digital signature method and random sequence generators are a few examples of hash-authentication-based cryptographic applications [14]. Turkanovic et al. [28] proposed a mutual authentication scheme, which was based on establishing a shared cryptographic key between the sensor node and the user outside the network. These authors, in their scheme, use only simple hash and XOR computations and adapt to the resource-constrained architecture of the WSN.

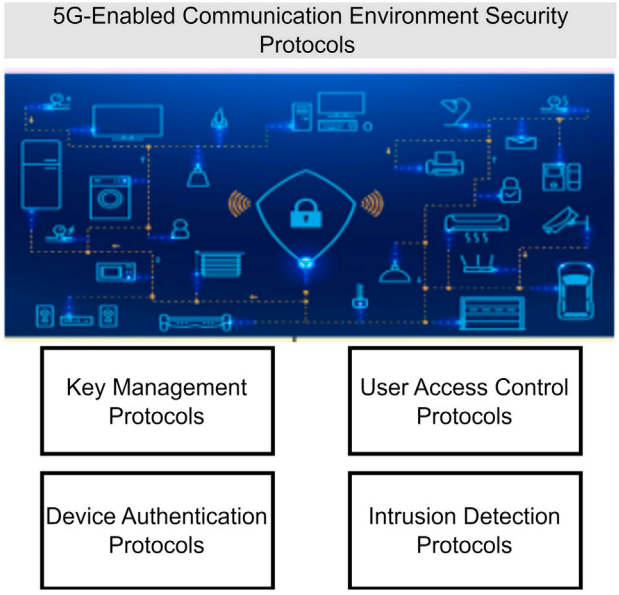


Fig. 4. Significant development in Quantum computing theories [A-Inventor; B-Significant contribution.

9. 5G-enabled IoT: Authentication and key management

In IoT communication, the user information will be encrypted using a private key the most popular methods are based on prime factorization and ECC. However, the 5G-enabled IoT framework suffers from various attacks. It is a complex task to secure a secret key and verify authorized users. Therefore, key management and authentication are the major research issues. As illustrated in Fig. 4, critical concerns such as key management, user access control, device authentication, and intrusion detection must be addressed to ensure secure communication within IoT [22]. Some authors proposed security protocols for different categories of key management issues such as "device authentication" and "User authentication" [15,46]. They also consider"intrusion detection" and "user access control".

The cyber-attacks identified in 5G IoT-enabled communication greatly impacted user privacy, identity authentication, and consequential security risks [82] to its powered devices. In an IoT environment, a key management scheme uses a cryptographic approach that keeps the records of trusted users and different smart devices involved.

Table 12 shows the protocols for key management and authentication examined by various researchers.

Moreover, the key management protocol generates, distributes, establishes and manages cryptographic keys to differentiate between malicious and legitimate entities [22]. 5G-enabled IoT attacks are not limited to domain name system (DNS) attacks [1], password guessing and cracking. Future attacks are more descriptive and vulnerable, which can be categorized based on three cryptological parameters:

(a) **Mutual Authentication** -In a 5G-enabled IoT, authentication is highly important as it involves both parties authenticating each other. Authentication between smart IoT devices is important to protect the network from Sybil, impersonation, identity-based attacks, etc.
(b) **Confidentiality** - Cyber-attacks are flourishing in digital communication; hence, hiding information from attackers like user private data and security keys is required. In the realm of IoT, 5G facilitates communication, but it also exposes vulnerabilities to passive attacks like packet sniffing and phishing, which can compromise user's sensitive data.
(c) **Integrity**- In a communication network, adversaries could access, monitor, and access authorized users' service requests. Additionally, this unauthorized person could also capture user credentials and user equipment. Integrity assures real and accurate data. It ensures user received data is not modified, deleted or illegally injected. An attacker may attack integrity by introducing session hijacking attacks, Man–In-Middle, etc. [86].

9.1. Classical mutual authentication schemes

In IoT communication, WSN plays a major role. The sensor nodes sense the data from the surroundings, which is beneficial for, decision-making. However, users must be authenticated to access data [87]. Many researchers proposed user authentication schemes for secure IoT communication. These schemes are one [12,14,15], two [54–57] and three-factor [59,67] authentication schemes. We discuss two and three-factor schemes in this section. As one parameter is based on knowledge factors like passwords, it is not considered secure due to heterogeneous and resource-constrained 5G-ad hoc wireless sensor networks.

Table 12
Existing key agreement and authentication protocols for 5G-enabled IoT.

Author	Authentication protocol	Resilience against attacks	Key areas covered	ASVT	Privacy	Protection	
					A	I	C
Seok et al. [25]	Lightweight Authentication Protocol With associated data (AEAD) Ciphers	Impersonation, Eavesdropping, Privacy,sensor node sniffing.	Covered Resource-Constrained IoT devices.	–	✓	✓	✓
Sharma et al. [83]	Authentication framework assisting user privacy-preserving and key authentication.	Attacks on confidentiality, Integrity and Availability	5G threats due to public access connectivity.	Scyther	✓	✓	✓
Basin et al. [84]	5G AKA protocol.	Location attacks, replay attacks	Mutually authenticating subscribers and their carriers and establishing a secure channel to protect subsequent communication.	Tamarin	✓	✓	✓
Gong et al. [85]	Mobile edge computing architecture.	Denial of Service(DoS)	-Secondary and data management authentication functions	–	✓	×	×
Chaudhary et al. [33]	Authenticated mobile devices Kerberos in the 5G communication framework.	Distributed Denial of Service (DDoS).	Authentication considers kerberos as a solution for attacks such as DDoS.	Python	✓	×	×

A-Authentication, I-Integrity, C-Confidentiality, ASVT-Automated Security Verification Tool.

Table 13
Security considerations by researchers based on two-factor authentication.

Authors	SC1	SC2	SC3	SC4	SC5	SC6	SC7	SC8	SC9	SC10	SC11
Watro et al. [86]	–	–	–	✓	–	–	–	–	–	–	–
Das et al. [88]	–	–	✓	✓	–	–	–	✓	–	–	–
Huang et al. [89]	✓	–	✓	–	–	–	–	–	✓	–	–
He et al. [90]	✓	–	–	–	–	–	–	–	✓	–	✓
Jiang et al. [74]	✓	–	✓	–	–	–	–	–	✓	–	✓
Chen et al. [91]	–	–	✓	✓	–	–	–	–	–	–	–
Vaidya et al. [92]	✓	–	✓	–	–	–	–	–	✓	–	✓
Kaur et al. [93]	✓	✓	✓	✓	–	–	–	–	✓	–	✓
Bakkiam et al. [94]	✓	–	✓	✓	✓	–	–	–	✓	–	✓
Khan et al. [89]	✓	–	✓	✓	✓	✓	–	–	✓	–	–

9.1.1. Two-parameter authentication scheme

The 2PAS uses the two parameters for authentication, such as entity password, biometric data or an E-Smart card. Watro et al. [85] use two parameters in their scheme password and smart card for authentication. The cryptography algorithms used in their work are RSA and Diffie Hellman. They worked on a lightweight sensor node to implement a public-key-based protocol called TinyPk. The weakness of their scheme is that they do not safeguard against user masquerading attacks, as reported by Das et al. [86]. A two-factor authentication protocol [56] provided a secure network communication considering stolen-verifier attacks based on the hash function. Table 13 outlines the security aspects addressed by various authors when employing a two-factor authentication scheme.

9.1.2. Three-parameter authentication scheme

The Two-factor communication schemes are secured until and unless passwords or smart cards are stolen [7]. Table 14 shows the security breaches examined by various authors in their respective research contributions.

Whilst 3PAS strengthens the security more than 2PAS. The added third parameter, biometric identification, makes it difficult for an eavesdropper to tamper the security. Chang et al. [95] proposed an authentication scheme based on key agreement. They aim to resolve weaknesses of 2PAS, such as lost E-Smart cards and impersonation attacks. However, later Jung et al. [96] reported that [95] it does not safeguard against impersonation and credential guessing attacks. Chang et al. [91] analysed that [28] is vulnerable to user impersonation, node capture and spoofing attacks.

Table 14
Security consideration of Three-Factor authentication schemes.

Authors	SC1	SC2	SC3	SC4	SC5	SC6	SC7	SC8	SC9	SC10	SC11
Das et al. [88]	✓	–	–	–	–	–	✓	–	–	–	✓
Althobaiti et al. [30]	–	–	✓	✓	–	–	✓	✓	–	–	✓
Lu et al. [78]	✓	–	–	✓	–	–	–	–	✓	–	✓
Om et al. [89]	✓	–	✓	✓	✓	–	–	–	✓	–	✓
Mishra et al. [95]	✓	–	✓	✓	✓	–	–	–	✓	–	✓
Chang et al. [91]	✓	–	–	–	–	–	–	–	–	–	✓
Jung et al. [96]	✓	✓	–	✓	–	–	–	–	✓	–	✓
Park et al. [97]	✓	–	–	–	–	–	✓	–	✓	–	✓
Le et al. [91]	✓	–	–	–	–	–	✓	–	–	–	✓
Turkanović et al. [88]	✓	✓	–	✓	–	–	–	–	✓	–	✓

✓Provided countermeasure against Security consideration; - Not covering the security consideration.

9.2. Security challenges of classical authentication schemes

The authentication schemes discussed above proposed by many researchers are based on two and three parameters. The recent practical implementation of these schemes proved that these are not secure under many classical and futuristic Quantum attacks. In a 5G-IoT environment, some recent and previously unresolved attacks motivate us to develop a four-step authentication model. The main loopholes and challenges may be summarized as follows:

9.2.1. Something you know: Passwords or pin

No security is infallible. The user credentials that provide two-factor authentication are saved in an E-smart card. Although the 2FAS scheme is more successful and powerful than a single factor, it is still inefficient. An adversary can easily get a user's personal information through fake websites by guessing passwords. In addition, they could also steal a user's phone from which they can access SMS and email information. The intruder gains effortless access to the text-based security code transmitted from the secure gateway to the user's mobile device.

9.2.2. Something you have: security token or smart card

For security, smartcards have been widely used for authentication purposes. The smart card is a proximity tag. These are used for many purposes, such as building access, identity confirmation, etc. Each card has a specific ID that differentiates it from others on the network. This identity is further passed on to the reader. Whenever any user accesses the system, this unique ID confirms their identity. Based on which their access is granted or denied. However, there are huge benefits of smart cards, but the overall estimated cost is between 75 and 100 dollars for implementing these E-smart cards and their subordinate devices. Suppose a company has thousands of customers or employees. For this, the implementation cost of smart cards could reach millions. Due to their small size, these smart cards are lost often and can be used by an unauthorized person.

9.2.3. Something you are: Biometrics

The authentication schemes based on biometrics provide more cost-effective solutions than OTP or token-based schemes. In biometrics, physical characteristics such as fingerprints, retinal data, or even facial features provide authentication if performed under controlled circumstances (e.g. guarded). However, it has been analysed that these biometric identity-based features are difficult to spoof. However, it is not feasible as this data may easily be replicated. The entire biometric information will become meaningless if there are security breaches on either the server side or the client side. However, considering the storage of biometric data could also be difficult as it raises the security implications of a data leak. The solution to the abovementioned problem is called cancellable biometrics to preserve user privacy. Cancellable biometrics is implemented using (1) Biometric Salting and (2) Noninvertible Transforms [97]. However, this solution has certain limitations in that it will only work under certain situations, such as when the client biometric data exact copy is available to the server.

9.2.4. Implementing authentication schemes: Time-consuming

Time may be a perfect metric, but it is also relative. The time lost in implementing 2PAS depends upon what form an organization adopts. For example, SMS-based 4-digit privacy codes are faster to navigate than others. The authentication schemes under 2FA increase the time to access accounts. It does that by adding a new factor to the authentication. Now, if we are considering each entity level, this would be minuscule. However, organizations with numerous employees add up to thousands of work hours lost each year in implementing this scheme.

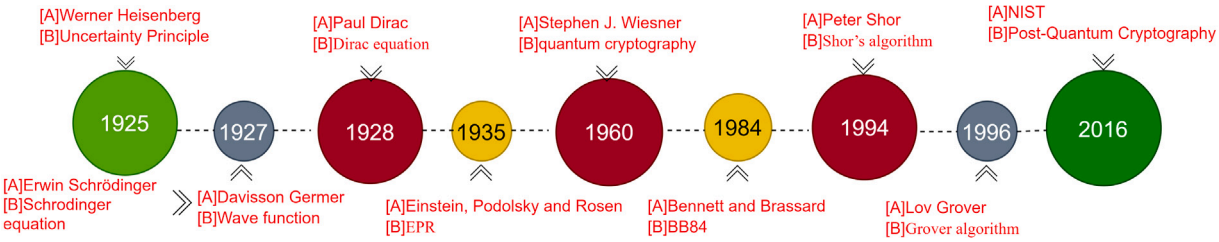


Fig. 5. Significant development in Quantum computing theories [A-Inventor; B-Significant contribution.

9.2.5. Error tolerance and nontrusted devices

The biometric data collection process is prone to various noises. Due to noise, it is impossible to reproduce biometric information as measured precisely. One issue with practical biometric authentication protocol is the comparison of the encryption of biometric templates. These protocols required an exact match. Another issue in biometric authentication is that the server should perform only biometrics verification instead of verifying any user devices. Such devices cannot be fully trusted due to their distance from the server.

9.2.6. Heterogeneity: WSNs

Heterogeneity is a congenital feature of mobile computing. Mobile devices, wireless networks and the cloud are the three different categories of mobile computing. Heterogeneity preserves the connectivity of mobile devices. However, in mobile computing, the main security concerns are during handover. So, while designing authentication schemes, the diversity of framework, technologies, hardware, infrastructure, and handover mechanisms must be considered.

9.2.7. Failure can be disruptive

In authentication schemes, the more factors you add, the more successfully we can prevent unauthorized access and increase complexity. However, these factors add more complexity to the systems but may also create hurdles for an attacker. The attacker would have to jump through many factors to gain access. Moreover, integrating IoT with different objects creates various issues in people's daily lives, and delivering robust security to the IoT is challenging during transmission. Even the lightweight and privacy-preserving two or three-factor algorithm was provided. Still, the major problems and the possible attacks such as node capture, routing threats, impersonation attacks, brute-force attacks, message-tampering attacks and replay attacks are possible. Therefore, more security is required for secure network communication for future IoT environments.

From the above discussion we identified, that there is an urgency to develop security protocols to safeguard the 5G-enabled IoT environment. Heterogeneous IoT devices are vulnerable to various classical and Quantum attacks. Therefore, in the next sections, we reviewed the Quantum-resistant solutions to safeguard IoT communication.

10. Assessment of quantum computing

The Internet of Things (IoT) applications include intelligent cities, IoT-based green agriculture, E-healthcare, UAV, and many more. The advancement of these applications connects many IoT devices by using the internet. However, Quantum algorithms jeopardized the security of classical security protocols [71] based on mathematical structures. An emerging paradigm of Quantum information and computing has created an unsafe environment for the current security algorithms, which are extensively utilized to secure the IoT communication framework. Therefore, Quantum-resistant techniques are required to provide security to current and futuristic IoT devices. These techniques must be implemented to solve security breaches.

10.1. Quantum computing: History of decades

Quantum information and computing were introduced by the famous American theoretical physicist Feynman [16] in the 1980s. However, the field gained attention after the article 'Simulating Physics with Computers [72]. Even before that, in the late 1960s, Stephen J. Wiesner [73] started work on Quantum cryptosystems. Later, Charles H. Bennet and Gilles Brassard [98] did the practical implementation of Quantum computing-based protocols in 1984. They examined the principles of quantum physics and information theory, leading to the creation of Quantum Computing. They proposed their first protocol based on Quantum mechanics known as BB84 [99]. Fig. 5 represents remarkable progress in Quantum computing and information. It represents the significant work done by researchers in Quantum computing, starting from the famous Heisenberg uncertainty principle towards the Post-Quantum cryptography era.

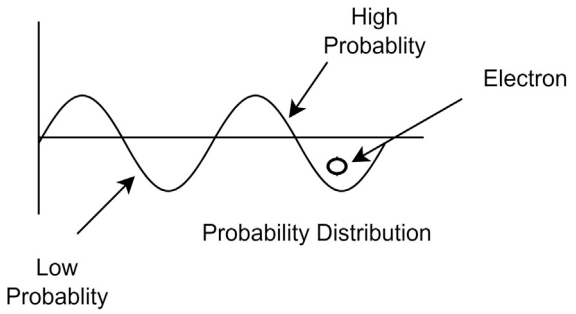


Fig. 6. Particle-Wave Duality.

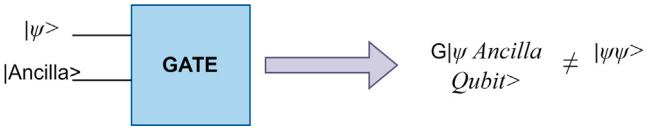


Fig. 7. Copying arbitrary Quantum State using ancilla qubit.

10.2. Quantum mechanics fundamentals

10.2.1. Wave function

The wave function is mathematically described as Ψ The wave defines the characteristics of a particle. According to the principle of Particle-Wave Duality, finding the particle described by a specific wave function has a certain probability at a given time. This probability is proportional to the value of ψ^2. Electrons move somewhere in the waveform area, which we do not know exactly about particles. However, what we know is only about probabilities. Certain mathematical operations are then applied to such a form of a particle, due to which even if an adversary tries to copy data, it becomes impossible. The experiment that confirms the particle Wave duality is the Double-slit experiment [100]. Fig. 6 represents the probability distribution of electrons. The high and low probability represents the value of the wave function of a particle at a given point of space and time is related to the likelihood of the particle's being there at the time.

10.2.2. Photon polarization

The Quantum system is based on the elementary particle of light referred to as a photon. These particles hold a constant amount of energy. Polarizing filter light may be polarized in a specific direction or fixed to the desired angle. Polarization is a Quantum physical property that comes out when light is considered an electromagnetic wave [101]. Polarized photons range from 0° to 180° [102].

10.2.3. Quantum no-cloning theory

It is impossible to make a perfect copy of the Quantum state. The concept was shared in 1982 by Wootters, Zurek, and Dieks [102]. It states [101] that the no-cloning principle does not allow the generation of identical Quantum states. The Quantum system is a composite system generated from different Quantum states. Such a system is manipulated through observation, collapsing the information in qubits. Another possibility is to control the Hamiltonian of the system [102]. The Quantum system's state is disturbed if an observation allows anyone to copy the state of the Quantum system. The eavesdropper cannot copy these qubits due to the no-cloning theorem. The No-cloning theorem can be demonstrated by involving ancilla qubits in the context of the Quantum gate, as depicted in Fig. 7. The results reflect that the ancilla qubit (duplicate state) along with the input qubit when processed through the copy gate do not produce the two copies of the actual state ψ.

10.2.4. Superposition

The multiple qubit state $|0>$ or $|1>$ is simultaneously represented using the superposition. Quantum information computing operations are performed on qubits. In Fig. 8 geometrical representation of the pure state of a two-level quantum system is represented. The quantum bits on the Bloch sphere [23] are represented as two orthogonal states.

The principle of coherent Quantum superposition states increased to fine-dimensional Hilbert Space [103]. Such phenomenon in Quantum computing is referred to as superposition, symbolized in the form as $|\phi>= \alpha|0> +\beta|1>$. Where α and β are complex numbers representing the probability of qubit to be in state $|0>$ or $|1>$. Additionally, to satisfy the normalization conditions, the total sum of these probabilities must equal one. Therefore, to satisfy the probability rule, we get that: $\|\alpha\|^2 + \|\beta\|^2 = 1$

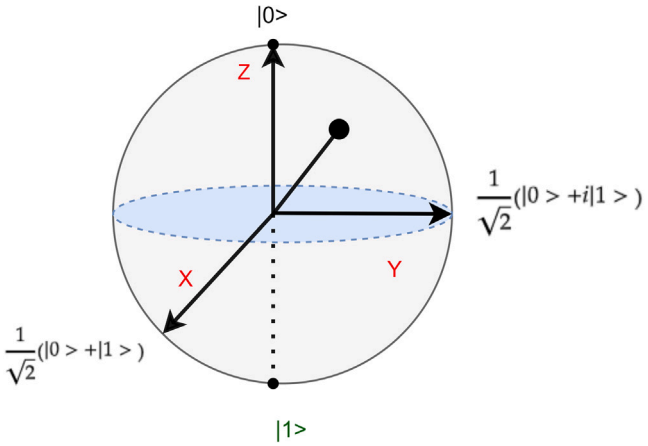


Fig. 8. Quantum state representation on Bloch Sphere.

10.2.5. Heisenberg uncertainty principle

The main logic behind this is a wave function that contains all of the positions of the moment of the electron. The uncertainty principle says it is impossible to simultaneously find the absolute measurement of a particle's position (x) and the momentum (p). It will give us probability distribution and electron always exists in different places with different probabilities for a specific wave. The probabilities create uncertainty in the Quantum system. Due to uncertainty, measuring a particle's position and momentum is impossible. Measurement creates a disturbance in the Quantum system. Therefore, the uncertainty principle is useful for Quantum Computing. It is useful for protecting user data and protecting communication networks from eavesdropping attacks [23]. Furthermore, multiplying the errors (represented as the delta in front of X and P) in the observation of these points has to give a value greater than or equal to half of a constant called " h-bar" as in Eq. (1):

$$\Delta X \Delta P \geq \frac{\hbar}{4\pi} \tag{1}$$

X is the position operator, P is the photon's momentum, and $\hbar$ is the plank's constant.

10.2.6. Quantum entanglement

It is a property that makes detecting intruders in network communication easy. When the simple independent probability multiplication rule does not work in identifying the exact probability value that means the particles are not independent. They are entangled. The changes in one state will affect the other state [104]. The local measurement will not work in entanglement. The only way to compute these bits is by joint operation. Two Quantum systems interact in such a way as to link both their spatial coordinates in a certain direction and their linear momenta (in the same direction). The phenomenon of entanglement creates the possibility of connecting particles widely separated in spaces. Entanglement is the teleportation feature that happens in composite systems and is absent in single-state systems [105]. The main concept of teleportation [98] is establishing a secure communication channel between two parties even without Quantum communication channels [106]. By utilizing Quantum teleportation, these entangled bits are shared among communicating parties. Entanglement is also referred to as Bell-states. Four mutually orthonormal Bell states are expressed as in Eq. (2):

$$|\Phi>^{\pm} = |00> \pm |11> \qquad |\psi>^{\pm} = |01> \pm |10> \tag{2}$$

10.3. Impact of quantum computing on classical cryptography

Although Quantum computing has enormous potential, the field is still in its nascent age. IBM's practical implementation of Quantum computers is up to 127 qubits [107], but its evolution is already threatening cybersecurity. Classical cryptosystems were based on symmetric and asymmetric algorithms for securing IoT communication. These cryptographically secured systems rely on the key distribution method. In the IoT framework, all key management and distribution are done over an insecure channel. Hence, it is difficult to identify authorized users. Quantum computing algorithms hugely impacted the present security structure of IoT communication systems. The security structure of IoT is based upon symmetric [108–110] and asymmetric [5,111] cryptosystems. Hence, there is an urgency to develop quantum-resistant cryptosystems. Two Quantum algorithms that will execute on Quantum computers [11] and are theoretically proven to break the security structure of classical algorithms are:

10.3.1. Shor's algorithm

The highly complex integer factorization problem was solved by a famous mathematician Peter Shor [99]. The author published his work in 1994. His theory is experimentally demonstrated in Bell Laboratories. Later he showed his contribution to solving the elliptic curve discrete logarithm problems.

10.3.2. Grover's algorithm

Complementing Shor's algorithm [99], the Grover search [34] algorithm affects the security of algorithms based on symmetric cryptography. It provides significant speedup for many problems such as optimization and factoring large numbers into a product of two prime numbers. Classically these types of problems are resolved using brute force search. Furthermore, large and small-scale startups such as IBM, Google, Rigetti, and D-Wave implemented large-scale Quantum computers [6]. Due to their availability, currently used public key infrastructures will become insecure. Recently researchers have been working on Quantum Fourier transform (QFT) [6] and the Quantum Walk (QW) [112] for solving complex searching problems. Some other research issues arise in the field of Quantum algorithms for optimization problems, such as adiabatic Quantum computing (AQC) [6]. Recently, various countries, such as the United States, Australia, Britain, Canada and New Zealand, used Secure Communication based on Quantum Cryptography (SECOQC) [104]. The SECOQC was proposed for a strategic defence against the Echelon intelligence gathering system [104]. Quantum computers are infant age, but their effect on classical cryptography based on symmetric (private-key) and asymmetric (or public key) cryptography will be very significant. The impact of Quantum algorithms such as Shor's and Grover's on symmetric and asymmetric cryptosystems is elaborated as follows:

(a) **Symmetric (Private-Key) -**AES is considered a secure algorithm for IoT-based communication. The algorithm is based on utilizing a single key for encryption and decryption. The National Institute of Standards and Technology (NIST) standardized it in 2001. However, the most widely used variant of AES is the AES-128 cryptosystem. However, AES allows applications to be more secure by using keys of 128, 196, and 256 bits. Brute force is the most recent attack on the security of AES cryptosystems. It executes by covering all possible keys. The complexity of the brute force attack is 2^n if the size of the key is n bits [16]. Therefore, until approximately 2030, by considering Non-Quantum compute availability, a minimum of 112–bit security is considered safe. The famous Grover [68] algorithm speeds up the process of brute force attack. Hence, we need 256 bits of the key to achieve the same level of security as the 128-bit key. The futuristic scenario is that a 128-bit key will offer roughly the same level of security as a 64-bit key today whenever the Quantum computer is available [6]. Another cryptographically secured symmetric key algorithm is DES [23]. The channel used for key exchange in this type of communication is insecure. This type of communication is vulnerable to passive eavesdropping. Therefore, it is impossible to achieve security with classical cryptographically secured algorithms.

(b) **Asymmetric (or public key) cryptosystems -** Quantum computing and algorithms hugely impacted classical public-key systems. The Diffie–Hellman (DH) key exchange, RSA and elliptic curve cryptosystems (ECC) algorithms are not secure with advancements in Quantum computing and information. The algorithms mentioned above are based on the complexity of finding solutions to problems such as integer factorization and discrete logarithms [6]. These systems operate under the One-Way Function (OWF). These functions had the property that they are easy to compute one way but difficult another way [16]. However, the Quantum Shor's algorithm [113] resolved the problem of integer factorization in polynomial time [101]. Therefore, as soon as Quantum-based Shor's algorithm is implemented, many cryptographically secure systems that rely on prime number factorization problems will become insecure. For example, Quantum Shor's [99] algorithm gave an approximately 80% probability [113] that the value of the divisor is a specific number, e.g. 213,432,237,905,197. In addition, the classical enumeration problem can also be solved using the Quantum Grover algorithm by quadratic speedup. Classical computers using classical algorithms, however, cannot solve this problem. For instance, recently attackers have manipulated and stolen millions of dollars from the German government to combat COVID-19 [16]. Therefore, Quantum-resistant communication channels are required to secure IoT applications to overcome the above-mentioned problems. Fig. 9 represents the classification of public and private key cryptographic algorithms [31]. These symmetric and asymmetric algorithms are classified based on key pairs and key factors. We have also mentioned the impact of Quantum computing algorithms on these cryptographic schemes.

Additionally, high security is the foremost vision of the 5G network. Whilst 5G networks exhibit multiple drawbacks, more work is required to overcome Quantum attacks. Therefore, core security features of 5G communication to ensure secure communication are [16]:

- The communication between the user and the network should be mutually authenticated.
- Detecting malicious codes by analysing traffic.
- Jamming unauthorized nodes by utilizing intrusion detection systems.
- Enhancing network-secured protocols by encrypting each slice after network slicing.
- Maintaining confidentiality of the user's location and privacy of user credentials.

By considering the limitations of classical cryptographic algorithms, Quantum-resistant authentication schemes gained huge attention [94]. These are named PQC schemes which are based on complex mathematical structures and QC based on the laws of Quantum physics.

11. Quantum–resistant authentication schemes for 5G-enabled IoT

Recent innovations in IoT towards a new digital era of communication provide a better lifestyle by embodying physical objects to the internet that give IoT eloquent importance. However, many implemented cryptographic solutions for securing IoT communication include symmetric and asymmetric key encryption, secure hash-based authentication methods and digital signature algorithms. Still, their recent advances provide threat and security challenges to existing IoT networks by breaking the cryptographic solutions. Therefore, Quantum-resistant solutions such as PQC and QC are required for securing futuristic IoT communication systems.

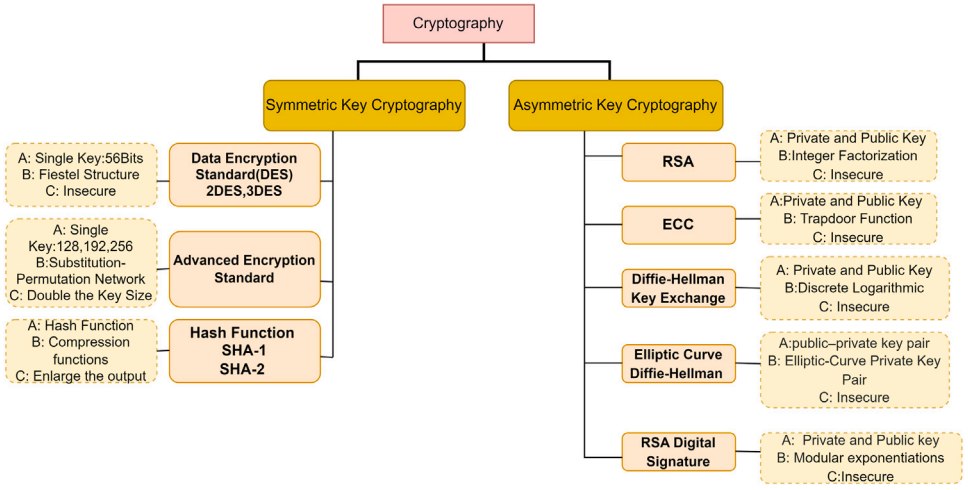


Fig. 9. (a) Quantum computing impact on symmetric and (b) asymmetric cryptographic systemsA: Key Pair; B: Key Factor, and C: Quantum Computers impact on these algorithms.

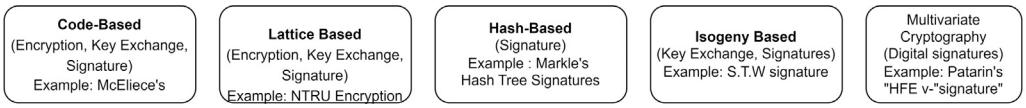


Fig. 10. An Overview of Post Quantum Family.

11.1. Post-quantum cryptography for IoT

Recent attacks on IoT security framework such as ransomware and DDoS has emerged from connected physical objects. The identity of these devices should be verified; therefore, the authentication framework can protect the identity of the devices and sensitive data during transmission. To avoid attacks such as Mirai [114], which brought many powerful DDoS attacks on existing authentication schemes. The authentication protocols must ensure the prevention of devices that are not legitimate from entering the system and becoming a part of the network. Therefore, changes are required in current cryptography protocols to ensure the security of 5G-enabled IoT communication. These changes ensure that Quantum computing does not affect the system's security [11]. According to NIST [115], Quantum-resistant solutions based on public-key cryptosystems are widely used. Post-Quantum-based security solutions (PQCS) are code-based cryptography, secure signatures based on a hash function, multivariate polynomial cryptosystems, and Lattice-based cryptography as depicted in Fig. 10. The in-depth discussion on the PQC family is presented in the next subsections.

11.1.1. Code-based cryptography

The PQC scheme based on code structure is a candidate for Quantum-resistant schemes. It was proposed in the 1970s. It provides solutions to secure current cryptography [116]. The public key-cryptosystem [117] was proposed in 1978. It was the first proposed cryptosystem based on code. The entire [6] scheme was based on assuming the authentic platform for communication. McEliece cryptosystem is binary Goppa codes considered as the base of this scheme. Their scheme also adds an error purposefully to safeguard messages against adversaries. The input messages are encrypted using binary Goppa [118]. However, the overall security depends on the syndrome-decoding problem. It says that decoding is performed without any knowledge of the coding scheme. The LDPC (Low-Density Parity-Check) codes and MDPC (Moderate-Density Parity-Check) codes are other variants of the McEliece scheme. In code-based cryptography, the following steps are to be considered for generating cypher-text:

(a) The input data is added with random errors.
(b) By analysing, forming a bit-error pattern by message encoding. The decoding recovers the original message by:
(c) Identifying and thereby eliminating mistakes from the input data.
(d) Removing the exact input message from the bit sequence of faulty codes.

It depends on the complexity of deciphering arbitrary linear codes, which positions it as a favourable option for safeguarding IoT devices from potential quantum threats. One important aspect of a code-based scheme is hiding the code structure is of utmost importance. Therefore, an attacker with access to the specific code used for encryption would decrypt the message easily.

Table 15
Lattice-based cryptographically secured authentication protocol.

Authors and Refs.	Authenticated protocols	Security issues	Verification tool
Cao et al. [15] and 2019	Lattice-Based Cryptography for NB-IoT	Signaling Congestion Avoidance Unlinkability, Protocol Attack Resistance	Scyther tool
Mitchell et al. [11] and 2020	5G-AKA	Mobile identity confidentiality,Session security	––
Boorghany et al. [121] and 2015	Authenticated key exchange (AKE) protocols	Confidentiality Authentication	JAVA, C++, ARM /THUMB compiler

11.1.2. Lattice-based cryptography

Lattice-based systems were formulated by Ajtai [16]. Lattice-based systems are formulated to solve problems such as the shortest vector in a high-dimensional lattice. A solution to these problems is computationally hard to find [10]. Göttert et al. [119] discussed and implemented in their work learning with errors (LWE) based cryptosystem. Lindner and Peikert propose these cryptosystems. Their scheme was based on matrix and polynomial-based variant comparisons of LWE. Boorghany et al. [120] in their work proposed lattice-based cryptographic authenticated key exchange (AKE) protocols. Furthermore, the provably secure lattice-based cryptography is better than NTRU in running time. The discrete Gaussian sampling and FFT are secure lattice-based cryptographic systems [28]. Cao et al. [15] discussed a group of NB-IoT devices. These devices are for Quantum resistance access authentication and data distribution schemes. Their scheme considered lattice-based homomorphic cryptographic technology. Mitchell et al. [11] discussed the impact of future Quantum information computing on 5G-enabled mobile security. They also proposed a 5G-AKA protocol to overcome the drawbacks of classical cryptographic algorithms. Hence, lattice-cryptographic mechanisms also provide fast, Quantum-resistant solutions. These solutions were earlier assumed to be impossible to solve [16]. Lattice-based cryptography relies on mathematical challenges associated with lattice structures, which are considered complex for both classical and quantum computers to address. It is considered a promising solution for securing IoT devices against the quantum threat. Some of the authenticated protocols implemented by different authors, security issues and tools used to verify their model are represented in Table 15

11.1.3. Hash-based signatures

These hash-based signatures are based on the one-time signature (OTS) scheme. The OTS scheme is based on a unique key pair [56]. The main flaw of this Quantum-resistant scheme is that two non-identical messages, such as a1 and a2, are signed using only one OTS key pair. In this case, by comparing these signed messages, the attacker can replicate this signature. An example is Merkle's hash tree based on the public-key signature system(1979) [122]. Lamport gave the idea of Merkle's signature system, and [123] is based upon a one-message signature. In this binary, hash trees are used. The leaf nodes of the binary tree represent the OTS public key hash values. However, the parent nodes are calculated by concatenating the hashes of their child nodes. The parent node's authentication of the public keys of OTS is accomplished using the collision-resistance hash function [116,121]. The hash-based signature scheme is in the process of standardization by the IETF [124], and it is referred to as the Extended Merkle Signature Scheme (XMSS) [118]. Hash-based signatures serve as a means to verify and protect communication integrity between IoT devices and gateways, guaranteeing that data shared among devices remains unaltered and resistant to interception.

11.1.4. Multivariate polynomial based cryptography

There is almost no way to solve random multivariate polynomial systems. Therefore, they are categorized as NP-hard. Their dependency is on the utilization of multivariate polynomial systems. These systems are used to protect data collected by IoT sensors. In 1996 the technique of Patarin's Hidden Fields [16] –Public key signature system (1996) became popular based on the multivariate scheme. The proposed scheme was the generalization of the approach discussed by Matsumoto and Imai [16,116,118]. Several multivariate cryptography schemes also exist based on Hidden Field Equations (HFE) trapdoor functions. These schemes are referred to as Unbalanced Oil and Vinegar encryption systems (UOV), Rainbow, and Tame Transformation Signature (TTS) [125]. The most popular signature scheme among these is HFEv due to its efficiency and ability to produce the shortest signatures among existing ones [6]. However, implementing a secure, robust and efficient multivariate cryptosystem scheme is an open challenge. Hence, many of those mentioned above multivariate public-key encryption schemes are not considered secure.

11.1.5. Supersingular elliptic curve isogeny cryptography

Based on classical ECC's concept, which works on defining points evaluated by computing addition and scalar multiplication operations. These points are defined on elliptic curves. Additionally, isogenies define operations between different elliptic curves [126]. Rostovtsev and Stolbunov [118] 2006 introduced public-key cryptosystems based on isogenies. However, more computation time required for encryption and decryption is the major drawback of this scheme. Later in 2010, Jao and Soukharev [127] identified

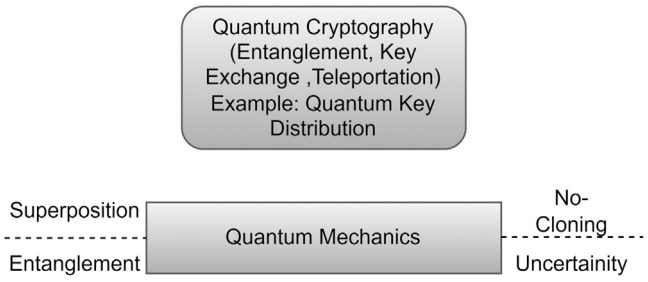


Fig. 11. Building blocks of Quantum Cryptography.

a sub-exponential Quantum computing attack on this system. Moreover, Isogeny-based schemes may serve as digital signatures or key exchanges [125], such as Supersingular Isogeny Diffie–Hellman (SIDH) and Supersingular Isogeny Key Encapsulation (SIKE) protocols [126]. These protocols are basic building blocks featuring different functionalities and levels of security. SIDH and SIKE protocols are to construct efficient and flexible authenticated key exchange schemes. Supersingular elliptic curve isogeny cryptography can facilitate secure key exchange among IoT devices, allowing them to create mutual cryptographic keys for ensuring encrypted communication.

11.2. Quantum cryptography

Post Quantum Cryptography (PQC) and Quantum Cryptography (QC) are two different worlds. PQC use hard mathematical techniques. The PQC techniques are unsolvable by Quantum computers. The development of Quantum computation and information relies on Quantum mechanics principles: Heisenberg and photon polarization principles [13]. QC was inspired by the limitations of the classical cryptography key distribution problem. Furthermore, Quantum cryptographic systems provide security over the internet on a large scale. As if anyone measures the Quantum channel, the state of the system collapses. The collapse prevents an experimenter from figuring out the Quantum state of a system. Many models are proposed for the Quantum-resistant world based on utilizing key ideas of QKD, such as Quantum coin tossing [103] and Quantum walks [4]. These provide a secure communication environment that can resist threats from Quantum computers. Fig. 11 represents the principles of Quantum mechanics around QC. The basic idea is to represent that QC is based on the fundamental laws of Quantum mechanics such as No-cloning and uncertainty so that no one other than the legitimate user can understand the actual message. Based on QC QKD, entanglement and teleportation provide unconditional security without imposing any restrictions on the abilities of an intruder.

Understanding the future needs of Quantum-secured solutions for securing IoT systems. Lohachab et al. [23] proposed a lightweight security authentication framework based on Quantum Key Distribution (QKD) and payload-based mutual authentication using ECC. Their scheme also considers how certain entities like the edge network layer, smart things network layer, middleware layer, physical objects, and cloud infrastructure layer interact using ECC and QC. Furthermore, the Quantum coin problem is discussed first by Blum [15], based on two mistrustful parties. These parties communicate at a distance without the help of a third party. They must agree on the probability of winning and losing the game of trust. In this type of communication, each party has approximately a 50% chance of winning. Both parties want to win the toss, which could let them cheat in various ways. Alice could attempt cheating at any step of communication. To cheat Einstein–Podolsky-Rosen(EPR) effect plays a major role. The coin directs the Quantum walk. It can go in both directions, like a wave spreading out in the Quantum world. This random process and randomness also make the walk difficult to control, making it more difficult to design systems. EL-Latif et al. [12] discussed efficient and highly Quantum walks. They proposed two authenticated hash functions, QWHF-1 and QWHF-2. These hash functions are based on QWs. These can either be continuous ($t \in R + \cup 0$) or discrete ($t \in NU0$) depending upon the way time is measured. [12] proposed Quantum hash functions centred on the discrete-time quantum walk model. Their focus is specifically on designing security mechanisms for 5G Networks. This coin and a walker are the elements of the coined discrete quantum walk. The authors also used the AQDC protocol and utilized Quantum hash functions. They also created a superposition of two linearly polarized states represented using the Hadamard operator (H) as [12] represented in: $|+>: \frac{1}{\sqrt{2}}(|0>+|1>)$. The confusion and diffusion test analysis were run to assess the usefulness of these protocols [128]. Compared to classical cryptosystems, their protocol provides maximum security, efficiency, and robustness against well-known attacks. However, high-level key security is important before exchanging data over the communication channel. Therefore, Arulet et al. [8] resolve limitations in the previous authentication [128] process by using QKD. They proposed Quantum Key GRID (QK-GRID). QK-GRID provides a secure LTE-based IoT framework for end-to-end communication. QKD defines two specific methods for single or entangled Quantum transmission. These two methods are Quantum Key-Server (QK-S) and Quantum Key GRID (QK-GRID) [8]. Table 16 represents the proposed Quantum authentication protocols by different researchers in the context of 5G.

Table 16
Quantum cryptography-based authentication protocols.

Author	Proposed model	Advantages	1	2	3	Simulation tool
Lohachab et al. [23] and 2018	QKD and ECC	DDOS attack	––	Model	×	AVISPA
EL-Latif et al. [12] 2019	Authenticated key distribution (AKD) protocol: QWHF-1 AQDC: QWHF-2.	AKD protocol: Secure from attacks: Impersonation Intercept–resend AQDC protocol: Secure from attacks: No-message, Measurement, Message capture, MITM	5G	Model	✓	––
Arul et al. [8] and 2020	QK-GRID AKA	Eavesdropper detection Rigid Security Adaptable to Key generation	5G	Model	✓	––
El-Latif et al. [12] and 2020	1D1PQWs Controlled Alternate Quantum Walks on a circle (CAQWs)	Brute Force attacks	5G	Model	✓	MATLAB
Mitchell et al. [11] and 2020	AKA protocol	Data integrity Confidentiality	5G	Model	✓	––
Amiri et al. [129] and 2021	QKD	Confidentiality of Keys Robustness	––	Survey	✓	––
Niemiec et al. [116] and 2013	QKD	Key Distillation,Privacy amplification	––	Survey	✓	QKD Simulator

1: Communication Standard; 2: Model/Survey; 3: Experimental/Statistical Analysis.

12. Advantages of quantum cryptography

12.1. Unconditional security

The most popular cryptography techniques are symmetric and asymmetric cryptography. Symmetric schemes resist Quantum attacks by increasing key and output sizes. However, Quantum computing greatly impacted most public-cryptosystems techniques like RSA and ECC. The security of these systems completely tampers with Shor's algorithm [11]. Therefore, Quantum algorithms such as Quantum Fourier Transform (QFT) and Quantum Walk (QW) for solving complex mathematical problems are popular research issues to provide Quantum-resistant systems. Adiabatic Quantum Computing (AQC) for optimization problems is useful for secure computation in futuristic IoT applications. Quantum key generation resolved the key distribution problem.

12.2. Sniffing detection

Classical cryptography provides a way to encrypt user data. A secured key is used among communicating entities to share data. In such a scenario, eavesdropping on the channel is also possible. Considering a scenario where the sender and receiver communicate on a Quantum-secured channel. According to Quantum no cloning, if an eavesdropper monitors the communication channel, he has a probability of 50% of getting the right basis vector. Therefore, the probability of the eavesdropper detecting Quantum information is: $1 - \frac{1}{2}n$ [98].

12.3. Eavesdropper detection

Secure communication requires mutual authentication between entities. As the number of devices is connected, there is a huge possibility of MITM and User impersonation (UI) attacks. The probability (pr) of the existence of an adversary on the Quantum channel can be identified as:

$Pr = PrBASE_A = BASE_B \bigwedge MEASURE_A \neq MEASURE_B$.

By applying laws of physics, QC provides an unrivalled security level: sender and receiver can reveal any modification of data by eavesdroppers.

12.4. Quantum entanglement

Entanglement is very useful in detecting eavesdropping. This process can be generated by splitting a single photon into two parts. Disturbance in one state will instantly disturb other states.

12.5. Confidentiality of keys

Classical cryptosystems suffer from insider threats. QKD is the first practical application of Quantum mechanics that provides the automatic distribution of keys and security for the overall system. Using QKD, the logistical burden of distributing keying material is resolved.

12.6. Authorization and trust

Only authorized people exchange information because of the Heisenberg Uncertainty principle photons will be destroyed once measured [104].

12.7. Bits error estimation

Once a thorough experiment examines the security of QKD, it ensures that the shared key consists of roughly half-of-the bits sent from sender to receiver. All other bits are rejected whose measurement basis is incorrect. Even if there is an adversary in the Quantum channel to obtain secret information, they can measure photons' polarization. Although they can do this using a random Rectilinear/Diagonal basis, any incorrect measurement changes the polarization. The Quantum Bit Error Rate (QBER) can be estimated based on QKD, and eavesdropping is impossible [113].

12.8. Key distillation process

Errors occur during communication over the network; however, these errors may not always occur because of eavesdropping. Still, some may also occur due to noise in detectors, optical misalignment, etc. The end-users identify and resolve (or delete) these errors if the values generated by QBER exceed a given threshold value. Therefore, a key reconciliation process is done between communicating parties [113].

12.9. Privacy amplification

This method ensures that the key is free of errors. The idea was based on entangled states, secure-hash methods (the BBCM protocol), and extractors. This method rejects the unmatched bits to create the secure key. The method hides shared key information from eavesdroppers.

12.10. Key size

Lattice-based cryptography becomes a perfect fit in terms of simplicity of implementation, key sizes and compactness. It is also considered safe when compared to other Quantum-safe alternative schemes.

13. Quantum key distribution (QKD)

The strength of the classical cryptography mechanism depends on the key distribution of secret key (symmetric) and public-key (asymmetric) cryptosystems. However, a problem with these is that the key transmission over an unreliable channel indicates that a third party may be on the communication path to intercept the key. Then adversaries use that secure key to decode secret user messages. Therefore, distributing keys without compromising the network's security is challenging. Quantum Key Distribution (QKD) provide solutions to the above-mentioned issues. The procedure is based on key distribution and not on message encryption. In QKD, two channels Quantum and classical, are used among communicating parties. The Quantum communication channel is used for secret key exchange. However, the classical way was used to demonstrate whether the shared key is distorted or not [102].

In this process, the photon polarization represents a particular state of a particle, which corresponds to each bit in a key. These bits are the foundation of QC. The QC is the principle of Quantum Key Distribution (QKD) [104]. In QKD, bits are transmitted using specific polarization angles. The two-bit values $|0>$ and $|1>$ are distinguished using polarization angle, also known as random basis. These bases are used each time to transfer keys securely. These two bases are more specifically referred to as rectilinear and diagonal basis. If we consider a rectilinear basis, photons are polarized at angles 0° or 90°. These photons, in turn, represent bit values 0 or 1, respectively. However, if the basis is diagonal, photons polarized at angle 45° represent bit 0°, and at angle 135°, it represents 1. Here, for ease of simplicity, these specific angles 0°, 90°, 45° and 135° can be written as H and V for horizontal and vertical. The D and A are used for diagonal and alternate. In polarization basis, the basis H, V is further denoted by + and D, A by X [130] as represented in Table 17.

However, the popularity of these conventions can be seen in the BB84 Protocol [99], categorized as the QKD protocol. According to the QKD scheme Alice and Bob communicate according to the procedure as represented in Fig. 12. A detailed explanation of this is as follows:

They exchange certain bits, which could lead to the final realization of the secret key. Unfortunately, some bits need to be discarded due to incompatible measurement settings. Those bits lost in transmission and would have never reached their destination, or the sender has not transmitted such bits are identified and deleted by the process known as sifting. Both communicating parties

Table 17
QKD random basis and its representation.

Basis	Bit value		Notation		Basis representation
	0	1	0	1	
Rectilinear	0°	90°	H	V	+
Diagonal	45°	135°	D	A	X

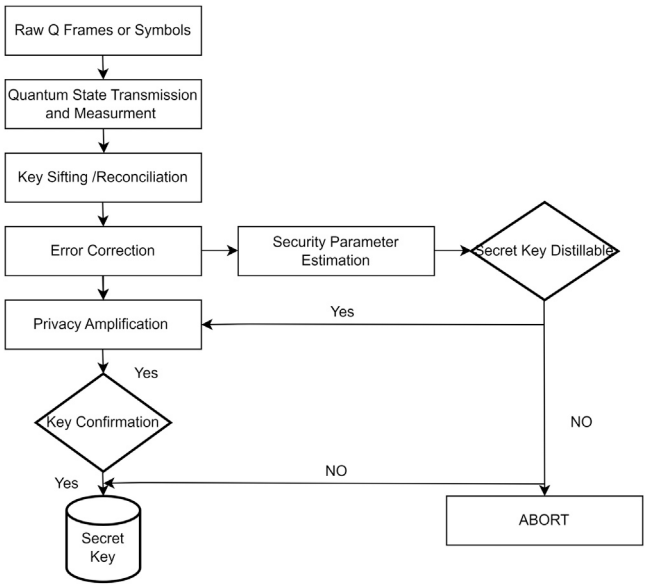


Fig. 12. Flowchart of Quantum Key Distribution.

perform error correction steps to identify the probability of information received by the eavesdropper. These correction steps estimate a security parameter that describes how much key content is lost in communication. The entire communication is aborted if this error amount exceeds a particular threshold value as they cannot guarantee any secrecy. However, if the error bits are below a certain acceptable level, then the privacy amplification process is applied. The process then identifies any unfinished information adversaries can utilize to decode shared secret keys [102]. The process of privacy amplification is based on an extractor, which can uniformly extract random bits from a random source.

13.1. QKD protocols

The inviolable laws of Quantum mechanics make QKD protocols an appropriate choice for ensuring security. The inability to duplicate(No-cloning) non-orthogonal states implies the perfect security of its protocols [101]. Vectors and matrices are the primary objects of Quantum computing, and their transformations and operations provide Quantum system states. In Quantum mechanics, this school of math's vector notation is that we do not write vectors with arrows. Instead, we write them with these brackets in BRA and KET form: $< v|; |v >$. Vector is a row or column vector that can be written as a Row vector (BRA): $< V|$; Column vector (KET): $|V >$. The Quantum states are also represented using those mentioned above $< Bra|$ and $|Ket >$ notation on the Bloch sphere. Compared to classical bit communication, these bits are replaced by a qubit in Quantum computing. In Quantum computing qubits can be manipulated using Quantum gates. Quantum gates are single and multi-qubits, as in Table 18. The QKD protocols are sorted based on the Uncertainty Principle (prepare-and-measure-based), including BB84, B92, SSP, SARG04, and S13. However, some others, such as E91, BBM92, DPS and COW, are Entanglement-based QKD protocols.

13.1.1. QKD protocols: uncertainty principle

The protocol was published in 1984 and named by Bennett and Brassard [23], who implemented this first. Nine QKD protocols have been proposed, out of which five are based on prepare-and-measure (uncertainty) [99], while the other four QKD protocols are based on entanglement [131], as:

(a) BB84 Protocol
Preliminaries

Table 18
Single and multiple qubits quantum gates.

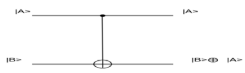
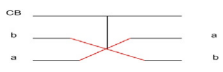
Single-qubit gates		
GATE	Matrix representation	Circuit representation
X-Gate	$X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$	
Z-Gate	$Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$	
H-Gate	$H = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$	
Multi-qubit gates		
CNOT	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$	
SWAP	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$	

Table 19
Representing polarization base of BB84 protocol.

	Degree	Bra-Ket Notation	Basis(B1)	Degree	Notation	Basis(B2)	B1 and B2
Rectilinear	0°	|0 >	↔	90°	|1 >	↕	+
Diagonal	45°	|0 >	↙	90°	|1 >	↘	×

Both parties must compute the total number of photons needed to form the secure key in advance. These numbers are used to identify the length of the message. The idea of the BB84 protocol is to deliver uninterrupted communication between sender and receiver by applying Quantum-computing notation (Bra-Ket). The BB84 protocol's four polarization states are represented in Table 19. The communication parties must agree that both:

(a) The sender (Alice) and Receiver (Bob) both must agree to share a sequence of randomly chosen photons based on one of the possible orientations: Rectilinear (0° or 90°) and Diagonal (45° or 135°).
(b) No secret communication or information is exchanged over the network rather than key exchange.
(c) As the final key is the product of both communicating parties, no one can individually determine the key in advance.

Assuming Alice (Sender) and Bob (Receiver) want to communicate with each other and Eve wants to intercept the basic message steps of the BB84 protocol as explained below:

(a) Alice prepares photons (qubits) randomly chosen in polarizations such as rectilinear 0° or 90° or diagonal polarization 45° and 135°. Before sending it to Bob, it keeps records of the polarization of each photon.
(b) The users can interchange faint flashes of polarized light [104] represented by a rectilinear and diagonal basis. The basis is shared between those users who have not shared secret information.
(c) The photons travel to a Bob. Then, for each photon, Bob considers a random basis.
(d) The horizontal/vertical or diagonal polarization is used by Bob to "read" the polarization angle of each particle sent by Alice.
(e) Bob keeps the log of the used basis and the resulting polarization measured.
(f) Bob does not know each photon's horizontal/vertical and diagonal beam splitter. He can only make a guess.
(g) If someone makes a wrong guess, he will get no information. For instance, if Bob analyses one photon with a diagonal filter sent by Alice using the rectilinear basis, then information about diagonally polarized photons is lost. Each measured photon selects one of the angles of polarization with a probability of 50%
(h) After making random choices, only the basis is informed by Bob Alice. Actual results are not shared between Alice and Bob.
(i) Bob will inform Alice only about the basis used to send measurement results.
(j) Alice publicly informs Bob about the correct measurement results. Then Bob keeps a log of its accurate measurement outcomes.
(k) The correctly measured bits create a one-time pad for sharing secret information. Both communicating parties discard the bits that do not match or are corrupted during transmission.
(l) The accurately measured photons are then transformed into a string of bits based on polarization using a notation: $D = 0$, $A = 1$, $H = 0$ and $V = 1$.
(m) QKD defines the procedure to generate the key by following this procedure securely.

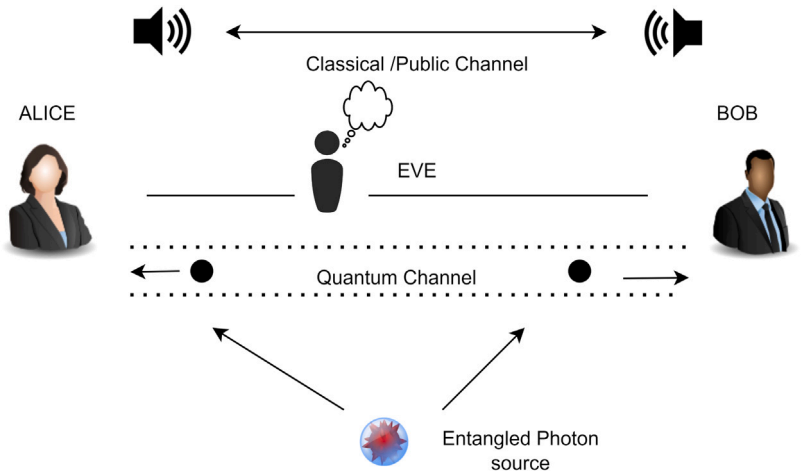


Fig. 13. Entanglement-based QKD protocol.

(n) The final key string consists of only those matched outcomes for which both the sender (Alice) and receiver (Bob) have the same measurement basis.
(o) The string of bits generated in the final step of the key generation process contains numerous errors. These errors are due to many factors such as decoherence, noise in detectors, and Quantum channel disturbances. Additionally, during Quantum communication, other errors or disturbances might be created due to the presence of an eavesdropper.
(p) Therefore, it is necessary to efficiently evaluate these errors to identify the adversary's presence in the Quantum communication channels. The Quantum bit error rate can be computed by comparing a small portion of the distributed Quantum key [113].

$$QBER = \frac{No.of\ Errors}{Total\ No.of\ Bits} * 100\%$$

13.1.2. QKD protocols: Entanglement

If two particles are entangled, measuring one attribute determines the state of the other. As a result, entanglement is one of Quantum theory's most perplexing phenomena. While communicating, these particles are entangled and will be connected during the entire communication process no matter how far apart the two are. Additionally, Ekert, in his seminal paper [1], considered an untrusted source of entangled photons between sender and receiver; the authors used this phenomenon for QC. Fig. 13 represents the entangled photon pair shared between two communicating entities. The photon pairs are exchanged using the Quantum channel however, classical channels are used for sharing measurement basis. It also claimed that eavesdropping could be easily detected by evaluating results by considering Bell's inequality [131]. One of the classic examples of entanglement-based protocols is E91, and there are some other entanglement protocols such as BBM92, DPS and COW [99,131].

(a) **E91 protocol**- The protocol was designed using entangled pairs of photons. Ekert proposed the protocol in 1991 [99]. Like BB84, chooses random measurement bases and utilizes the classical channel to identify the correct output string. The presence of an adversary on the Quantum channel could also be detected by computing the bell test.
(b) **BBM92**- This protocol was proposed by Bennett, Brassard, and Mermin [132] in 1992. The raw key exchange mechanism, key sifting, and privacy amplification are identical to BB84. It can be considered an entanglement-based version of BB84.
(c) **Differential-Phase-Shift (DPS)**- K. Inoue et al. [133] in their published work in 2003, they proposed entanglement-based QKD (DPS-QKD) named as DPS protocol. In comparison to the protocols above, DPS-QKD has a simple configuration. Moreover, the protocol is vigorous to photon number splitting (PNS) attacks.
(d) **Coherent One-Way protocol (COW)**- In 2004, Nicolas Gisin et al. [134] in their work consider time function for information encoding. Based on the concept of entanglement, the protocol has certain advantages. It provides high efficiency on distilled secret bits per qubit and tolerance to reduced interference visibility. COW protocol is also robust to photon number splitting (PNS) attacks like DPS.

13.2. Technical challenges of QKD

13.2.1. Control and organization of systems using quantum cryptography

QC provides huge security benefits to end-users. The limitation considers the end-user's control and organization of Quantum security protocols. The end-users cannot tailor the Quantum security services to fulfil their requirements. However, popular Quantum-based solutions allow customers to measure their particular security requirements. The end-users then choose the right level of data protection [104] for their devices.

13.2.2. Maximum distance for successful QKD transmission

Long distances Quantum communication is an important research issue. The challenge is due to the regeneration of the Quantum signal. The overall maximum distance covered by QKD transmissions [130] is currently over 200 km, successfully implemented. Additionally, the bit rate of QKD systems reaches only a few Mbit/s in a telecom metropolitan area network. Still, this disturbance is too high beyond 50 kilometres, which increases error rates. Therefore, it leaves the channel vulnerable to eavesdroppers and makes it virtually impossible to send information [113].

13.2.3. Quantum implementation cost for high security

Until recently, only large companies, final institutions, and public administrations utilized the services of QKD techniques. Due to the huge cost involved in its implementation, QKD techniques are used by specific end-users [135]. However, the major cost incurred in implementing Quantum devices is the usage of optical fibre. These optical fibre links create a Quantum channel for transmitting Quantum bits.

13.2.4. Quantum entanglement-based computing

Quantum entanglement is a method that can be very useful for detecting eavesdropping, but the cost associated with such a process is huge. However, considering the needs of the real world, keeping qubits entangled long enough is challenging.

13.2.5. Different end-user requirements

The level of security used by different platforms such as E-financial services, public administration and large companies varies greatly. It is complex to provide different end-user-specific services based on their changing demands. Users can only choose those services offered by network service providers. They can access the right level of data protection specific to their application. They cannot tailor these services according to their needs [113].

13.2.6. Quantum communication

In two-sided communication, the sender can always obtain more secret information about the receiver. However, the communication took almost ten years, but the fact mentioned above was noticed by Colbeck [129]. Similar work was done by Vein, Salvail, Schaffner and Sotakova [127]. In this, the authors demonstrated that any Quantum information scheme can always leak to an unauthorized user. The worst scenario of any Quantum system is the leakage of information. This information leakage in any Quantum protocol is essentially as worst as one can assume. Moreover, considering probabilistic authorization and protection, one of the two communicating entities, sender or receiver, is considered secure.

13.2.7. "Quantum rewinding": Zero-knowledge against quantum adversaries

The zero-knowledge property is used in the classical world and is referred to as rewinding. The theoretical aspect of this is proved by implementing the given verifier. The verifier makes arbitrary choices that sometimes are not consistent with the desired outcomes. In this scenario, such computation paths are culled. This selection is based on tracing interactions. If the inappropriate path is followed, the entire computation is reset or rewind. A rewinding approach in the Quantum setting is impossible. The system cannot maintain the alternative copy of the transcript or reproduce the state. Implementing and deploying Quantum rewinding in the context of zero-knowledge proofs is a complex cryptographic task that demands specialized cryptographic skills. IoT developers might encounter difficulties when integrating and managing these advanced cryptographic techniques.

13.2.8. Quantum security notions: Superposition access to oracles

The security of classical schemes relies on mathematical cryptosystems. However, one can assume the Quantum adversary on such systems. Network security is often considered an interactive game between an adversary and a challenger. Therefore, to prove Post-Quantum resistance, the adversary can only win the game with zero probability. In this scenario, one must consider Quantum adversaries coordinating Quantum with the challenger. The chosen plaintext attack (CPA) learning phase is an example of a game-based security system. In this example, superposition is of plaintexts to be encrypted query by an attacker. The challenger then returned superposition-according ciphertext to the attacker. The superposition is for defining the indistinguishability (IND) security of encryption schemes.

13.2.9. Position-based quantum cryptography

The main intention of position-based cryptography is to utilize the geographical position of an entity as a cryptographic credential. Consider the financial institution scenario, where the employee suffices as a credential to initiate the interchange of susceptible information. The main feature is the position verification of employees. The ultimate goal is to demonstrate the set of verifiers at a specific geographical location. Multiple attackers can break classical position-verification protocols. Classically, entities could easily simulate being at certain positions while their actual site is elsewhere. However, due to QM laws such as no-cloning on Quantum information, for collaborative attacks, one could come up with protocols that are impenetrable by these attacks, which later became insecure [127]. Integrating Position-based Quantum Cryptography into IoT systems could potentially lead to challenges related to managing devices, distributing cryptographic keys, and verifying their physical locations.

14. Security analysis of quantum resistant schemes

14.1. Security analysis

The security analysis of Quantum protocols based on mutual authentication such as Quantum Walk, QK-GRID, QWHF-1 and QWHF-2 are reviewed in our work.

14.1.1. Authentication and key issues

QK-GRID and QKG-AKA [31] were proposed to resolve a fixed and single-key problem. The proposed QKD method is to resolve a single invariable security interconnection issue through a reliable key management mechanism. The QK-GRID [8] protocol provides key generation and authentication methods. Additionally, on five dissimilar servers executing Ubuntu and Windows, the practical performance of the proposed mechanism is tested. In [4], the main feature is that the key is generated dynamically rather than having a fixed key length for encryption or decryption. This large enough key space is considered for the data encryption algorithm. Therefore, the key space for all control parameters would be $253 \times 5+40 = 2305$. Hence, it withstands numerous attacks, even brute force, on powerful computers. Furthermore, Quantum walks authentication functions utilized the features of Quantum laws and mechanics. In the 5G IoT scenario [4] QW mechanism is required, which provides secure block cipher techniques. For this purpose, a new S-box mechanism [128] that depends on the one-dimensional one-particle Quantum walks (1D1PQW's) is utilized. Another proposed key interchange protocol is Authenticated Key Distribution (AKD) based on QW [12]. Additionally, the author's work on Quantum Walk Hash Function (QWHF-1) [4] resolves the key exchange problem of classical symmetric algorithms. Utilizing the properties of Quantum mechanics for creating authenticated key streams between two communicating parties, designed a protocol based on Controlled Alternate Quantum Walks(CAQWs). The hash values constructed by CAQWs are shared with others using a communication channel. In this protocol, the master key parameters are to be shared before establishing communication. No one can deduce any information about the (α, θ_0, θ_1, and θ_2), which are master key parameters and are utilized for constructing these values or the final secret key. Furthermore, Authenticated Quantum Direct Communication (AQDC) is based on QWHF-2 [12]. Rather than classical hash functions based on mathematical functions, QWHF is based on Quantum laws. The scheme mentioned above uses single photons based on a Quantum hash-based function. This message authentication aims to verify the shared message's integrity and origin once it reaches its destination.

14.1.2. Security attacks

El-Latif et al. [12] encryption scheme based on QW resists DoS, data modification and replay attacks. They presented a solution for histogram analysis and entropy attacks. They cover threats imposed due to Quantum computers. In [12], AKD and AQDC protocols based on QWHF-1 and QWHF-2 detect any active attack. Therefore, the protocol resists intercept–resend, message, and no-message attacks. However, Arul et al. [8] proposed Quantum Key GRID (QK-GRID) based on QKD. They resolve attacks that mainly arise due to massive device connectivity. The author' proved that their protocol withstands attacks such as network jamming and privacy attacks. These attacks are menacing innovative technologies. In addition, their protocol successfully solved previous key management issues, so ejecting secret information from the system in polynomial time is impossible. They provide security for smart applications such as Internet of Vehicles (IoV) domain secure and reliable transportation systems.

15. Comparison of quantum-resistant schemes with classical authentication

In comparison with classical schemes, the resource requirements of Quantum-resistant schemes differ. Table 20 represents the juxtaposition of the secret key and message sizes, proposed algorithms, their impact on Quantum computers and the time taken to execute in a classical computer of selected Post-Quantum schemes. Quantum-resistant cryptography provides alternatives to classical cryptographic primitives and solutions to future Quantum computer attacks. Classical cryptographic algorithms such as RSA, ECC, and DH are not Quantum-resistant. Therefore, PQC families must be taken into account. The Quantum-resistant schemes require larger key sizes and signature or ciphertext or message sizes than classical schemes. However, this incurs costs and provides security against Quantum computer attacks [125]. By defining specific parameters, the European research group PQCRYPTO [16] has given initial recommendations for Quantum-resistant schemes [118]. Quantum-resistant schemes are referred as multivariate cryptography [16], secure code-based [125], hash-based signatures or lattice-based schemes [16] and QC.

15.1. Public-key signatures

The asymmetric cryptosystem has been extensively used in all secret communications. It provides authenticity and non-repudiation. The asymmetric signature scheme is based on traditional RSA and ECC signatures and uses a small public key of 64 to 1,056 [6]. Merkle [114], in his work, proposed a hash-based signature scheme that utilizes a binary hash tree [6] to create the signatures. Later, the extended Merkle Signature [16] scheme (XMSS) was implemented by considering weak security assumptions. However, the proposed scheme faces challenges due to several devices used for signature generation and synchronization is required for these devices each time after signature generation. SPHINCS [118] has been proposed to avoid this, referred to as a multi-tree version of XMSS. ElGamal [40], another digital signature scheme, has also gained much popularity.

Table 20
Comparison of Post-Quantum cryptography schemes with classical cryptography.

Proposed algorithms	Purpose	Public key - size(Bytes)	Data– size	Time in a classical computer	Impact on Quantum computer	Time in Quantum computer	Refs.
Public-key signatures							
Hash-based							
XMSS (stateful)	Reduces the signature size of MSS	64	2500 – 2820	————————	Secure, Not broken by Shor in polynomial time.	The run time <t,x E {0,1}ˆn was selected on a random basis with a uniform distribution.	[136]
SPHINCS (state free)	provide security even against attackers.	1056	41,000	————————-	Against Quantum attacks, expendable 2ˆ128 security is provided.	Signature time depends upon the number of Layers.	[137]
Multivariate based							
HFEv-*	Ability to produce the shortest signature.	500,000 – 1,000,000	25-32	——-	Most promising and Secure but has a large signature as compared to classical schemes.	polynomial Time	[137]
Lattice-based							
BLISS	Based on preimage sample table functions.	——	———	—————	Quantum Resistant for constrained IoT Devices.	Ring-LWE provides security higher than 156 bit: 68 ms and 18.8 ms for encryption. and decryption	[6]
Public-Key Encryption							
Code based:							
McEliece	Provide codeword based Encryption	958,482 – 1,046,739	187-194	————————-	Withstood all proposed attacks, no Quantum attack.	————————	[6]
Lattice-based							
NTRUEncrypt with binary Goppa	It gives us very good error correcting capabilities.	1495 - 2062	1495 – 2062	Difficult to break (Time Undetermined)	Provide Security and concerning parameters size classical schemes grow linearly.	Non-Deterministic Polynomial Time.	[6,16]
KEY-EXCHANGE							
Lattice-based							
NewHope	Resist Quantum attacks as it is based on Mathematical problem RLWE.	——-	1824 – 2048	Undetermined	hard to break	Non-Deterministic polynomial Time hardness.	[122]
Supersingular isogenies							
SIDH	Secure and attack resistant: classical and Quantum	———	564	——-	Secure, provide small message size for key exchange.	——	[16], [122]
Quantum Cryptography	Based						

(*continued on next page*)

Table 20 (*continued*).

Quantum Walk	Secured Key exchange	——	———	————	Quantum Resistant	2^305 Keyspace required.	[6]
Classical schemes:							
Symmetric key encryption							
AES-256	Faster and secure Encryption	———	———	Time complexity: - Brute-force attack O(2n). -AES-128 bit encryption O(2^128).	Doubling the Key – Size	Time complexity: -AES-128 bit encryption: O(2^64), - Easy to crack by Quantum computer. -Cracked using Brute Force attack O(2^n/2).	[6]
Salsa20	symmetric stream ciphers for authenticated encryption construction	———	———	———-	Insecure, the Grover algorithm can break security Salsa 20.	——-	[138]
Asymmetric-Key Encryption							
RSA	Operates on authentication function and Public key cryptosystems.	For RSA-2048: 256 RSA-4096:512	256 and 512	Sub-Exponential Time analysis: C*10^8 for 512 bits and C*10^17 for 1024 bits. Here, C constant.	Completely insecure, solved by Shor's[92]algorithm.	0.5*10^9 and calculated to be 4n^3	[6] ,[122]
ECC	Operate on Elliptic Curve equations	For 256 bit –32 512 bit –64	32 and 64	Sub-Exponential Time analysis: C*10^8 for 512 bits and c*10^17 for 1024 bits.	Insecure, Shor's algorithm can break ECC.	360n^3;n: Number of bits	[6]

15.2. Public key encryption

McEliece and Nieder-reiter [118] used Goppa codes as hard encryption techniques. The asymmetric keys required are large compared to classical schemes, which are about 1 MB, and the ciphertext is only around 190 bytes as in the scope of classical approaches. They also compared lattice and code-based cryptography schemes implemented for PQC but found lattice less trusted than code-based cryptography. In 1998, Hoffstein, Pipher, and Silverman [116] proposed the NTRUEncrypt [6] scheme, which needed 1.5–2.0 kB for both keys and ciphertext. The lattice scheme based on asymmetric encryption has gained popularity due to its efficiency and compact keys. Lattice-based signature schemes are another milestone that has been achieved by lattice-based cryptography. One popular signature scheme based on preimage sample functions proposed by Gentry, Peikert and Vaikuntanathan [138] is the GPV scheme. Other schemes consider preimage sample functions such as the decisional ring-LWE (Learning with Errors), GLP, and ring-SIS problems like Bimodal Lattice Signature (BLISS) schemes. A signature validation program (SVP) is also used in lattice-based cryptosystems, including NTRUEncrypt and NTRUSign [119]. Additionally, a public-key cryptosystem such as RSA utilizes the complexity of integer factorization [6] for data encryption and decryption.

15.3. Key-exchange

Classical secure key exchange methods are based on the hardness of mathematical formulation discrete logarithm problem (DLP) [139,140] and elliptic curve discrete logarithm problem (ECDLP) [16,132]. These schemes were considered safe before considering Quantum computing. These classical problems are based on public-key cryptosystems, which could easily be broken with a Quantum computer. Quantum computers are also threatening the security of the classical secure Diffie–Hellman key interchange algorithm, which works by generating and distributing new keys for each session. However, some specific key-interchange schemes provide better performance considering the problems such as the time required in key generation and bandwidth demand. Quantum-based cryptographic schemes work towards finding solutions to these problems. For example, the lattice-based key exchange scheme that is gaining much research attention is NewHope [100]. This scheme requires the transmission of network packets of approximately 2 KB compared to the traditional ECDH of 32 to 64 bytes. SIDH [141], based on another PQC scheme, the supersingular isogeny, is very close to the traditional scheme, as it only requires 564 bytes. However, this scheme is gaining attention, but the scheme is new and cannot be believed. Moreover, QC-based key exchange provides better security by using Quantum physics to improve the key exchange method. Quantum cryptographic algorithms such as Quantum Walk, QWHF, and Quantum–GRID provide a new way to exchange and secure key exchange methods robust in the Post-Quantum world [6,118].

15.4. Symmetric encryption

The most popular AES scheme is based on a private key structure. The National Institute of Standards and Technology [136] (NIST) gave standardization to this scheme in 2001. The AES algorithm cryptographic scheme allows plain text to be encrypted. AES enhances security by providing a variant of secure keys 128, 196, and 256 bits. These bits are often referred to as AES-128, AES-196, and AES-256 [6]. The scheme is well-known for securing IoT communication. The best-known attack against AES is brute-force [16]. Hence, to attain a security level of 128 bits against Quantum computing and information, we need an AES to secure a key size of 256 bits [118].

16. Conclusion and future research directions

16.1. Conclusion

This survey primarily emphasized enhancing the security of IoT communication, given the increasing popularity of IoT devices in conjunction with high-speed 5G internet connectivity. The usage of 5G-enabled IoT makes people's lives better and more convenient. The data emerging from these billions of IoT devices are vulnerable to attacks. The data generated and transferred from these IoT devices is valuable and must be secured due to their usage in various applications. The cryptographic analysis of making IoT communication safe and sound needs to be addressed. In this study, we examined the recent forge ahead in Quantum computing to yield a promising data and network security future. In comparison with classical cryptography, the PQC family provides unconditional security. Quantum cryptography-based QKD protocols can solve cyberspace security and key exchange issues for future secure internet communication. Exploring the benefits of 5G-powered IoT in our work, we delved into the implications of Quantum-resistant cryptographic schemes on the security of the IoT infrastructure. We also examined in detail the impact of the Quantum era on the effectiveness of classical security protocols for IoT communication. This paper also discusses the 5G-IoT security requirements, Quantum-resistant authentication schemes and challenges.

16.2. Future research directions

We have identified that for the futuristic 5G-IoT communication scenario, more security and mutual authentication are necessary; therefore, a change in classical cryptography is also required. However, much more work is required to prepare the 5G-IoT system for the Quantum world. We reviewed many companies that are working on providing 5G-enabled IoT services to users and researchers who are developing new Quantum-resistant security solutions for the IoT framework. The research on Quantum-resistant cryptography experimental demonstration is still at the prototype level. A Quantum blind signature, QC, satellite-based QKD and adiabatic quantum computing are gaining huge popularity. However, our theoretical cognizance of the potential and limitations of PQC is ceaselessly expanding. The PQC and QC have received much recognition from smart medical, smart industry, smart academia, and smart agriculture. Hence, more attention and effort are required to make ready 5G-enabled IoT for the futuristic secure communication network. Hence, overcoming challenges such as Quantum imperfections, Quantum Internet, Quantum key distance, and cost and achieving the goal of preparing IoT for a Quantum world will take a rapid pace in years to come. Additionally, to prevent unauthorized access and secure IoT communication cryptography stands as the most effective practice. PQC has reached its pinnacle, and extensive research has been conducted on topics like holomorphic encryption, zero-knowledge proofs, oblivious transfer to bit commitment, and device-independent cryptography, especially in the context of resource-limited IoT environments. These fields hold significant promise for various applications in the coming 5 to 10 years, including secure multi-party computation, confidential voting, safeguarding data sharing while preserving privacy, and strengthening device authentication. These areas are likely to remain active research domains in the near future. These methods are applied to enhance security, safeguard privacy, and improve efficiency within the IoT communication framework. This survey is anticipated to function as a valuable resource for enhancing security in forthcoming IoT applications.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] L. Chettri, R. Bera, A comprehensive survey on Internet of Things (IoT) toward 5G wireless systems, IEEE Internet Things J. 7 (1) (2019) 16–32.
[2] S. Li, L. Da Xu, S. Zhao, 5G Internet of Things: A survey, J. Ind. Inf. Integr. 10 (2018) 1–9.
[3] D. Sey, A survey on authentication methods for the internet of things, PeerJ Preprints 6 (2018) e26474v2.

[4] R. Melki, H.N. Noura, A. Chehab, Lightweight multi-factor mutual authentication protocol for IoT devices, Int. J. Inf. Secur. 19 (2020) 679–694.
[5] A. Adeel, M. Gogate, S. Farooq, C. Ieracitano, K. Dashtipour, H. Larijani, A. Hussain, A survey on the role of wireless sensor networks and IoT in disaster management, in: Geological Disaster Monitoring Based on Sensor Networks, Springer, 2019, pp. 57–66.
[6] C. Cheng, R. Lu, A. Petzoldt, T. Takagi, Securing the Internet of Things in a quantum world, IEEE Commun. Mag. 55 (2) (2017) 116–120.
[7] D. Singh, B. Kumar, S. Singh, S. Chand, Evaluating authentication schemes for real-time data in wireless sensor network, Wirel. Pers. Commun. 114 (2020) 629–655.
[8] R. Arul, G. Raja, A.O. Almagrabi, M.S. Alkatheiri, S.H. Chauhdary, A.K. Bashir, A quantum-safe key hierarchy and dynamic security association for LTE/SAE in 5G scenario, IEEE Trans. Ind. Inform. 16 (1) (2019) 681–690.
[9] K.S. Roy, H.K. Kalita, A quantum safe user authentication protocol for the internet of things, Int. J. Next-Gener. Comput. 10 (3) (2019).
[10] L. Malina, P. Dzurenda, S. Ricci, J. Hajny, G. Srivastava, R. Matulevičius, A.-A.O. Affia, M. Laurent, N.H. Sultan, Q. Tang, Post-quantum era privacy protection for intelligent infrastructures, IEEE Access 9 (2021) 36038–36077.
[11] C.J. Mitchell, The impact of quantum computing on real-world security: A 5G case study, Comput. Secur. 93 (2020) 101825.
[12] A.A. Abd EL-Latif, B. Abd-El-Atty, S.E. Venegas-Andraca, W. Mazurczyk, Efficient quantum-based security protocols for information sharing and data protection in 5G networks, Future Gener. Comput. Syst. 100 (2019) 893–906.
[13] S.S. Gill, A. Kumar, H. Singh, M. Singh, K. Kaur, M. Usman, R. Buyya, Quantum computing: A taxonomy, systematic review and future directions, Softw. - Pract. Exp. 52 (1) (2022) 66–114.
[14] A.P. Bhatt, A. Sharma, Quantum cryptography for internet of things security, J. Electron. Sci. Technol. 17 (3) (2019) 213–220.
[15] J. Cao, P. Yu, X. Xiang, M. Ma, H. Li, Anti-quantum fast authentication and data transmission scheme for massive devices in 5G NB-IoT system, IEEE Internet Things J. 6 (6) (2019) 9794–9805.
[16] O.S. Althobaiti, M. Dohler, Cybersecurity challenges associated with the internet of things in a post-quantum world, IEEE Access 8 (2020) 157356–157381.
[17] V.K. Ralegankar, J. Bagul, B. Thakkar, R. Gupta, S. Tanwar, G. Sharma, I.E. Davidson, Quantum cryptography-as-a-service for secure UAV communication: applications, challenges, and case study, IEEE Access 10 (2021) 1475–1492.
[18] Y. Xu, C. Wang, X. Wang, H. Zhu, Scalable multiple GHZ states equations and its applications in efficient quantum key agreement, Quantum Inf. Process. 21 (3) (2022) 91.
[19] X. Li, K. Zhang, L. Zhang, X. Zhao, A new quantum multiparty simultaneous identity authentication protocol with the classical third-party, Entropy 24 (4) (2022) 483.
[20] I.W. Primaatmaja, K.T. Goh, E.Y.-Z. Tan, J.T.-F. Khoo, S. Ghorai, C.C.-W. Lim, Security of device-independent quantum key distribution protocols: a review, Quantum 7 (2023) 932.
[21] D. Chawla, P.S. Mehra, A survey on quantum computing for internet of things security, Procedia Comput. Sci. 218 (2023) 2191–2200.
[22] M. Wazid, A.K. Das, S. Shetty, P. Gope, J.J. Rodrigues, Security in 5G-enabled internet of things communication: issues, challenges, and future research roadmap, IEEE Access 9 (2020) 4466–4489.
[23] A. Lohachab, et al., Using quantum key distribution and ECC for secure inter-device authentication and communication in IoT infrastructure, in: Proceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT), 2018, pp. 26–27.
[24] A. Zhang, J. Chen, R.Q. Hu, Y. Qian, Seds: Secure data sharing strategy for D2D communication in LTE-advanced networks, IEEE Trans. Veh. Technol. 65 (4) (2015) 2659–2672.
[25] L. Spalazzi, S. Tacconi, Classification of attacks on cryptographic protocols, in: Handbook of Electronic Security and Digital Forensics, World Scientific Publishing, Singapore, 2010, pp. 47–82.
[26] K. Shafique, B.A. Khawaja, F. Sabir, S. Qazi, M. Mustaqim, Internet of things (IoT) for next-generation smart systems: A review of current challenges, future trends and prospects for emerging 5G-IoT scenarios, Ieee Access 8 (2020) 23022–23040.
[27] G. Avoine, S. Canard, L. Ferreira, IoT-friendly AKE: forward secrecy and session resumption meet symmetric-key cryptography, in: Computer Security–ESORICS 2019: 24th European Symposium on Research in Computer Security, Luxembourg, September 23–27, 2019, Proceedings, Part II 24, Springer, 2019, pp. 463–483.
[28] M. Turkanović, B. Brumen, M. Hölbl, A novel user authentication and key agreement scheme for heterogeneous ad hoc wireless sensor networks, based on the Internet of Things notion, Ad Hoc Netw. 20 (2014) 96–112.
[29] H. Nagdewani, P.S. Mehra, A complete internet of things based home security system, in: 2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA), IEEE, 2022, pp. 1–5.
[30] O. Althobaiti, M. Al-Rodhaan, A. Al-Dhelaan, An efficient biometric authentication protocol for wireless sensor networks, Int. J. Distrib. Sens. Netw. 9 (5) (2013) 407971.
[31] L. Wei, R.Q. Hu, Y. Qian, G. Wu, Key elements to enable millimeter wave communications for 5G wireless systems, IEEE Wirel. Commun. 21 (6) (2014) 136–143.
[32] A.A. Zaidi, R. Baldemair, H. Tullberg, H. Bjorkegren, L. Sundstrom, J. Medbo, C. Kilinc, I. Da Silva, Waveform and numerology to support 5G services and requirements, IEEE Commun. Mag. 54 (11) (2016) 90–98.
[33] R. Chaudhary, N. Kumar, S. Zeadally, Network service chaining in fog and cloud computing for the 5g environment: Data management and security challenges, IEEE Commun. Mag. 55 (11) (2017) 114–122.
[34] A.A. Zaidi, R. Baldemair, V. Molés-Cases, N. He, K. Werner, A. Cedergren, OFDM numerology design for 5G new radio to support IoT, eMBB, and MBSFN, IEEE Commun. Stand. Mag. 2 (2) (2018) 78–83.
[35] Y. Niu, Y. Li, D. Jin, L. Su, A.V. Vasilakos, A survey of millimeter wave communications (mmwave) for 5G: opportunities and challenges, Wirel. Netw. 21 (2015) 2657–2676.
[36] S. Sharma, S. Satapathy, S. Singh, A.K. Sahu, M.S. Obaidat, S. Saxena, D. Puthal, Secure authentication protocol for 5G enabled IoT network, in: 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), IEEE, 2018, pp. 621–626.
[37] M. Irfan, H. Jawad, B.B. Felix, S.F. Abbasi, A. Nawaz, S. Akbarzadeh, M. Awais, L. Chen, T. Westerlund, W. Chen, Non-wearable IoT-based smart ambient behavior observation system, IEEE Sens. J. 21 (18) (2021) 20857–20869.
[38] I.-P. Chang, T.-F. Lee, T.-H. Lin, C.-M. Liu, Enhanced two-factor authentication and key agreement using dynamic identities in wireless sensor networks, Sensors 15 (12) (2015) 29841–29854.
[39] A.C. Jose, R. Malekian, B.B. Letswamotse, Improving smart home security; integrating behaviour prediction into smart home, Int. J. Sens. Netw. 28 (4) (2018) 253–269.
[40] A. Braeken, P. Porambage, M. Stojmenovic, L. Lambrinos, eDAAAS: Efficient distributed anonymous authentication and access in smart homes, Int. J. Distrib. Sens. Netw. 12 (12) (2016) 1550147716682037.
[41] S. Yu, N. Jho, Y. Park, Lightweight three-factor-based privacy-preserving authentication scheme for iot-enabled smart homes, IEEE Access 9 (2021) 126186–126197.
[42] A. Dorri, S.S. Kanhere, R. Jurdak, P. Gauravaram, Blockchain for IoT security and privacy: The case study of a smart home, in: 2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), IEEE, 2017, pp. 618–623.
[43] S. Kailasam, S.D.M. Achanta, P. Rama Koteswara Rao, R. Vatambeti, S. Kayam, An IoT-based agriculture maintenance using pervasive computing with machine learning technique, Int. J. Intell. Comput. Cybern. 15 (2) (2022) 184–197.

[44] M. Gupta, M. Abdelsalam, S. Khorsandroo, S. Mittal, Security and privacy in smart farming: Challenges and opportunities, IEEE Access 8 (2020) 34564–34584.
[45] M.A. Ferrag, L.A. Maglaras, H. Janicke, J. Jiang, L. Shu, Authentication protocols for internet of things: a comprehensive survey, Secur. Commun. Netw. 2017 (2017).
[46] M.A. Ferrag, L. Maglaras, A. Derhab, Authentication and authorization for mobile IoT devices using biofeatures: Recent advances and future trends, Secur. Commun. Netw. 2019 (2019).
[47] L. Guo, C. Zhang, J. Sun, Y. Fang, A privacy-preserving attribute-based authentication system for mobile health networks, IEEE Trans. Mob. Comput. 13 (9) (2013) 1927–1941.
[48] A. Sampath Dakshina Murthy, T. Karthikeyan, R. Vinoth Kanna, Gait-based person fall prediction using deep learning approach, Soft Comput. (2021) 1–9.
[49] P.S. Mehra, Y.B. Mehra, A. Dagur, A.K. Dwivedi, M. Doja, A. Jamshed, COVID-19 suspected person detection and identification using thermal imaging-based closed circuit television camera and tracking using drone in Internet of Things, Int. J. Comput. Appl. Technol. 66 (3–4) (2021) 340–349.
[50] W. AL-mawee, et al., Privacy and security issues in IoT healthcare applications for the disabled users a survey, 2012.
[51] S. Butt, J. Diaz-Martinez, T. Jamal, A. Ali, E. De-La-Hoz-Franco, M. Shoaib, IoT smart health security threats, in: Proceedings-2019 19th International Conference on Computational Science and Its Applications, ICCSA 2019, 10, IEEE, 2019, pp. 26–31.
[52] I. Yusof, A.-S.K. Pathan, Preventing persistent cross-site scripting (XSS) attack by applying pattern filtering approach, in: The 5th International Conference on Information and Communication Technology for the Muslim World (ICT4M), IEEE, 2014, pp. 1–6.
[53] S. Rathore, P.K. Sharma, J.H. Park, Xssclassifier: an efficient XSS attack detection approach based on machine learning classifier on SNSs, J. Inf. Process. Syst. 13 (4) (2017) 1014–1028.
[54] R.M. Thiyab, M. Ali, F. Basil, et al., The impact of SQL injection attacks on the security of databases, in: Proceedings of the 6th International Conference of Computing & Informatics, School of Computing, 2017, pp. 323–331.
[55] L. Ntagwabira, S.L. Kang, Use of query tokenization to detect and prevent SQL injection attacks, in: 2010 3rd International Conference on Computer Science and Information Technology, Vol. 2, IEEE, 2010, pp. 438–440.
[56] R. Latif, H. Abbas, S. Latif, A. Masood, Distributed denial of service attack source detection using efficient traceback technique (ETT) in cloud-assisted healthcare environment, J. Med. Syst. 40 (2016) 1–13.
[57] C.-M. Yu, Y.-T. Tsou, C.-S. Lu, S.-Y. Kuo, Constrained function-based message authentication for sensor networks, IEEE Trans. Inf. Forensics Secur. 6 (2) (2011) 407–425.
[58] I. Mistry, S. Tanwar, S. Tyagi, N. Kumar, Blockchain for 5G-enabled IoT for industrial automation: A systematic review, solutions, and challenges, Mech. Syst. Signal Process. 135 (2020) 106382.
[59] K. Tsiknas, D. Taketzis, K. Demertzis, C. Skianis, Cyber threats to industrial IoT: a survey on attacks and countermeasures, IoT 2 (1) (2021) 163–186.
[60] M. Chandrasekaran, R. Chinchani, S. Upadhyaya, Phoney: Mimicking user response to detect phishing attacks, in: 2006 International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM'06), IEEE, 2006, p. 5.
[61] K. Demertzis, P. Kikiras, N. Tziritas, S.L. Sanchez, L. Iliadis, The next generation cognitive security operations center: network flow forensics using cybersecurity intelligence, Big Data Cogn. Comput. 2 (4) (2018) 35.
[62] E. Stalmans, B. Irwin, A framework for DNS based detection and mitigation of malware infections on a network, in: 2011 Information Security for South Africa, IEEE, 2011, pp. 1–8.
[63] M. Al-Hawawreh, F. Den Hartog, E. Sitnikova, Targeted ransomware: A new cyber threat to edge system of brownfield industrial Internet of Things, IEEE Internet Things J. 6 (4) (2019) 7137–7151.
[64] O.M. Alhawi, J. Baldwin, A. Dehghantanha, Leveraging machine learning techniques for windows ransomware network traffic detection, Cyber Threat Intell. (2018) 93–106.
[65] A.O. Almashhadani, M. Kaiiali, S. Sezer, P. O'Kane, A multi-classifier network-based crypto ransomware detection system: A case study of locky ransomware, IEEE Access 7 (2019) 47053–47067.
[66] R. Muraleedharan, L.A. Osadciw, Cross layer denial of service attacks in wireless sensor network using swarm intelligence, in: 2006 40th Annual Conference on Information Sciences and Systems, IEEE, 2006, pp. 1653–1658.
[67] M. Jebalia, A.B. Letaifa, M. Hamdi, S. Tabbane, A revocation game model for secure cloud storage, in: 2014 International Conference on High Performance Computing & Simulation (HPCS), IEEE, 2014, pp. 1016–1017.
[68] M. Usman, S. Raponi, M. Qaraqe, G. Oligeri, KaFHCa: Key-establishment via frequency hopping collisions, in: ICC 2021-IEEE International Conference on Communications, IEEE, 2021, pp. 1–6.
[69] P. Radanliev, D. De Roure, K. Page, J.R. Nurse, R. Mantilla Montalvo, O. Santos, L. Maddox, P. Burnap, Cyber risk at the edge: current and future trends on cyber risk analytics and artificial intelligence in the industrial internet of things and industry 4.0 supply chains, Cybersecurity 3 (1) (2020) 1–21.
[70] S. Raza, L. Wallgren, T. Voigt, SVELTE: Real-time intrusion detection in the Internet of Things, Ad hoc Netw. 11 (8) (2013) 2661–2674.
[71] J. Cui, L.S. Liew, G. Sabaliauskaite, F. Zhou, A review on safety failures, security attacks, and available countermeasures for autonomous vehicles, Ad Hoc Netw. 90 (2019) 101823.
[72] Y. Ming, X. Shen, PCPA: A practical certificateless conditional privacy preserving authentication scheme for vehicular ad hoc networks, Sensors 18 (5) (2018) 1573.
[73] Y. Hao, Y. Cheng, C. Zhou, W. Song, A distributed key management framework with cooperative message authentication in VANETs, IEEE J. Selected Areas Commun. 29 (3) (2011) 616–629.
[74] R. Amin, G. Biswas, A secure light weight scheme for user authentication and key agreement in multi-gateway based wireless sensor networks, Ad Hoc Netw. 36 (2016) 58–80.
[75] P.S. Mehra, M.N. Doja, B. Alam, Codeword authenticated key exchange (CAKE) light weight secure routing protocol for WSN, Int. J. Commun. Syst. 32 (3) (2019) e3879.
[76] S. Wang, J. Wang, Z. Yu, Privacy-preserving authentication in wireless IoT: applications, approaches, and challenges, IEEE Wirel. Commun. 25 (6) (2018) 60–67.
[77] S.F. Abbasi, J. Ahmad, J.S. Khan, M.A. Khan, S.A. Sheikh, Visual meaningful encryption scheme using intertwinning logistic map, in: Intelligent Computing: Proceedings of the 2018 Computing Conference, Volume 2, Springer, 2019, pp. 764–773.
[78] Y. Lu, G. Xu, L. Li, Y. Yang, Anonymous three-factor authenticated key agreement for wireless sensor networks, Wirel. Netw. 25 (2019) 1461–1475.
[79] M.S. Farash, M. Turkanović, S. Kumari, M. Hölbl, An efficient user authentication and key agreement scheme for heterogeneous wireless sensor network tailored for the Internet of Things environment, Ad Hoc Netw. 36 (2016) 152–176.
[80] P. Gope, R. Amin, S.H. Islam, N. Kumar, V.K. Bhalla, Lightweight and privacy-preserving RFID authentication scheme for distributed IoT infrastructure with secure localization services for smart city environment, Future Gener. Comput. Syst. 83 (2018) 629–637.
[81] B. Di Ying, D. Makrakis, H.T. Mouftah, Anti-traffic analysis attack for location privacy in WSNs, EURASIP J. Wireless Commun. Networking 2014 (1) (2014) 1–15.
[82] J. Ni, X. Lin, X.S. Shen, Efficient and secure service-oriented authentication supporting network slicing for 5G-enabled IoT, IEEE J. Sel. Areas Commun. 36 (3) (2018) 644–657.

[83] D. Basin, J. Dreier, L. Hirschi, S. Radomirovic, R. Sasse, V. Stettler, A formal analysis of 5G authentication, in: Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, 2018, pp. 1383–1396.
[84] S. Gong, A. El Azzaoui, J. Cha, J.H. Park, Secure secondary authentication framework for efficient mutual authentication on a 5G data network, Appl. Sci. 10 (2) (2020) 727.
[85] R. Watro, D. Kong, S.-f. Cuti, C. Gardiner, C. Lynn, P. Kruus, TinyPK: securing sensor networks with public key technology, in: Proceedings of the 2nd ACM Workshop on Security of Ad Hoc and Sensor Networks, 2004, pp. 59–64.
[86] M.L. Das, Two-factor user authentication in wireless sensor networks, IEEE Trans. Wirel. Commun. 8 (3) (2009) 1086–1090.
[87] A.K. Dwivedi, P.S. Mehra, O. Pal, M.N. Doja, B. Alam, EETSP: Energy-efficient two-stage routing protocol for wireless sensor network-assisted Internet of Things, Int. J. Commun. Syst. 34 (17) (2021) e4965.
[88] A.K. Das, P. Sharma, S. Chatterjee, J.K. Sing, A dynamic password-based user authentication scheme for hierarchical wireless sensor networks, J. Netw. Comput. Appl. 35 (5) (2012) 1646–1656.
[89] S. Kumari, H. Om, Authentication protocol for wireless sensor networks applications like safety monitoring in coal mines, Comput. Netw. 104 (2016) 137–154.
[90] Q. Jiang, J. Ma, F. Wei, Y. Tian, J. Shen, Y. Yang, An untraceable temporal-credential-based two-factor authentication scheme using ECC for wireless sensor networks, J. Netw. Comput. Appl. 76 (2016) 37–48.
[91] C.-C. Chang, H.-D. Le, A provably secure, efficient, and flexible authentication scheme for ad hoc wireless sensor networks, IEEE Trans. Wirel. Commun. 15 (1) (2015) 357–366.
[92] B. Vaidya, D. Makrakis, H.T. Mouftah, Improved two-factor user authentication in wireless sensor networks, in: 2010 IEEE 6th International Conference on Wireless and Mobile Computing, Networking and Communications, IEEE, 2010, pp. 600–606.
[93] M.K. Khan, S. Kumari, An improved user authentication protocol for healthcare services via wireless medical sensor networks, Int. J. Distrib. Sens. Netw. 10 (4) (2014) 347169.
[94] B.D. Deebak, Secure and efficient mutual adaptive user authentication scheme for heterogeneous wireless sensor networks using multimedia client–server systems, Wirel. Pers. Commun. 87 (2016) 1013–1035.
[95] D. Mishra, P. Vijayakumar, V. Sureshkumar, R. Amin, S.H. Islam, P. Gope, Efficient authentication protocol for secure multimedia communications in IoT-enabled wireless sensor networks, Multimedia Tools Appl. 77 (2018) 18295–18325.
[96] H. Li, F. Li, C. Song, Y. Yan, Towards smart card based mutual authentication schemes in cloud computing, KSII Trans. Int. Inf. Syst. (TIIS) 9 (7) (2015) 2719–2735.
[97] N. Park, N. Kang, Mutual authentication scheme in secure internet of things technology for comfortable lifestyle, Sensors 16 (1) (2015) 20.
[98] C.H. Bennett, G. Brassard, Quantum cryptography: Public key distribution and coin tossing, 2020, arXiv preprint arXiv:2003.06557.
[99] P.K. Panda, S. Chattopadhyay, A secure mutual authentication protocol for IoT environment, J. Reliable Intell. Environ. 6 (2020) 79–94.
[100] M.S. Sharbaf, Quantum cryptography: a new generation of information technology security system, in: 2009 Sixth International Conference on Information Technology: New Generations, IEEE, 2009, pp. 1644–1648.
[101] K. Dong, Y. Cao, Q. Yang, S. Zhang, H. Xing, Q. Ren, Role of hydrogen bonds in ionic-liquid-mediated extraction of natural bioactive homologues, Ind. Eng. Chem. Res. 51 (14) (2012) 5299–5308.
[102] V. Kalaivani, et al., Enhanced BB84 quantum cryptography protocol for secure communication in wireless body sensor networks for medical applications, Pers. Ubiquitous Comput. (2021) 1.
[103] M.S. Sharbaf, Quantum cryptography: An emerging technology in network security, in: 2011 IEEE International Conference on Technologies for Homeland Security (HST), IEEE, 2011, pp. 13–19.
[104] E. Rodríguez, J. Arqués, R. Rodríguez, M. Nuñez, M. Medina, T. Talarico, I. Casas, T. Chung, W. Dobrogosz, L. Axelsson, et al., We are IntechOpen, the world's leading publisher of open access books built by scientists, for scientists TOP 1%, Intech 32 (tourism) (1989) 137–144.
[105] V. Hassija, V. Chamola, V. Saxena, V. Chanana, P. Parashari, S. Mumtaz, M. Guizani, Present landscape of quantum computing, IET Quant. Commun. 1 (2) (2020) 42–48.
[106] V. Teja, P. Banerjee, N. Sharma, R. Mittal, Quantum cryptography: state-of-art, challenges and future perspectives, in: 2007 7th IEEE Conference on Nanotechnology (IEEE NANO), IEEE, 2007, pp. 1296–1301.
[107] A. Lamas-Linares, C. Kurtsiefer, Breaking a quantum key distribution system through a timing side channel, Opt. Express 15 (15) (2007) 9388–9393.
[108] L. Xu, M. Wang, A QDS scheme based on superdense teleportation, Quantum Inf. Process. 21 (6) (2022) 220.
[109] T. Zhou, J. Shen, X. Li, C. Wang, J. Shen, Quantum cryptography for the future internet and the security analysis, Secur. Commun. Netw. 2018 (2018) 1–7.
[110] M.A. Ghonaimy, An overview of quantum information systems, in: 2013 8th International Conference on Computer Engineering & Systems (ICCES), IEEE, 2013, pp. xx–xxxii.
[111] W.H. Wong, Timing attacks on RSA: revealing your secrets through the fourth dimension, XRDS: Crossroads, the ACM Magazine for Students 11 (3) (2005) 5.
[112] Z.A. Abdulkader, et al., A secure IoT system using quantum cryptography with block cipher, J. Appl. Sci. Eng. 24 (5) (2021) 771–776.
[113] A.B. Dorothy, S.B.R. Kumar, An approach for iot security using quantum key distribution.
[114] J. Li, N. Li, Y. Zhang, S. Wen, W. Du, W. Chen, W. Ma, A survey on quantum cryptography, Chin. J. Electron. 27 (2) (2018) 223–228.
[115] M. Moizuddin, J. Winston, M. Qayyum, A comprehensive survey: quantum cryptography, in: 2017 2nd International Conference on Anti-Cyber Crimes (ICACC), IEEE, 2017, pp. 98–102.
[116] A.A. Abd El-Latif, B. Abd-El-Atty, S.E. Venegas-Andraca, H. Elwahsh, M.J. Piran, A.K. Bashir, O.-Y. Song, W. Mazurczyk, Providing end-to-end security using quantum walks in IoT networks, IEEE Access 8 (2020) 92687–92696.
[117] M. Niemiec, Error correction in quantum cryptography based on artificial neural networks, Quantum Inf. Process. 18 (6) (2019) 174.
[118] V. Hassija, V. Chamola, V. Saxena, D. Jain, P. Goyal, B. Sikdar, A survey on IoT security: application areas, security threats, and solution architectures, IEEE Access 7 (2019) 82721–82743.
[119] A. Kumar, S. Garhwal, State-of-the-art survey of quantum cryptography, Arch. Comput. Methods Eng. 28 (2021) 3831–3868.
[120] D.J. Bernstein, Introduction to Post-Quantum Cryptography, Springer, 2009.
[121] N. Göttert, T. Feller, M. Schneider, J. Buchmann, S. Huss, On the design of hardware building blocks for modern lattice-based encryption schemes, in: Cryptographic Hardware and Embedded Systems–CHES 2012: 14th International Workshop, Leuven, Belgium, September 9-12, 2012. Proceedings 14, Springer, 2012, pp. 512–529.
[122] Z. Liu, K.-K.R. Choo, J. Grossschadl, Securing edge devices in the post-quantum internet of things using lattice-based cryptography, IEEE Commun. Mag. 56 (2) (2018) 158–162.
[123] R. Niederhagen, M. Waidner, Practical post-quantum cryptography, Fraunhofer SIT (2017).
[124] A. Boorghany, S.B. Sarmadi, R. Jalili, On constrained implementation of lattice-based cryptographic primitives and schemes on smart cards, ACM Trans. Embed. Comput. Syst. (TECS) 14 (3) (2015) 1–25.
[125] R. Overbeck, N. Sendrier, Code-based cryptography, Post-Quant. Cryptogr. (2009) 95–145.
[126] S. De Capitani di Vimercati, S. Foresti, G. Livraga, P. Samarati, Practical techniques building on encryption for protecting and managing data in the cloud, in: The New Codebreakers: Essays Dedicated to David Kahn on the Occasion of His 85th Birthday, Springer, 2016, pp. 205–239.

[127] G.C. Pereira, C. Puodzius, P.S. Barreto, Shorter hash-based signatures, J. Syst. Softw. 116 (2016) 95–100.
[128] S. Suhail, R. Hussain, A. Khan, C.S. Hong, On the role of hash-based signatures in quantum-safe internet of things: Current solutions and future directions, IEEE Internet Things J. 8 (1) (2020) 1–17.
[129] R. Amiri, R. Stárek, D. Reichmuth, I.V. Puthoor, M. Mičuda, L. Mišta Jr., M. Dušek, P. Wallden, E. Andersson, Imperfect 1-out-of-2 quantum oblivious transfer: bounds, a protocol, and its experimental implementation, PRX Quantum 2 (1) (2021) 010335.
[130] L. Malina, L. Popelova, P. Dzurenda, J. Hajny, Z. Martinasek, On feasibility of post-quantum cryptography on small devices, IFAC-PapersOnLine 51 (6) (2018) 462–467.
[131] P. Longa, A note on post-quantum authenticated key exchange from supersingular isogenies, Cryptology ePrint Archive (2018).
[132] A. Broadbent, C. Schaffner, Quantum cryptography beyond quantum key distribution, Des. Codes Cryptogr. 78 (2016) 351–382.
[133] K. Inoue, E. Waks, Y. Yamamoto, Differential-phase-shift quantum key distribution using coherent light, Phys. Rev. A 68 (2) (2003) 022317.
[134] V. Scarani, A. Acin, G. Ribordy, N. Gisin, Quantum cryptography protocols robust against photon number splitting attacks for weak laser pulse implementations, Phys. Rev. Lett. 92 (5) (2004) 057901.
[135] K. Kadian, S. Garhwal, A. Kumar, Quantum walk and its application domains: A systematic review, Comp. Sci. Rev. 41 (2021) 100419.
[136] S. Streit, F. De Santis, Post-quantum key exchange on ARMv8-a: A new hope for NEON made simple, IEEE Trans. Comput. 67 (11) (2017) 1651–1662.
[137] U. Vazirani, T. Vidick, Distribución de clave cuántica totalmente independiente del dispositivo, Phys. Rev. Lett. 113 (140501) (2014) 10–1103.
[138] J. Buchmann, E. Dahmen, A. Hülsing, XMSS-a practical forward secure signature scheme based on minimal security assumptions, in: Post-Quantum Cryptography: 4th International Workshop, PQCrypto 2011, Taipei, Taiwan, November 29–December 2, 2011. Proceedings 4, Springer, 2011, pp. 117–129.
[139] M. Lopes, N. Sarwade, Cryptography from quantum mechanical viewpoint, 2014, arXiv preprint arXiv:1407.2357.
[140] M. Campagna, L. Chen, Ö. Dagdelen, J. Ding, J.K. Fernick, N. Gisin, D. Hayford, T. Jennewein, N. Lütkenhaus, M. Mosca, et al., Quantum safe cryptography and security, ETSI White Paper 8 (2015).
[141] S.M. Safi, A. Movaghar, M. Mohammadizadeh, A novel approach for avoiding wormhole attacks in VANET, in: 2009 Second International Workshop on Computer Science and Engineering, Vol. 2, IEEE, 2009, pp. 160–165.

View publication stats

Physica Scripta

INCLUSIVE PUBLISHING
TRUSTED SCIENCE

PAPER

A single ended, single port configuration based 9 T SRAM cell for stability enhancement

To cite this article: Vansh Singhal *et al* 2023 *Phys. Scr.* **98** 115035

View the article online for updates and enhancements.

You may also like

- A 32 nm single-ended single-port 7T static random access memory for low power utilization
Bhawna Rawat and Poornima Mittal
- Robust transmission gate-based 10T subthreshold SRAM for internet-of-things applications
Erfan Abbasian and Morteza Gholipour
- Zirconium Aided Epitaxial Growth of In_xSe_y on InP(111) Substrates
Cheng Zheng, , Dapeng Zhao et al.

This content was downloaded from IP address 14.139.251.98 on 03/11/2023 at 05:57

Physica Scripta

PAPER

A single ended, single port configuration based 9T SRAM cell for stability enhancement

RECEIVED
5 April 2023

REVISED
13 September 2023

ACCEPTED FOR PUBLICATION
6 October 2023

PUBLISHED
23 October 2023

Vansh Singhal[1], **Bhawna Rawat**[1], **Poornima Mittal**[1] and **Brijesh Kumar**[2,*]

[1] Delhi Technological University, Delhi-110042, India
[2] Indira Gandhi Delhi Technological University, Delhi-110006, India
[*] Author to whom any correspondence should be addressed.

E-mail: singhalvansh30@gmail.com, bhawnarawat12@gmail.com, poornimamittal@dtu.ac.in and brijeshkr@igdtuw.ac.in

Keywords: 9T SRAM, single-ended, static noise margin, write margin, low voltage

Abstract

The growing demand for power efficient devices and high-density memories has pushed researchers to develop low power SRAMs. The main objective for these researches is to reduce power consumption and enhances battery life and scaling of technology node. Consequently, in this paper a 9T SRAM bit cell with enhanced stability and single ended, single port configuration is proposed. The cell is designed and simulated at 180 nm technology node with a voltage supply of 1V. The cell proposed has low power consumption owing to single bitline, higher read stability due to isolated read port, better write margin due to disconnected feedback connection and resistant to soft errors because of half select disturbance free architecture. To assess the performance of the proposed cell its performance is compared against existing 6T, 8T, 9TST, SB 9T, TRD 9T, and NTV 9T bit cells. The HSNM (RSNM) and WM values for the proposed cell are equal to 364 mV and 378 mV respectively. The cell is designed to be half select disturbance free and supports bit interleaving. The reliability of the proposed cell is further analysed for local, global and temperature variation. While, the area footprint for the cell is 24.91 μm^2.

1. Introduction

Recent advancements in AI, such as facial recognition and video recognition, have significantly eased the burden on human intelligence. These improvements have been largely driven by neural networks, which utilize massive datasets to train their processors leading to enhanced accuracy [1]. These processors incorporate an on-chip random access memory (RAM) which necessitates low dynamic power and high-speed static RAM (SRAM). Researchers have developed robust machine learning algorithms which - because of their extensive utilization, necessitates devices with prolonged battery endurance [2, 3]. Thus, battery life of a gadget is a crucial concern. So, it is imperative to ensure low power consumption while designing a system-on-chip (SoC). With a rising demand for devices with faster operation and larger memory capacity, there is an urgent need to mitigate power consumption and boost reliability while enhancing speed for SoC. Memory is an important component of most SoCs, and its performance is strongly dependent on its memory performance. Cache memory is a major component of every system; it is formed by arranging SRAM bit cells in an array configuration [4, 5]. The quantity of embedded SRAM in any modern-day microprocessor system and recent SoCs has improved over the years so as to fulfil the performance requirements in emerging technological advancement [6]. Our proposed design, characterized by a single-ended architecture, exhibits a low dynamic power consumption, making it highly suitable for low power applications. In any microprocessor, it is estimated that SRAM cells roughly consume 30% of the total power requirement [7, 8]. Dynamic and static power consumption make up the majority of an SRAM bit cell's power requirements. Static power consumption exhibits a linear dependency on supply voltage (V_{DD}) [9, 10], while dynamic power consumption exhibits a quadratic dependency on V_{DD} [11].

The industry has been compelled to move toward nano-meter technology size range, in order to fulfil consumer demand for sleek designs with greater battery life. In order to meet the growing demand for greater

© 2023 IOP Publishing Ltd

battery life, it is essential to ensure that power consumption in a bit cell is kept considerably low. Reducing V_{DD} for a cell is the most evident way to minimize power usage. But reducing V_{DD} in the deep sub-micrometer region is a challenging problem owing to the process voltage temperature (PVT) variations [12, 13]. Downscaling V_{DD} results in an increase in operational delay, which increases total energy per read/write cycle significantly [14]. There is a certain limit to how far V_{DD} can be minimized, because doing so will significantly raises the bit error rate while also resulting in a slowdown in memory operations [15]. The (V_{DD}-V_{TH}) factor, where V_{TH} is the transistor threshold voltage, decreases with a decrease in V_{DD}. As the V_{DD}-V_{TH} decreases, noise margin significantly degrades thus, making the circuit susceptible to operational failure [16]. Furthermore, because random dopant fluctuations and random PVT variations are so common in the sub-micrometre domain, the stability of the design has been significantly compromised [17, 18]. As a result, cache memories, comprised of SRAM cells, are probably more susceptible to these failures [19]. Consequently, the implementation of conventional 6 T cell is getting more difficult for low-voltage devices as the 6 T bit cell fails for scaled V_{DD} due to excessive leakage and significant static power loss [20]. These variations lead to degradation in the memory performance.

The combined effect of the above-mentioned factors in the conventional cell has resulted in an increase in the occurrence of failure during hold, read and write operation. As a result, various cell configurations have been reported over the years to address the aforementioned limitations. Various techniques are used to enhance noise margin and minimize power consumption in memory cell design. Improvement in read stability is ensured [21] by employing a Schmitt-trigger, that alters the switching threshold of an inverter resulting in greater read stability. A significant improvement in read stability is proposed in [22, 23], which isolates cross coupled inverter from read bitline. But these exhibit high power dissipation. The adoption of single ended bitline structure is most effective in minimizing the consumption of dynamic power as the activity factor is reduced to half, due to its single bitline structure [24].

The 9 T cells in [25, 26] and 7 T cell in [24] use an additional transistor inside the core of bit cells to enable/disable feedback connection of two inverters, which significantly improve their write margin (WM). In the absence of any decoupling device, bit cells in [24, 25] may suffer from frequent read upsets. The 9 T bit cell in [27] also exhibits improvement in read stability by employing read-decoupling mechanism. The bit cells in [28, 29] show improvement in write stability by using asymmetrical inverter sizing in the core. However, when exposed to PVT variations, these bit cells are quite vulnerable to failure. A single-ended 7 T cell free from read disturbances, was first proposed in [30], but it suffers from high chances of write failure. Some researchers [31, 32] have even used different devices for performance enhancement for bit cell.

Here in this paper, a single-ended nine transistors (9 T) SRAM bit cell is proposed with the aim of overcoming the above-mentioned limitations. The proposed design has the following prominent characteristics. The design greatly boosts the read stability due to the formation of isolated read port and simultaneously improves the write ability for the cell owing to the presence of a MOS transistor between the feedback connection of two cross-coupled inverters. The existence of single bitline in the proposed architecture results in significant reduction in its consumption of dynamic power. The stacking effect of the transistors in the discharge path results in significant reduction in consumption of leakage power. This paper is segmented into the following sections. The detailed functioning of proposed 9 T cell is outlined in section 2. The literature review of pre-existing SRAM bit cell topologies, including their pros and cons, is detailed in section 3. Section 4 presents comparison of static performance of proposed cell against other cells. Section 5 addresses the impact of temperature variation on all cells. In section 6 the issues related to half select disturbances and soft errors experienced by the proposed cell are explained. The impact of process variation on the proposed cell is examined in section 7. The read and write time analysis is explained in section 8, followed by power consumption analysis in section 9. The layout and area for the bit cells are elaborated in section 10. In section 11 the electrical quality metric is explained. Finally, the conclusive statements are summed up in section 12.

2. Proposed 9T cell and its functioning

In this paper, a single bitline SRAM bit cell with nine transistors (9 T) is proposed. The memory core for the cell is depicted in figure 1. The memory core of this design comprises of two cross-coupled inverters, INV1 (N1-P1) and INV2 (N2-P2). The cross-coupled feedback loop is disconnected by a MOS gate N3, which is placed between INV2 and INV1. The feedback connection between the two inverters is enabled/disabled using N3 transistor controlled via signal WE. The access transistor N4 gets turned ON by the wordline WL, whereas the write access transistor N5 gets turned ON by the chip-select-line CSL. For a static noise margin (SNM) free read operation, an isolated read port is created via two additional stack transistors (N6-N7). The read decoupling transistor N6, positioned between the nodes Q2 and X1, is gated by node QB (output of N1-P1 inverter). During read operation, N7 is activated using a read wordline RWL. The memory core of the proposed architecture can be

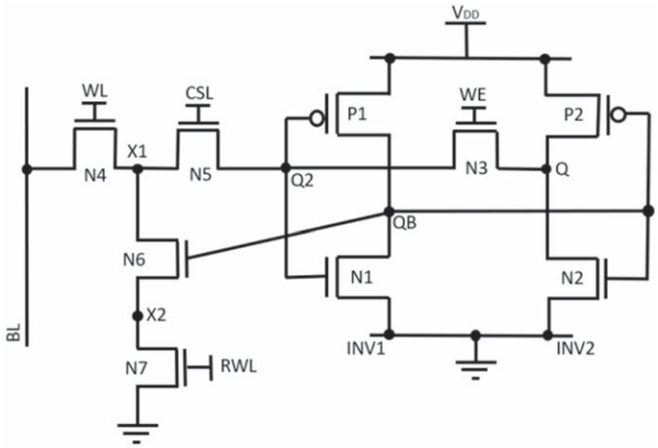


Figure 1. Circuit diagram for the proposed 9 T SRAM bit cell.

Table 1. Signal combination for proposed 9T SRAM bit cell.

	WL	CSL	WE	BL	RWL
Hold	GND	GND	V_{DD}	GND	GND
Read	V_{DD}	GND	V_{DD}	Pre-charged	V_{DD}
Write −1	V_{DD}	V_{DD}	GND	V_{DD}	GND
Write—0	V_{DD}	V_{DD}	GND	GND	GND

accessed via transistors N4 and N5 during write operation and via transistor N6 during read operation. The truth table corresponding to the different operations for the proposed 9T bit cell is given below in table 1.

2.1. Hold operation

During this operation, the cell needs to preserve the data value stored in it. The access signals - RWL, CSL and WL are maintained at value '0' while transistors (N4, N5, N7) are maintained in OFF state. The transistor N3 is turned ON by keeping WE input at V_{DD} to establish the feedback connectivity between the two inverters in the memory core. During this mode, the cell only retains the data, while the read and write ports for the cell are disabled. Therefore, this operation observes the least amount of power consumption. As the cell is in static mode, the power consumption by the cell in the hold mode is referred to as the leakage power consumption. For the proposed 9T cell the power consumption during the hold operation is obtained to be 5.46 nW.

2.2. Read operation

During the read operation, the feedback connection between INV1 and INV2 (between node Q and Q2) is ensured by maintaining the feedback cutting MOS gate in ON state (set WE signal to V_{DD}) so that data should be preserved. The CSL input is set to GND, which isolates the pre-charged bitline BL from the memory core of the cell. Transistors N7 and N4 are turned ON via RWL and WL signals respectively. Bitline (BL) is precharged to $V_{DD.}$ The gate of transistor N6 is coupled with node QB. Bitline will therefore either be discharged or remain charged depending on the data value of internal node 'QB as shown in figures 2(a) and (b). Consider node QB stores logical '1' which turned ON the transistor N6. This completes the path consisting of transistors (N4, N6, N7) and BL will start discharging to ground via this path. For successful read operation, discharge in the BL voltage will be sensed by sense amplifier to generate output. The read discharge current for the proposed cell is registered only when Q = '0' and QB = '1'. The read discharge current path for Q = '0', and Q = '1', for the proposed cell is depicted in figures 2(a) and (b), respectively.

During the read '1' operation, the QB node equals '0', thus resulting in N6 transistor being OFF. This disrupts the read discharge path, thus for QB = '0', no read discharge current is registered. Whereas, for Q = '0', the QB = '1' this connects the BL to ground through transistor N6, and N7 transistor. Thus, read discharge current is registered. Consequently, the pre-charge voltage on the BL registers a drop. The voltage drop in BL is expected to be low enough to drive the read sense amplifier. The drop in BL voltage during the read '0' operation is depicted in figure 2(c). It can also be inferred from figure 2(c), that a pulse width in 70 ps is required by the proposed bit cell. Also, as read operation will result in discharge current only corresponding to Q = '0', and not for Q = 1. The power consumption for the proposed 9T cell is also reduced, as read power is consumed only for

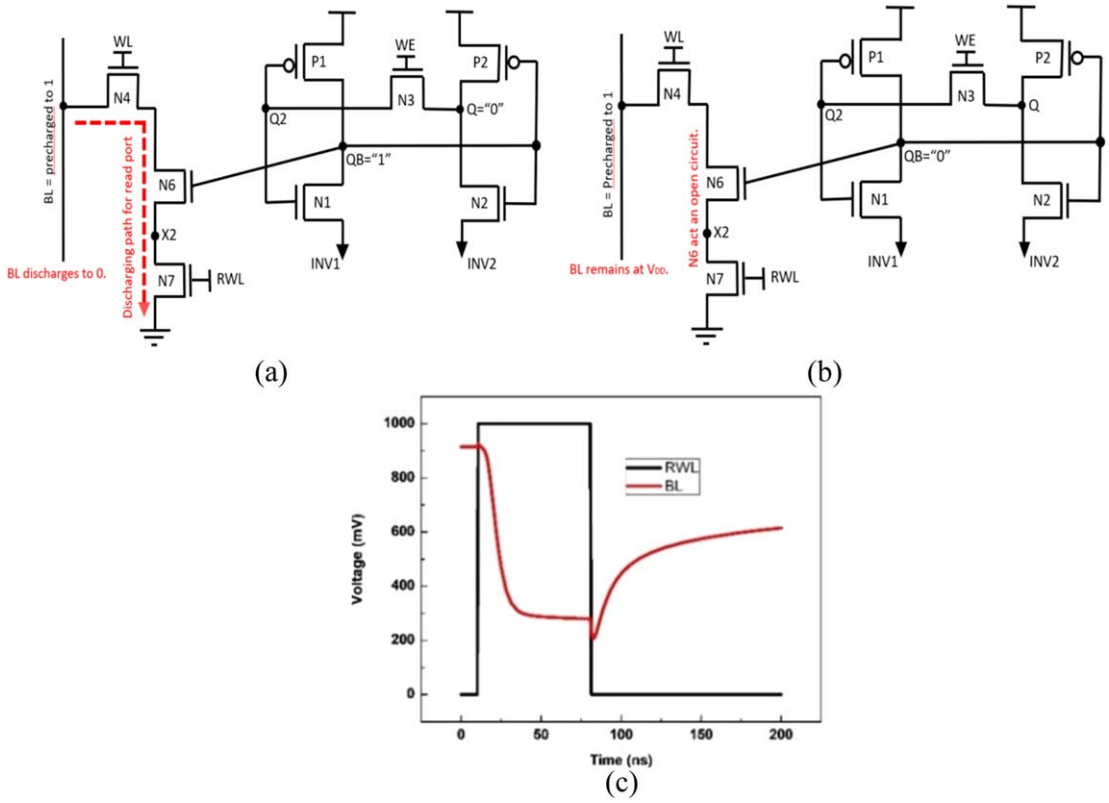


Figure 2. Effective circuit during read operation for the proposed 9 T SRAM bit cell (a) read 0 operation,(b) read 1 operation, and (c) voltage drop in BL during read '0' operation.

Q = '0' and for Q = '1' only a small amount of leakage power is registered for the circuit. Thus power consumed by the cell during read '0', and read '1' is obtained to be 3.67 μW, and 12.68 pW respectively.

2.3. Write operation

The feedback loop between the nodes Q2 and Q is interrupted by keeping WE signal low during the onset of write operation, which in turn disables the N3 transistor. Thus, INV2 is totally driven by INV1. The desired data value is then placed onto BL. Thereafter, N5 and N4 transistors are turned ON via CSL and WL signals respectively and N7 is turned OFF by maintaining RWL signal low as shown in figure 3(a). This allows transfer of data from bitline BL to node Q2. Q2 then drives the inverter P1-N1 to generate the complementary value of input data at node QB, which in turn drives the inverter P2-N2 to generate the desired data value at node Q. BL is changed to '1'/'0' in accordance with the operation.

2.3.1. Write-1 operation

Initially node Q/QB stores '0'/'1'. BL is set to V_{DD}. When CSL and WL signals are enabled, voltage at node Q2 starts rising. The rise in voltage of node Q2 will further drive the output voltage of INV1 to ground. As node QB is directly connected to the input of INV2, the fall in voltage of node QB will further drive the output of INV2 to V_{DD}. Thus, voltage at node QB becomes '0'. This in turn makes node 'Q' as '1'. The successful write 1 operation for the proposed cell is depicted in figure 3(b). The pulse width required for the proposed 9T cell for write '1' operation is determined to be 0.7 ns.

2.3.2. Write-0 operation

Initially node Q/QB stores '1'/'0'. BL is set to GND. As WL and CSL signals are enabled, voltage at node Q2 starts decreasing. The fall in voltage of node Q2 will further drives the output voltage of INV1 to V_{DD}. As node QB is directly connected to the input of INV2, the rise in voltage of node QB will further drive the output of INV2 to ground. Thus, voltage at node QB becomes '1'. This in turn makes node 'Q' as '0'. The successful write '0' operation for the proposed cell is depicted in figure 3(c). It can be inferred from figures 3(b), and (c) that the write pulse width needed for a successful write operation is 0.7 ns.

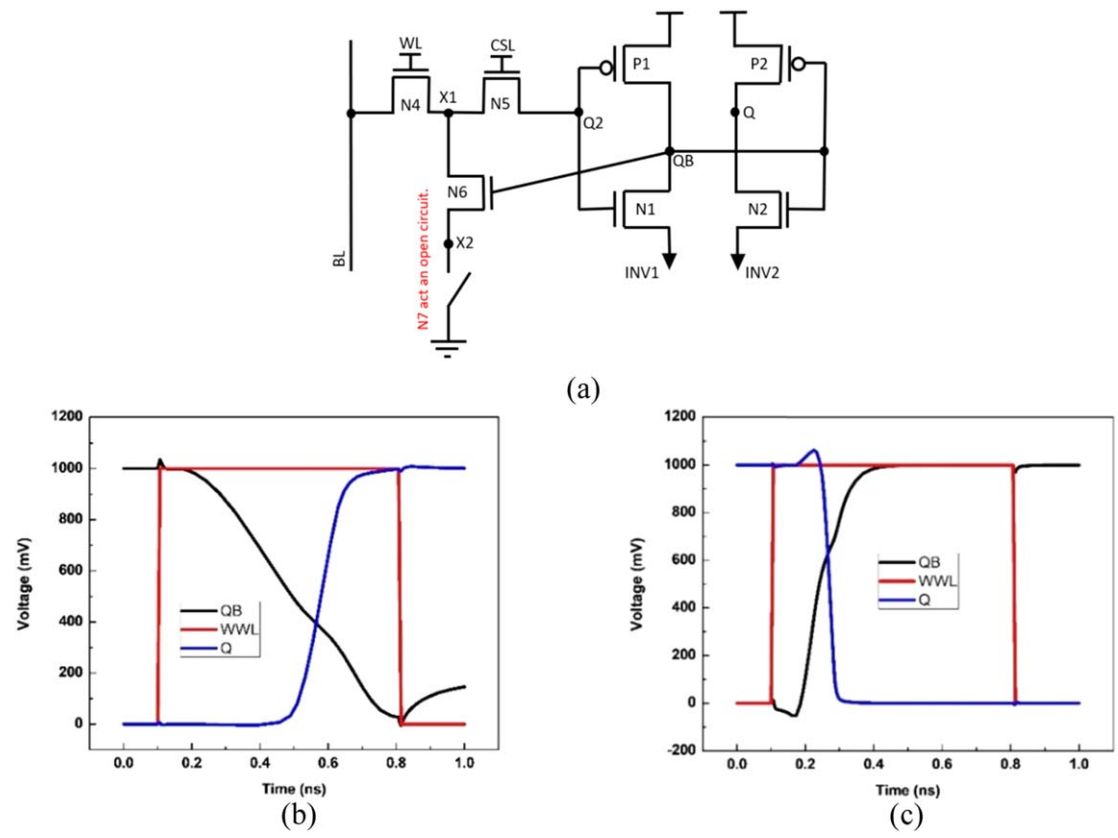


Figure 3. (a) Effective circuit during write operation, (b) write 1 operation, and (c) write 0 operation for the proposed 9T cell.

It may be observed for figures 3(b), and (c) that write '0' occurs faster in comparison to write '1' operation for the bit cell. This is an implication of the single ended write nature for the proposed cell. But, pulse width is maintained same for the two operations as in a single the operation clock pulse is defined globally and cannot be altered as per requirement of each operation. But, this does impact the write power consumption for the bit cell and higher power is consumed for write '1' in comparison to write '0'. The former consumes 9.55 μW power, whereas the latter consumes 7.23 μW of power.

3. Static margin analysis for the proposed cell

The stability of the proposed 9 T cell for the hold and read operation is measured with the help of the butterfly curve. The butterfly curve for a bit cell is drawn by overlapping the voltage transfer characteristics for the inverters that form the inverter core for the cell. Additionally, the read port for the proposed cell is designed to exclude the data node from the read discharge path. Such cells that isolate the data node from the read discharge current path are referred to as read SNM free cell [33]. Thus, the read SNM (RSNM) for the cell is equal to its hold SNM (HSNM). The hold and read butterfly curve for the proposed 9T cell are overlapping and are depicted in figure 4(a). The HSNM, and RSNM value for the proposed 9T cell is simulated to be 364 mV. While, the stability of a cell during the write operation is measured in terms of write margin (WM). WM can be computed as the voltage difference between supply voltage (V_{DD}) and voltage at WL when data stored in internal node get inverted. It is estimated via the WL sweep technique suggested by Hasan and Islam [34]. The WM value for the proposed cell is simulated to be 378 mV. Whereas, the WM curve for the proposed cell is depicted in figure 4(b).

The recent advancements in AI, such as facial recognition and video recognition, have significantly eased the burden on human intelligence. These improvements have been largely driven by neural networks, which utilize massive datasets for to train their processors and have greatly enhanced accuracy. These processors incorporate an on-chip RAM which necessitates low dynamic power and high-speed SRAM. Researchers has developed more robust machine learning algorithms which, because of their extensive utilization, necessitates devices with prolonged battery endurance. The proposed 9T cell design, characterized by a single-ended architecture, exhibits improvement in performance (explained subsequently), making it highly suitable for these low power applications that require high packaging density for memories.

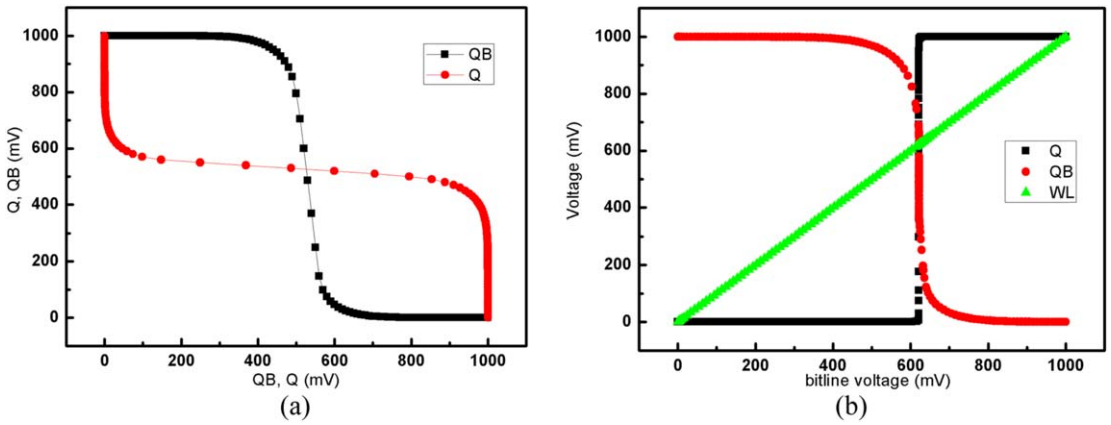


Figure 4. (a) Butterfly curve obtained for 9 TP cell during read (hold) operation and (b) WM curve, for proposed 9TP cell.

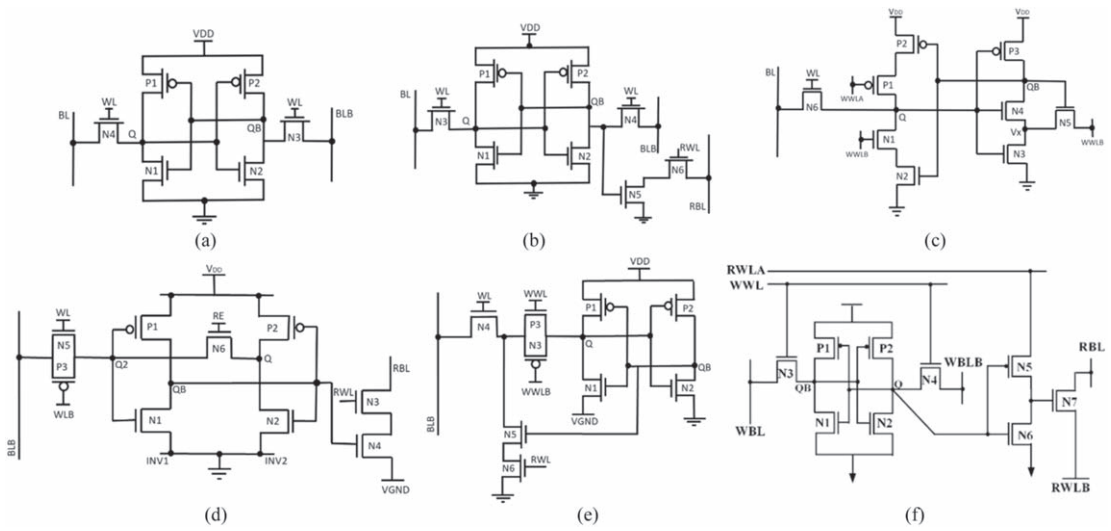


Figure 5. Schematic Diagram for (a) 6T, (b) 8T, (c) 9T ST, (d) TRD 9T, (e) SB 9T, and (f) NTV 9TSRAM bit cells.

4. Comprehensive description of different pre-exisiting bitcells

Over the years, many different SRAM bit cells have been presented to tackle the different irregularities observed in the performance of the conventional 6T SRAM cell encountered because of low voltage operation and declining technology node. Some of them are reviewed in this section and a comparative study of these cells is presented and their performance is then compared with the proposed design.

4.1. 6T SRAM cell

The 6 T SRAM bit cell majorly comprises of two identical inverter pairs (N1-P1) and (N2-P2), as shown in figure 5(a). The data core for the cell formed by two cross-coupled inverter pair is accessed by two NMOS access transistors (N4-N3). However, in this conventional 6 T cell design, the bit lines are directly coupled to the storage nodes via access transistors during the read operation. The nodal voltage of the storage node that holds logical '0' start rising during read operation, because of the voltage division act along the access transistors (N3, N4) and pull down transistors (N2, N1), making the cell very susceptible to noise. This may result in a destructive read operation of the cell. This error sensitivity of 6T cell is further aggravated with reduction in V_{DD} and technology node.

4.2. 8T SRAM Cell

The simplest mechanism to elevate the performance of the 6T bit cell is to isolate its read port from write port [35, 36]. The same was done by researchers in designing the 8T cell, as depicted in figure 5(b). An isolated read

port is created by means of two additional stack transistors (N5, N6). A separate read bitline (RBL) is used to access the memory core of the cell. To ensure read stability, the read port is designed with read SNM free configuration. The data of internal node of 8T cell is not affected while performing read operation, thus resulting in higher RSNM value for the cell. Additionally, the stacked transistor configuration of the read port aids in minimising bitline leakage [37]. A major issue with this topology is its high power dissipation due to dual port configuration.

4.3. 9T ST SRAM Cell

The core of 9T SRAM cell is composed of a back-to-back inverter pair, connected in feedback loop, formed by a Schmitt Trigger (ST) based inverter (P3, N3, N4, N5) and a stacked transistor-based inverter (N1, N2, P1, P2), and a NMOS access transistor (N6), as shown in figure 5(c). The switching threshold of a Schmitt trigger-based inverter can be changed depending on which way the input signals are transitioning. The conflict between read stability and write stability is a significant designing bottleneck with SRAM bit cell. A SRAM cell must maintain a high bit cell ratio to ensure a reliable read operation. Instead, this reduces the write ability by weakening the access transistor, leading to rise in write access time. Likewise lowering the pull-up ratio will boost the write performance but at the expense of read performance. But, this cell is able to bridge the gap between read and write performance for the cell with the help of Schmitt-Trigger based inverter.

4.4. TRD 9 T SRAM Cell

A transmission gate based 9T cell is depicted in figure 5(d). The core of the transmission gate based 9T cell includes two CMOS inverters connected in a feedback loop. The feedback path for the cell is interrupted by a MOS transistor (N6); controlled by RE signal, to facilitate write operation. This design is based on single bitline, which accounts for low dynamic power dissipation. Its architecture is dual port, where write operation is initiated by bitline (BLB) through transmission gate (P3 and N5) and read operation is performed by separate isolated port, consists of N3 and N4, which alleviate read stability and solves the problem of access transistor sizing. Virtual ground signal (VGND) is used to mitigate leakage current through read decoupled path by leaving the source of N4 floating. The use of transmission gate increases the area footprint for the bit cell.

4.5. SB 9T SRAM Cell

The single bitline nine transistor (SB 9T) SRAM cell comprised of two identical inverters (P1-N1 and P2-N2) connected in a feedback loop. The schematic for the SB 9T cell is depicted in figure 5(e). The read path consists of three transistors (N4, N5, and N6) while write path comprised of a write access transistor N4 and a transmission gate (P3-N3). The source of transistor N1 is left floating (VGND) to facilitate write '1' operation which weakens the pulling effect of NMOS. The WWL signal is de-asserted during read operation which isolate the memory core from bitline thus making RSNM same as HSNM. During write operation, WL and WWL signals are asserted to transfer the bitline data to the internal node. The utilization of transmission gate resolves the issue of access transistor sizing during read/write operation and allows independent adjustment of pull-up and bit-cell ratio.

4.6. NTV 9T SRAM Cell

A cell arrangement having nine transistors (NTV 9T) was proposed with an L-shaped layout, different from the usual rectangle shape [38]. To avoid the sharing of RWLA signal among the neighbouring cells, the layout for the cell is asymmetric. It has two inverters, cross coupled with each other, just like in a conventional 6T bit cell. It also consists of write-access transistors (N3-N4) and a separate read port having transistors—N5, N6, and N7. Few additional row-based signals are used in the circuit which are Write Wordline (WWL), Read Wordline A (RWLA) and Read Wordline B (RWLB). An additional inverter connected to the Q node is formed by the transistors N5 and N6. The cell uses transistors—N5, N6, and N7 to dissociate RBL from the storage node. This SRAM cell can have a notable increase in the sensing delay, when compared with the other cells having a separate read port, due to its enhanced current through the cell [38]. For read operation of the cell, RWLA and RWLB signals are enabled whereas for the write operation, signal WWL is activated.

5. Static analysis

All the cells in this paper are designed and simulated for 180 nm technology node using the predictive technology models. The cells are simulated at 180 nm technology to ensure a common platform for a fair comparison between the cells. Also, the V_{DD} for the cell is maintained at 1 V. The first metric for determining stability, noise immunity and robustness of a SRAM bit cell is the SNM for read, write and hold operations [39]. SNM is commonly defined as the maximum amount of DC noise voltage that a cell can withstand before the data

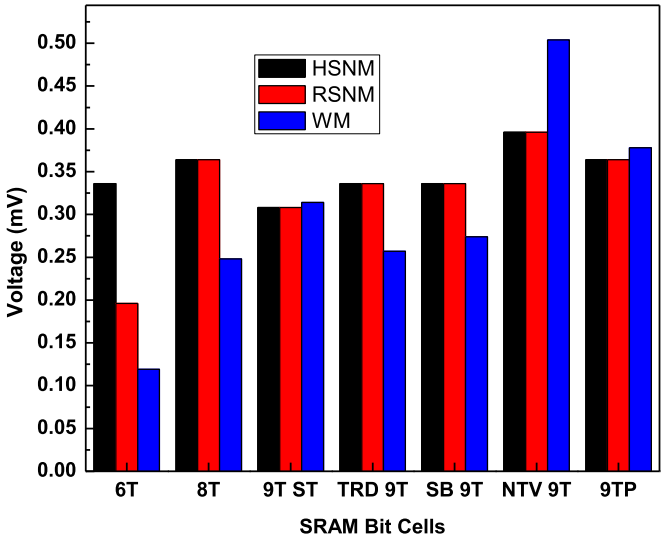


Figure 6. Graphical comparison for HSNM, RSNM & WM values for different SRAM cells WM.

Table 2. Hold SNM, Read SNM and WM for different SRAM bit cells.

	HSNM	RSNM	WM
6T	0.336	0.196	0.119
8T	0.364	0.364	0.248
9T ST	0.308	0.308	0.314
TRD 9T	0.336	0.336	0.257
SB 9T	0.336	0.336	0.274
NTV 9T	0.396	0.396	0.504
9T Proposed	0.364	0.364	0.378

value stored in the bit cell gets flipped. The SNM is used to assess the bit cell's stability and noise tolerance during read, write and hold operation [40]. The voltage transfer characteristics of the both inverters are graphically superimposed to give a two-lobe butterfly curve, which is used to assess SNM [41]. The side of the biggest square that can fit inside the smaller lobe of butterfly curve is used to calculate SNM's value.

5.1. HSNM

The initial analysis is to ensure hold stability, which is determined using the HSNM. It is an important parameter as for major part of its operation, the SRAM cell remains in hold state. A graphical comparison of the HSNM values obtained for the different SRAM bit cells is depicted in figure 6. While, a summary of the quantitative HSNM results obtained for the various bit cells is given in table 2. The HSNM values for the proposed 9T cell is 364 mV, which shows an improvement of 8.34%, 18.18%, 8.34% and 8.34% in HSNM value as compared to 6T, 9T ST, TRD 9T and SB 9T cells respectively. The only cell with HSNM value higher than the proposed cell is the NTV 9T cell, with HSNM of 0.396 mV. While the HSNM value for the 8T cell is the same as the proposed cell. Thus, on the basis of HSNM the performance of the proposed cell may be deemed improved in comparison to most its counterparts.

5.2. RSNM

RSNM is considered to be a key factor to assess read stability of a cell. Read stability is the ability of a bit cell to preserve the data stored inside its memory core at the time of read operation. The pre-charged bitline is discharged or not, based upon the data stored in a cell. RSNM for a conventional cell is measured by introducing noise source at the internal node. This is done to check the extent to which the voltage at data node can be altered before an erroneous flip in data node value is observed. But, erroneous flipping of data during the read operation is possible only for cells which have read discharge current passing through the data node. In that case, due to the discharge current passing through the data node, its voltage gets altered, which may bias the inverters on trip point resulting is erroneous data flip. Such cells tend to have RSNM lower than their HSNM. But, the proposed

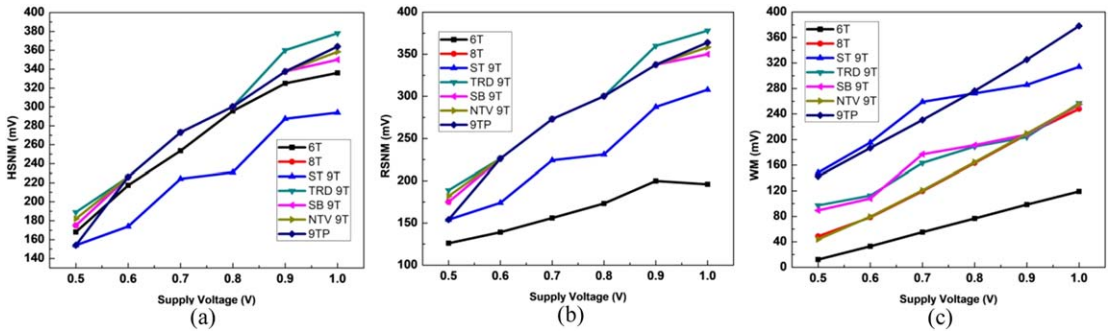


Figure 7. Graphical comparison for the HSNM, RSNM, and WM values obtained for the different SRAM bit cell for different values of V_{DD}.

9T cell is designed to not include the data node in the read discharge current path. Thus, during the read operation the voltage on the data node is not altered. For the proposed 9T cell, the data node QB only turns ON transistor N6. To steer the cell into read operation RWL signal is set high and if at that instant QB = '1'. The read discharge current flows from BL to ground through transistors N4, N6, and N7. The representation for the same is depicted in figure 2(a). Whereas, the lowering of BL voltage during the read operation for QB = '0' is depicted in figure 2(c). While, for QB = '0', the N6 transistor remains OFF and no read discharge current is registered. This is depicted in figure 2(b). But for neither of the two read operations corresponding to QB = '0' or '1', does the data node value change thus, ensuring the read integrity for the bit cell. Cells with such a read port are referred to as read-SNM free cells [33].

The RSNM values obtained for the different cells are compared in figure 6 and values are summarized in table 2. The RSNM value of the proposed cell is 364 mV, which shows an improvement of 85.71%, 18.18%, 8.34% and 8.34% in RSNM values as compared to 6T, 9T ST, TRD 9 T and SB 9T cells respectively. Major drawback with the 6T cell is high noise interference, which is improved by employing a Schmitt trigger (in 9T ST cell) and further enhanced by adopting port isolation technique in TRD 9T, SB 9T and conventional 8T cell. Thus, the use of port isolation technique has greatly enhanced static margin for the proposed cell and enables the cell to outperform all other cells.

The write operation is basically the flipping of the current data value stored in memory cell. The data value is placed onto BL and thereafter it gets transferred to the internal node using access transistors. Write margin (WM) is considered to be a reliable parameter to assess the cell's stability, while executing write operation. WM can be computed as the voltage difference between V_{DD} and voltage at WL when data stored in internal node get inverted. It is estimated via the WL sweep technique suggested by Hasan and Islam [34]. The WM values for all the cells are then summarised in table 2 and compared in figure 6. The write trip point voltage for the proposed cell is 622 mV, as is depicted in figure 4(b). Thus, WM for the proposed cell is estimated as the difference of V_{DD}, which is 1 V, and write trip point voltage which is estimated from figure 4(b) to be 622 mV. The WM value for proposed cell is thus calculated to be 378 mV, which shows an improvement of 217.64%, 52.41%, and 20.38% in WM value as compared to 6T, 8 T, 9T ST cell respectively. The disconnected feedback connection between two inverters has enhanced the write margin and enables the cell to outperform all other cells. The proposed 9T cell has better performance in all three parameters (HSNM, RSNM, and WM) as compared to other cells. The HSNM, RSNM, and WM for the 9T cell are 364, 364, 378 mV respectively. Thus, the 9T cell proved to be most immune to noise interference among all the cells being simulated. The static noise margin values for the different SRAM bit cells are measured at V_{DD} values ranging from 0.5 V to 1 V. This is done to analyze the impact of voltage decline on the performance of the bit cells. The graphical comparison of the HSNM, RSNM and WM values obtained for the different SRAM bit cells for V_{DD} ranging from 0.5 V to 1 V is presented in figures 7(a)–(c), respectively. It can be observed from the curves in figure 7 that the 6T cell has the weakest performance in comparison to the other bit cells. It is followed by the 7TST cell that also records an inferior performance for all values of V_{DD}. The proposed 9T cell on the other hand registers high HSNM, RSNM, and WM values for all V_{DD} values. This further validates that the proposed 9T cell registers a reliable performance against noise throughout its operation.

6. Temperature variation analysis

In a microprocessor, cache memory is considered to be a crucial component and its stability and reliability must be assured across a wide range of temperature because of external or internal influences. Thus, a cell must sustain

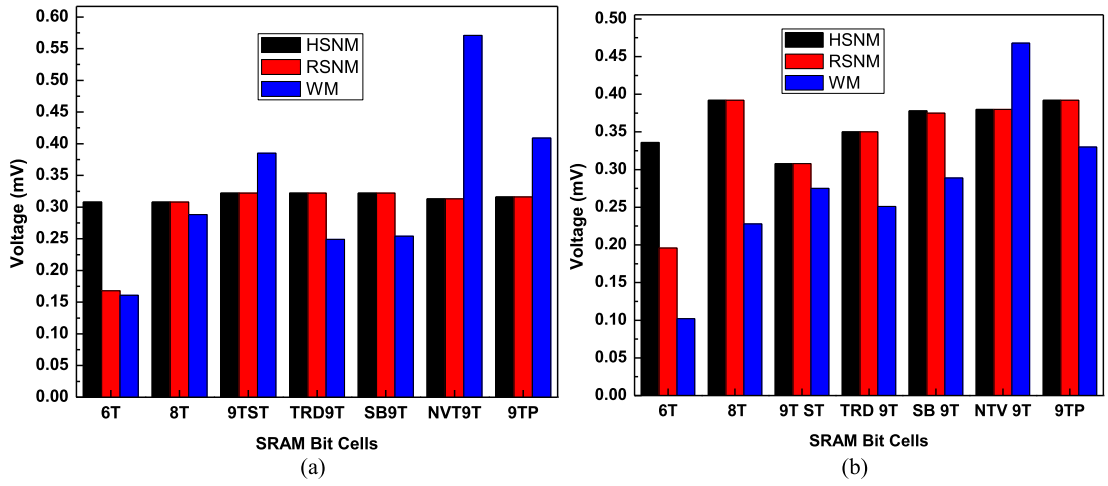


Figure 8. Comparison of SNM values (HSNM, RSNM, WM) obtained for each SRAM cell (a) at 100 °C and (b) at −10 °C.

Table 3. Hold SNM, Read SNM and WM values for all SRAM cells at −10 °C.

Bit Cells	HSNM	RSNM	WM
6 T	0.336	0.196	0.102
8 T	0.392	0.392	0.228
9 T ST	0.308	0.308	0.275
TRD 9 T	0.35	0.35	0.251
SB 9 T	0.378	0.375	0.289
NTV 9 T	0.380	0.380	0.468
9 T Proposed	0.392	0.392	0.330

Table 4. Hold SNM, Read SNM and WM values for all SRAM cells at 100 °C.

Bit Cells	HSNM	RSNM	WM
6 T	0.308	0.168	0.161
8 T	0.308	0.308	0.288
9 T ST	0.322	0.322	0.385
TRD 9 T	0.322	0.322	0.249
SB 9 T	0.322	0.322	0.254
NTV 9 T	0.313	0.313	0.571
9 T Proposed	0.316	0.316	0.409

admirable performance over a wide variety of temperatures. The reliability of the 9T cell is thus demonstrated by analysing it at temperature ranging from −10 °C to 100 °C. The impact of temperature variation on all the cells is analysed in this section. SNM and WM values for each SRAM bit cell at −10 °C and 100 °C are displayed in tables 3 and 4, respectively. In figure 8, a comparison between the SNM values of each cell have been drawn.

The HSNM, RSNM, WM values for the proposed cell are 316 mV, 316 mV, 330 mV respectively at −10 °C and are 392 mV, 392 mV, 409 mV respectively at 100 °C. Thus, it is reasonable to assert that performance of proposed 9T cell is most reliable against temperature variation. In general, it is observed that the HSNM and RSNM values for the different bit cells are high for −10 °C. This is an implication of two major factors—first, at lower temperature the leakage current is also low [42], and second, the mobility of charge carriers (electrons and holes) increases with lowering of temperature [43]. Higher mobility leads to improved transistor performance and reduced thermal noise, resulting in a more robust memory cell.

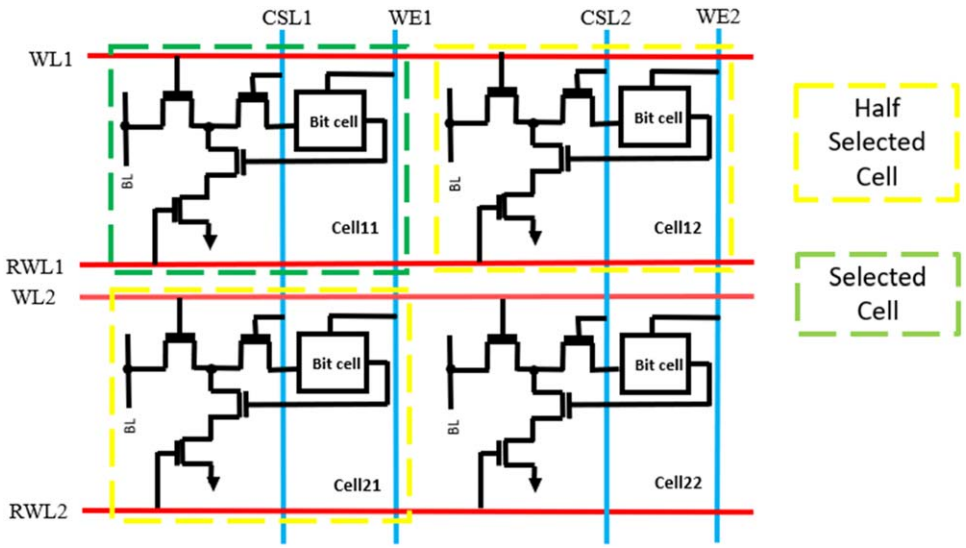


Figure 9. A 2 × 2 array of proposed 9 T cell to show half selected cell when a SRAM bitcell is selected.

7. Half select disturbance and soft error

When a bit cell in the array of SRAM cells is selected, other cells that are in the same row and same column will get half selected (HS). So, whenever a write/read operation is performed on selected cell, these half-selected cells may report an unintended write/read action. This unintentional read/write action in half selected cell is called half select disturbance [44]. Thus, stability of a SRAM cell must be ensured during HS condition as well. For a bit cell to be HS disturbance-free, it must incorporate exclusive row-based and column-based control signals to govern its write/read operation.

Of all the SRAM cells listed in section 2, only proposed 9T has reported HS disturbance free operation. 6T and ST 9T cell are susceptible to both read and write half select disturbance, whereas the 8T and TRD 9T cells only experience a write HS disturbance. Use of exclusive row-based and column-based signals enables the proposed 9T cell and SB 9T cell to register a HS disturbance free operation. Using RWL and CSL signals during read operation and WL and WE signals during write operation will make the 9T cell HS disturbance free while SB 9T uses VGND and WL signals during write operation while WWL and RE signals during read operation to ensure a HS free operation. Thus, validates that the proposed 9T cell is robust to HS disturbance.

Ideally, only selected cell should take part in operation, but because of common row based or column-based signals, they get partially selected [44]. The performance of the proposed 9 T cell is assessed by constructing a 2 × 2 array and simulating it against the HS condition. Figure 9 shows a depiction of the 2 × 2 array for the proposed 9T cell. At any given time, only one cell in the array is selected, and other cells in same row and column as the selected cell are half-selected; these cells are also indicated in figure 8. The cells in yellow box are half selected as they shared a common row based or column-based control signal but they do not take part in any operation. These half selected cells register a decline in its SNM value. The SNM curve for the half selected condition for the proposed cell is depicted in figure 10(a). To SNM for the cell reduces slightly and is recorded to be 235 mV.

To further evaluate the stability of the proposed cell under half select condition, SNM is evaluated using 5000 Monte Carlo simulation. The simulation results for Monte Carlo simulation of the proposed cell under half select condition is depicted in figure 10(b). The resilience of the proposed cell to HS condition is demonstrated with the help of the timing diagram. The timing diagram for the proposed cell for different conditions is depicted in figure 10(c). As can be observed that during the time frame between 3 μs and 6 μs, the WL signal is high multiple timed while the WE signal is also high, this is the HS condition for the proposed cell. But it can be observed that during this time frame no write operation is registered for the proposed 9 T cell.

Soft errors appear in a SRAM cell as a result of speeding cosmic neutron or alpha-particle radiation. The occurrence of these errors is most prominent in circuits with nanoscale technology node. The speeding cosmic neutron or alpha-particle generates an extra charge in the circuit because of direct or indirect ionization. This might result in a voltage change from one level to another. This unintentional flipping of data caused by the radiation of speeding cosmic neutron or alpha-particles is known as soft error. This results in multi bit error in a single word. These errors usually arise in bursts, affecting subsequent bit cells. As a result, different studies [45, 46] have identified bit interleaving (BI) as an efficient way for ensuring stability of a cell against soft error. In BI,

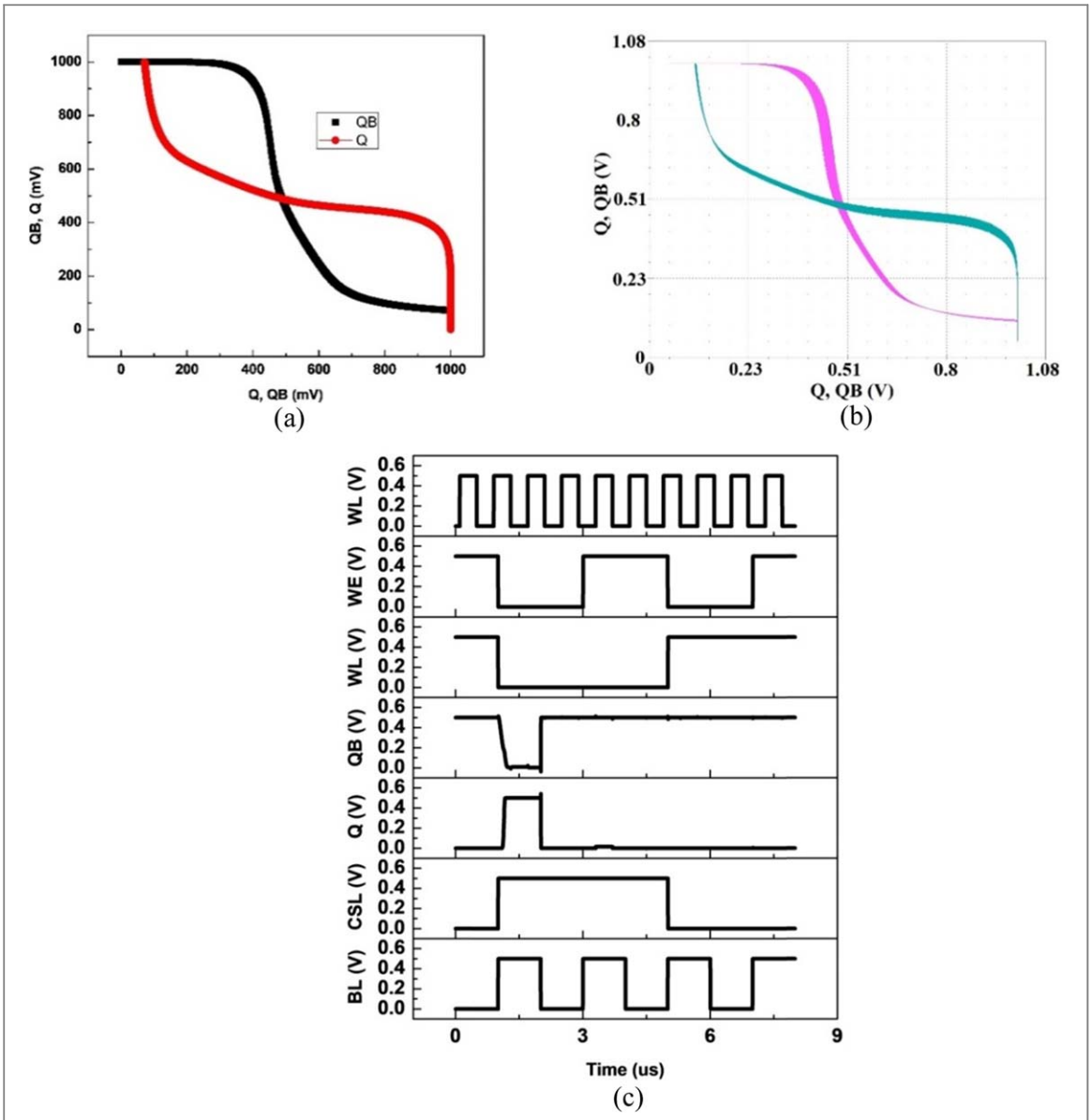


Figure 10. (a) SNM curve for a half selected cell, (b) SNM curve for 5000 Monte Carlo simulations, and (c) Timing curve to demonstrate HS resilience of the proposed bit cell.

each corresponding bit of every word is grouped and placed separately resulting in only single bit error in each word. Therefore, constructing array architecture for the proposed 9T cell, using the BI approach, will make the cell robust and ensure that cell is shielded against such errors. It is essential that a cell should possess HS disturbance free operation so as to design an architecture based on the BI approach, for the SRAM array. As the proposed 9T cell possess HS disturbance free operation, allowing implementation of BI architecture, which improves its immunity. Thus, the proposed cell may exhibit reliable performance against soft errors.

8. Local and global variation analysis

Technology node reduction has become necessary due to growing demand for high-density memories. However, it makes the cell more prone to local and global mismatch variations. Therefore, analysing the influence of variation is important, as scaling of technology node severely affects the circuit performance [47, 48]. It is imperative to ensure cell's performance against local variation, which is assessed using Monte Carlo simulations. These simulations are conducted via statistical techniques to determine the degree of variation in cell performance owing to local variations. V_{TH} is considered as a prominent parameter for this analysis, having an independent Gaussian distribution with a variation of 6σ around the mean V_{TH} value of the device. This methodology varies V_{TH} in the 6σ range around its mean value. The performance of 9T cell is analysed against process variation for 5000 data points. The simulation results (butterfly curves) for HSNM (RSNM) and WM

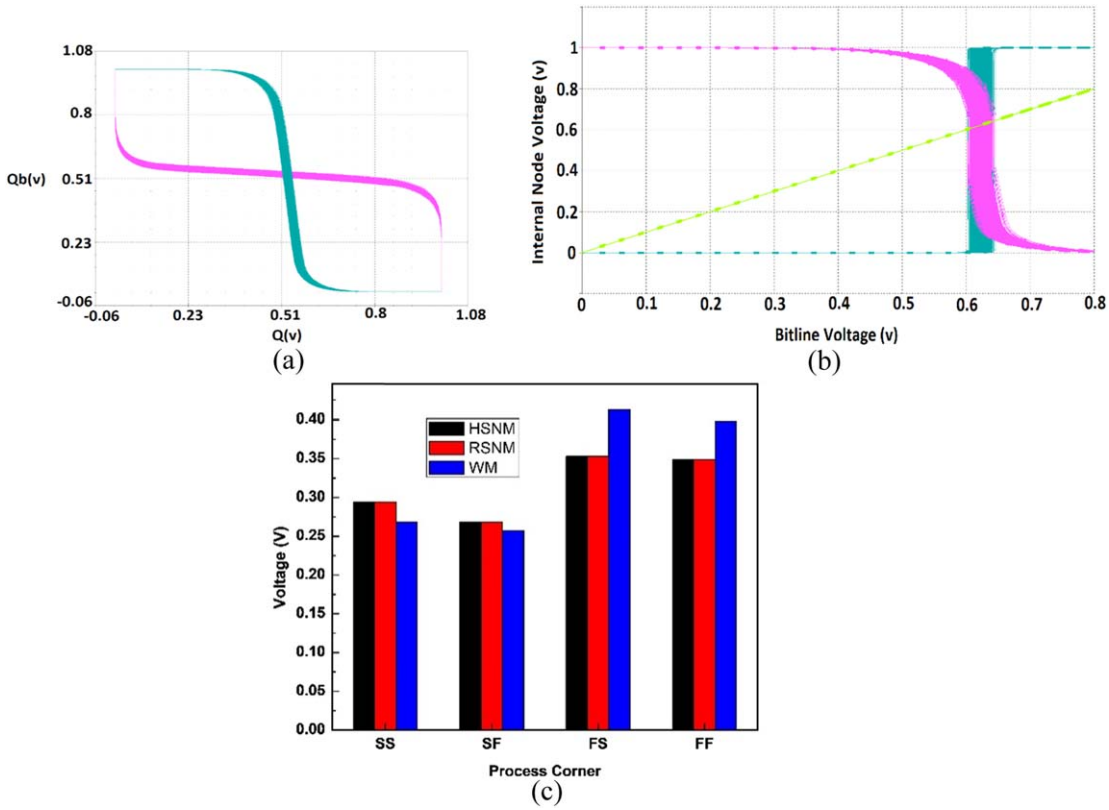


Figure 11. Butterfly curves for 5,000 MC simulations results during the (a) hold (read) operation (b) write operation for 9 T proposed cell, and (c) and (c) comparison for noise margin at different process corner for the proposed 9 T cell.

Table 5. Static margin values (HSNM, RSNM, WM) for Monte Carlo simulations for all cells.

Cells	HSNM (min)	HSNM (max)	RSNM (min)	RSNM (max)	WM (min)	WM (max)
6T	0.336	0.378	0.182	0.210	0.101	0.144
8T	0.336	0.378	0.336	0.378	0.251	0.286
9T ST	0.308	0.336	0.308	0.336	0.287	0.348-
TRD 9T	0.308	0.35	0.308	0.35	0.234	0.296
SB 9T	0.336	0.392	0.336	0.392	0.249	0.312
NTV 9T	0.319	0.375	0.319	0.375	0.485	0.523
9T Proposed	0.336	0.392	0.336	0.392	0.355	0.398

variation, obtained using Monte Carlo simulations, are shown in figures 11(a) and (b), respectively. The 9T ST cell registers least variation in static noise margin values for read and write stability due to high noise margin compared to other cells. 6T suffers from read and write upset due to low noise margin while 8T cell might report write disturbance. The minimum and maximum values for HSNM, RSNM, and WM obtained due to local variation are tabulated in table 5. Thus, it is reasonable to assert that performance of proposed 9T cell is most reliable against process variation.

The impact of global variation on the performance of a cell is analysed by evaluating the performance for a cell at the different process corners. The results for Monte Carlo simulations have helped identify the resilience of the proposed cell against local variation. Thus, to ensure that the cell is also resilient to global variations, its static margin analysis is performance for all the process corners. The values obtained at the different process corners are graphically compared and presented in figure 11(c). As can be observed for figure 11(c), the best results are obtained for FS corner, closely followed by the FF corner.

This is because the performance of the n-type and p-type transistors improves at the two corners whereas, for SS and SF corner the performance for the cell deteriorates. The highlight of the global variation analysis is that the maximum variation in performance due to global variation is only 60 mV for HSNM and RSNM. While for WM, the maximum variation is registered to be 120 mV. Thus, it may be inferred that though there is slight

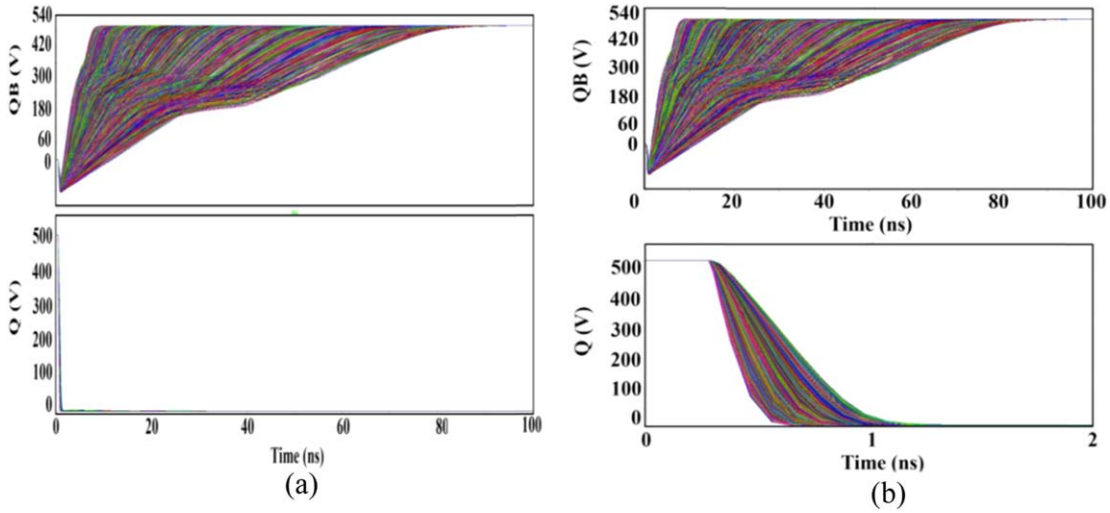


Figure 12. Write operation for the proposed 9 T cell for 5000 Monte Carlo simulation at the (a) SF, and (b) FS corner for −40 °C.

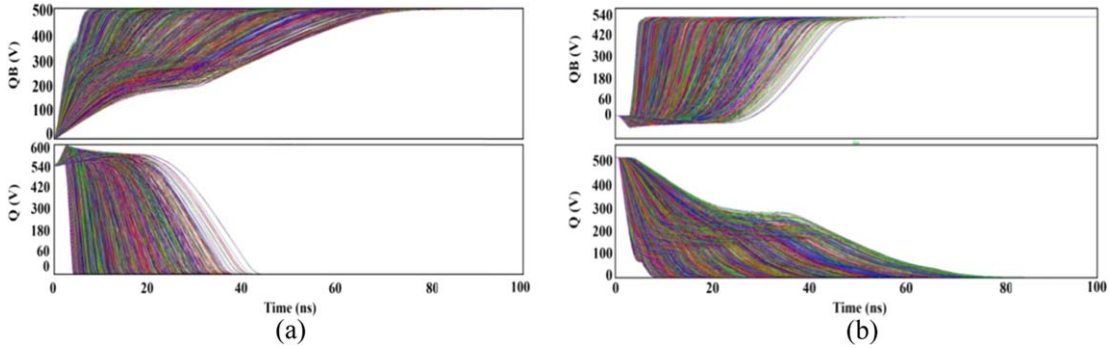


Figure 13. Write operation for the proposed 9 T cell for 5000 Monte Carlo simulation at the (a) SF, and (b) FS corner for 100 °C.

variation in performance due to process corners, the variation is within bounds and will not have a dramatic impact on the cell performance.

To demonstrate the resilience of the write operation for the proposed cell towards the global variations the write timing curve for the cell is simulated for both the SF and FS corner. Additionally, to check for performance at extreme temperature variations the same is simulated at −40 °C and 100 °C. The write operation curves for the SF and FS corner at −40 °C are depicted in figures 12(a) and (b) respectively. While the write operation curves for the SF and FS corner at 100 °C are depicted in figures 13(a) and (b) respectively.

9. Read and write time analysis

The analysis performed and explained thus far are all static in nature and do not account for the time required to performance the read and write operation for the bit cell. But real time operations are time bound, and the access operation for a bit cell (read and write) are key operations that determine the operational speed for the bit cell. Thus, it is essential to determine the width of the pulse that is required to be high to successfully perform the two operations. The read operation for a bit cell is measured as the time required to lower the BL value for the cell to switching voltage for an inverter [49]. Whereas, the write time is defined the time required to completely flip the data stored inside a cell [41].

The read and write time values obtained for the different cells are compared against the proposed 9T cell in figures 14(a), and (b) respectively. It is observed that the 9T ST cell and the SB 9T cell require high read time for V_{DD} lower than 0.7 V. Whereas, the other bit cells are able to perform their operation comparatively faster for all values of V_{DD}. In terms of write time the ST9T cell has the weakest performance as its write time requirement is

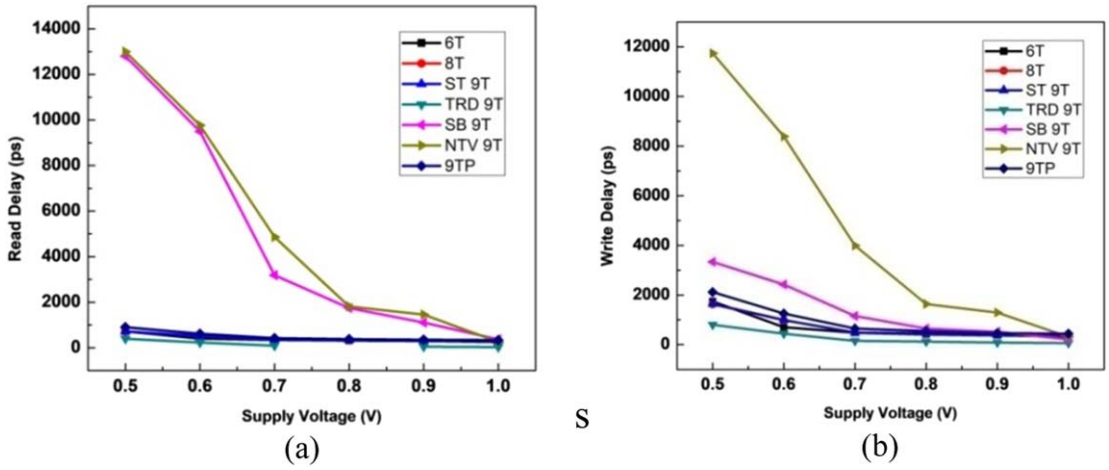


Figure 14. Comparison of (a) read, and (b) write time required for the different SRAM bit cells for V_{DD} ranging from 0.5 V to 1 V.

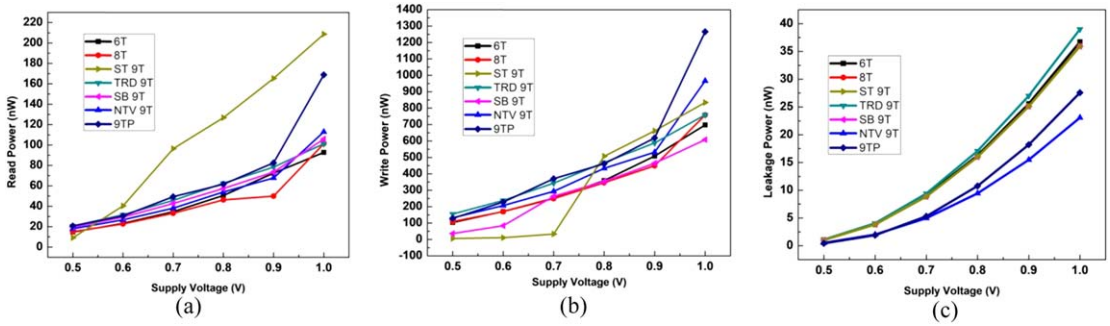


Figure 15. Comparison of (a) read, (b) write, and (c) leakage power consumption for the different SRAM bit cells at different V_{DD} values.

comparatively higher than the other bit cell. Whereas, the read and write time requirement for the proposed 9T cell are significantly lower, thereby ensuring its fast operation.

10. Power concumption analysis

Another highly essential parameter for a bit cell is its power consumption. In keeping with the three operations for a bit cell, its power consumption can also be divided into three categories—read power, write power and hold (leakage) power. With the growing demand for consumer devices with longer battery life, it is essential to have power consumption for the cell as low as possible. Thus the read, write and leakage power for all the bit cells are calculated for different V_{DD} values and compared in figures 15(a)–(c) respectively.

It can be inferred from the curves in figure 15(a), that read higher power consumption values are higher for the ST 9T cell for all values of V_{DD} above 0.6 V. Whereas, the least read power values are obtained for the 8T cell during the read operation. The power consumption values for the proposed 9T cell are comparable to the existing cell for all V_{DD} values except 1 V. In terms of write power consumption, all the cells seem to have comparable values except for 1 V, where the proposed 9T cell records the highest value. Whereas for leakage power consumption, the best performance is recorded for the NTV 9T cell; followed by the proposed 9T cell. Thus, based on the power consumption analysis it may be inferred that if low power applications are to be targeted the proposed 9 T cell should be operated at V_{DD} value lower than 1 V.

As the V_{DD} is lowered, the impact of leakage power on the performance of the bit cell starts to dominate. Therefore, leakage is an essential performance parameter. Additionally, at lower voltages if the circuit encounters temperature variations, it becomes critical to analyze the leakage power performance for variation in temperature. In keeping with the temperature variation analysis presented in section 5, the leakage power consumption values for the different cells at varied V_{DD} values are also calculated for −10 °C, and 100 °C. The

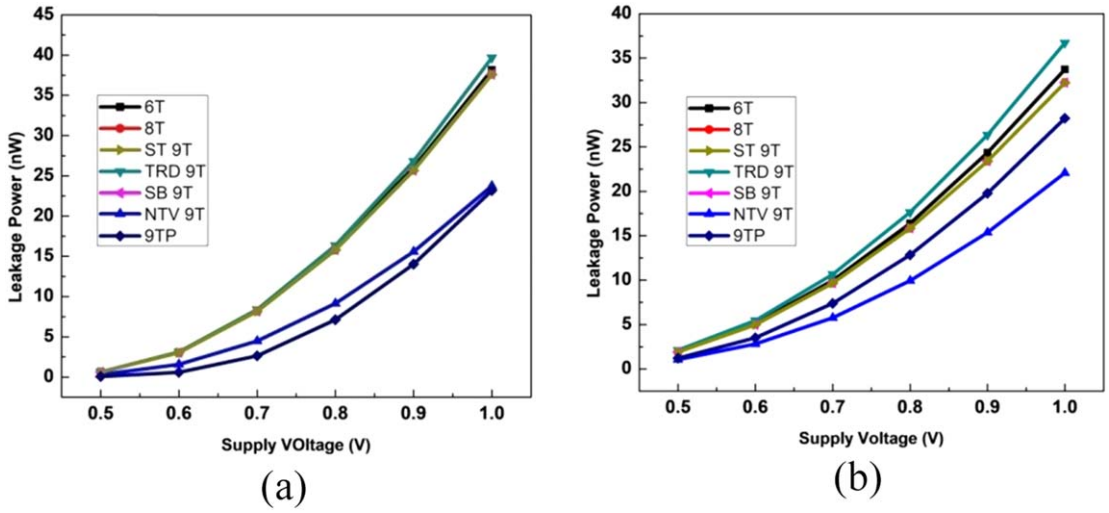


Figure 16. Comparison for leakage power values for the different cells at varied V_{DD} for (a) − 10 °C, and (b) 100 °C.

results obtained for the leakage current at −10 °C, and 100 °C for the different cells are graphically compared in figures 16(a), and (b) respectively.

11. Layout

The increasing processing load is increasing the demand for high density memories. Bit cells are arranged in an array configuration to form the complete SRAM integrated into a microprocessor. SRAM tend to occupy a significant portion of the total area on chip. Therefore, it is essential to determine the layouts area for a bit cell. Lower the area for a bit cell, higher will be its integration density. Thus, layouts for all the bit cells are designed and presented in figure 17. For ease of comparison, the length, width, and area for the different bit cells are tabulated in table 6.

The 6T bit cell has the least area occupancy with area of 14.56 μm^2. While, the area for the proposed 9T cell is 24.91 μm^2, the layout for the cell is depicted in figure 11(g). The layout area for the 8T, ST 9T, TRD 9T, SB 9T, and NTV 9T cell is 23.98, 30.84, 24.03, 31.71, and 26.07 μm^2, respectively. The area values corresponding to 6T, 8T, and TRD 9T are lower than the proposed cell with area of 14.56, 23.98, and 24.03 μm^2 respectively. But the area for 9T ST, SB 9T, and NTV 9T cells is greater than the proposed 9T cell by 23.8, 27.29, and 4.65%, respectively. Thus, it may be inferred that the area footprint for the proposed cell is optimal in comparison to most 9T bit cells.

12. Electrical quality metric

The different analysis' explained previously are isolated from one to another. These analysis' also reflect the different performance trade-offs between the performance parameters for a cell. Thus, for a comprehensive analysis for a cell, the parameter in use is electrical quality metric (EQM). The EQM for a bit cell is a representative of its overall performance and is mathematically defined as follow –

$$EQM = \frac{RSNM * HSNM * WM}{Read \quad Delay * Area * P_{Hold} * P_{Read} * P_{Write}}$$

Here the RSNM, HSNM, and WM are the static margin values for the cell obtained during the static analysis, P_{Hold}, P_{Read} and P_{Write} are the power consumption values obtained for the cell during the hold, read, and write operation, respectively. The EQM values obtained for the different bit cells are compared in figure 18.

The EQM is a wholesome performance parameter for a bit cell. It accounts for all the major performance parameters for a cell. Thus, it is considered a reliable parameter to predict the quality of a cell. The higher the EQM well, the better the performance of the bit cell. The EQM values for the different bit cells are compared with the proposed 9T cell in figure 18 for varied values of V_{DD}. It can be inferred from figure 15 that EQM values for the 9TP cell is high for values 0.5 and 0.6 V, thereby validating that its performance is better suited for low voltage applications. Also, it is closely followed by the NTV 9T cell. The cell with the least values for EQM is the SB 9T cell.

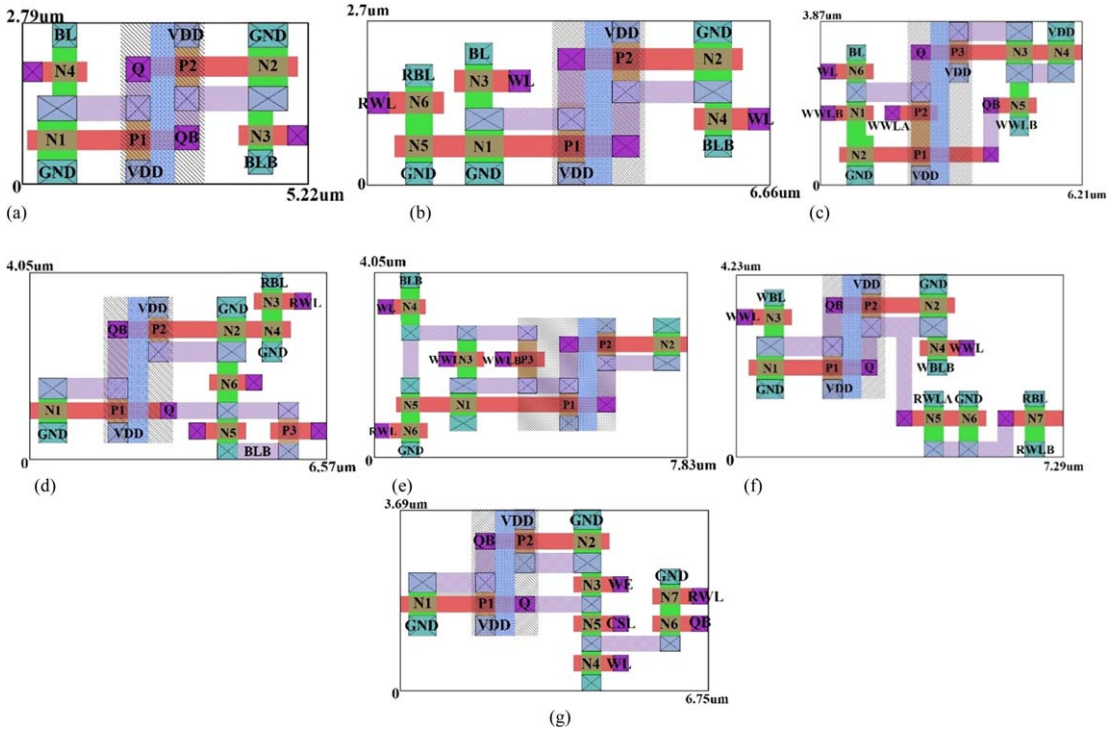


Figure 17. Layout footprint for (a) 6 T, (b) 8 T, (c) 9 T ST, (d) TRD 9 T, (e) SB 9 T, (f) NTV 9 T, and (g) proposed 9 T SRAM bit cells.

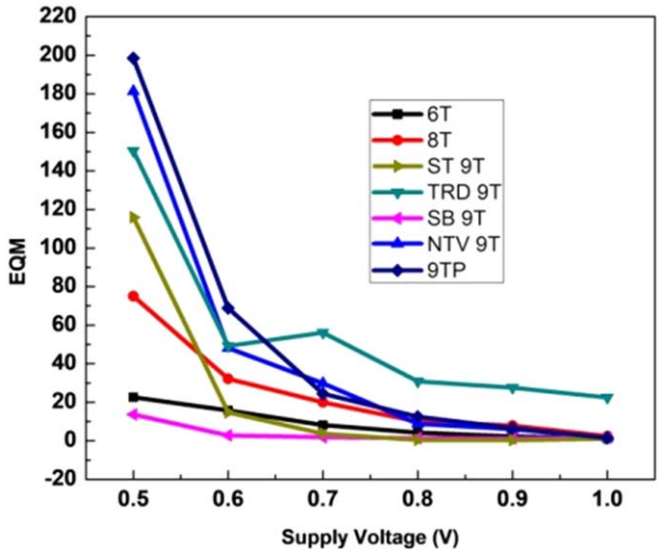


Figure 18. Comparison of Electrical Quality Metric ($x10^{28}$) for the different SRAM bit cells for varied V_{DD} values.

Table 6. Length, width, and area measurement for the different SRAM bit cells.

Cells	Length (μm)	Width (μm)	Area (μm^2)
6 T	5.22	2.79	14.56
8 T	6.66	2.7	23.98
9 T ST	6.57	4.05	30.84
TRD 9 T	6.21	3.87	24.03
SB 9 T	7.83	4.05	31.71
NTV 9 T	7.29	4.23	26.07
9 T Proposed	6.75	3.69	24.91

13. Conclusion

In this research work, a single-ended 9T (9 Transistor) SRAM cell is proposed. All the cells discussed throughout this study are designed and simulated at 180 nm technology node with a voltage supply of 1V. The proposed cell performs all of its operations (read or write) via a single bitline. The adoption of single ended bitline in the architecture of the proposed design effectively mitigates the activity factor thus minimizing the dynamic power consumption. The performance of 9T cell is analysed and equated against the existing 6T, 8T, 9TST, SB 9T, TRD 9T, and NTV 9T bit cells to showcase the improvement in its overall performance. The cell exhibits a comprehensive performance as its static performance is best. The analysis divulges that the HSNM, RSNM and WM values for the proposed cell are 364, 364, 378 mV respectively. The cell is analysed over the temperature ranging from −10 °C to 100 °C and has a reliable performance against temperature variation as its impact on static performance is minimal. The analysis divulges that the static performance for 9 T cell has an improvement of 9.09%, 42% in HSNM (RSNM), WM respectively at 100 °C and 44.7% in WM at −10 °C as compared to 8 T cell, which outperforms all other reviewed cells. Moreover, the proposed cell is designed to be resilient to half select disturbance, making it possible to implement its BI architecture, which furthermore shield the cell against soft. The area footprint for the cell is 24.91 μm^2 that is 23.8, 27.29, and 4.65% lower than the area obtained for its 9 T counter parts − 9T ST, SB 9T, and NTV 9T.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Statements and declarations

Funding

Not Applicable

Conflicts of interest/competing interests

Not Applicable

Availability of data and material

All data generated or analyzed during this study are included in this published article.

Code availability

Not Applicable

ORCID iDs

Poornima Mittal https://orcid.org/0000-0002-9479-8628

References

[1] Kim Y, Li S, Yadav N and Choi K K 2021 A novel ultra-low power 8T SRAM-based compute-in-memory design for binary neural networks *Elecronics* **10** 2181–94
[2] Sachdeva A and Tomar V 2020 Design of multi-cell upset immune single-end SRAM for low power applications *AEU-International Journal of Electronics and Communications* **128** 1434–8411
[3] Tomar V K and Sachdeva A 2022 Design of a soft error hardened SRAM cell with improved access time for embedded systems *Microprocess. Microsyst.* **90** 104445
[4] Ruhil S, Khanna V, Dutta U and Shukla N K 2021 A study of emerging semi-conductor devices for memory applications *International Journal on Nano Dimensions* **12** 186–202
[5] Ayush, Mittal P and Rohilla R 2023 Comparative analysis of current sense amplifier architectures for SRAM at 45 nm technology node *Advances in Data-Driven Computing and Intelligent Systems. ADCIS 2022. Lecture Notes in Networks and Systems* 698, pp 633–40

[6] Kursun V, Tawfik S A and Liu Z 2007 *Leakage-Aware Design of Nanometer SoC Proc. of the IEEE Int. Symp. on Circuits and Systems* 3231–4
[7] Pourbakhsh S A, Chen X, Chen D, Wang X, Gong N and Wang J 2016 Sizing-priority based low-power embedded memory for mobile video applications *17th Int. Symp. on Quality Electronic Design (ISQED)*
[8] Ahmad S, Gupta M K, Alam N and Hasan M 2016 Single-ended schmitt-trigger-based robust low-power SRAM cell *IEEE Transactions on Very Large Integration (VLSI) Systems* **24** 2634–42
[9] Gupta S, Member S, Gupta K and Pandey N 2018 Pentavariate vmin analysis of a subthreshold 10T SRAM Bit cell with variation tolerant *IEEE Trans. Circuits Syst. I* **6** 1–12
[10] Sachdeva A and Tomar V K 2021 Stability and dynamic power analysis of novel 9t sram cell for iot applications *Communication and Intelligent Systems. Lecture Notes in Networks and Systems* 204, pp 945–53
[11] Ahmad S, Iqbal B, Alam N and Hasan M 2018 Low leakage fully half-select-free robust SRAM cells with BTI reliability analysis *IEEE Transactions on Device and Material Reliability* **18** 337–49
[12] Toh S O, Guo Z, Liu T J K and Nikolic B 2011 Characterization of dynamic SRAM stability in 45 nm CMOS *IEEE Journal on Solid-State Circuits* **46** 2702–12
[13] Chang L *et al* 2005 Stable SRAM cell design for the 32 nm node and beyond *Digest of Technical Papers. 2005 Symposium on VLSI Technology (Kyoto, Japan)* pp 128–9
[14] Jiang J, Xu Y, Zhu W, Xiao J and Zou S 2019 Quadruple cross-coupled latch-based 10T and 12T SRAM Bit-cell designs for highly reliable terrestrial applications *IEEE Trans. Circuits Syst. I* **66** 967–77
[15] Kumar M and Ubhi J S 2019 Design and analysis of CNTFET based 10T SRAM for high performance at nanoscale *International Journal on Circuit Theory and Applications* **47** 1775–85
[16] Maroof N and Kong B 2017 10T SRAM using half- VDD precharge and row-wise dynamically powered read port for low switching power and ultralow RBL leakage *IEEE Trans. Very Large Scale Integr. VLSI Syst.* **25** 1193–203
[17] Oh T W, Jeong H, Kang K, Park J, Yang Y and Jung S O 2017 Power-gated 9T SRAM cell for low-energy operation *IEEE Transactions on Very Large Scale Integration Systems* **25** 1183–7
[18] Divya and Mittal P 2022 A low-power high-performance voltage sense amplifier for static RAM and comparison with existing current/voltage sense amplifiers *Int. J. Inf. Technol.* **14** 1711–8
[19] Mukhopadhyay S, Mahmoodi H and Roy K 2005 Modeling of failure probability and statistical design of SRAM array for yield enhancement in nanoscaled CMOS *IEEE Transactions on Computer Aided Design (CAD) Integrated Circuit Systems* **24** 1859–80
[20] Ruhil S, Dutta U, Khanna V and Shukla N K 2022 Design of a 30 Nm Novel 3D quad gate stacked nano-sheets FinFET *Silicon* **14** 11859–568
[21] Cho K, Park J, Oh T W and Jung S O 2020 One-sided schmitt trigger based 9T cell for near threshold operation *IEEE Trans. Circuits Syst. I* **67** 1551–61
[22] Rawat B and Mittal P 2021 A 32 nm single ended single port 7T SRAM for low power utilization *Semicond. Sci. Technol.* **36** 095006–22
[23] Pal S and Islam A 2016 9-T SRAM cell for reliable ultralow-power applications and solving multibit soft-error issue *IEEE Transactions on Device and Material Reliability* **16** 172–82
[24] Aly R E and Bayoumi M A 2007 Low-power cache design using 7T SRAM cell *IEEE Transactions on Circuits Systems II: Express Briefs* **54** 318–22
[25] Pal S, Gupta V, Ki W H and Islam A 2019 Transmission gate-based 9T SRAM cell for variation resilient low power and reliable internet of things applications *IET Circuits Devices Syst.* **13** 584–95
[26] Pal S, Bose S, Ki W H and Islam A 2019 Characterization of half-select free write assist 9T SRAM cell *IEEE Trans. Electron Devices* **66** 4745–52
[27] Meshram N, Prasad G, Sharma D and Mandi B C 2022 *Low Power and High Speed SRAM for Ultra Low Power Applications IEEE Int. IOT, Electronics and Mechatronics Conf. (IEMTRONICS)*
[28] Tu M H, Lin J Y, Tsai M C, Jou S J and Chuang C T 2010 Single-ended subthreshold SRAM with asymmetrical write/read-assist *IEEE Trans. Circuits Syst. I* **57** 3039–47
[29] Farkhani H, Peiravi A and Moradi F 2014 A new asymmetric 6T SRAM cell with a write assist technique in 65 nm CMOS technology *Microelectron. J.* **45** 1556–65
[30] Takeda K, Hagihara Y, Aimoto Y, Nomura M, Nakawaza Y, Ishii T and Kobatake H 2006 A read-static-noise-margin-free SRAM cell for low-VDD and high-speed applications *IEEE J. Solid-State Circuits* **41** 113–21
[31] Abbasi A, Setoudeh F, Tavakoli M G and Horri A 2021 Design and simulation of high speed and low power memcapacitor based non-volatile static cells using FinFET transistors *Semicond. Sci. Technol.* **37** 075027
[32] Abbasi A, Setoudeh F, Tavakoli M G and Horri A 2022 A novel design of high performance and robust ultra low power SRAM cell based on memcapacitor *Nanotechnology* **33** 165202
[33] Singh J, Mohanty S P and Pradhan D K 2013 *Robust SRAM Designs and Analysis.* (Springer)
[34] Islam A and Hasan M 2012 Leakage characterization of 10T SRAM cell *IEEE Trans. Electron Devices* **59** 631–8
[35] Chaturvedi M, Garg M, Rawat B and Mittal P 2021 8T SRAM bit cell with improved read stability for 90 nm technology node *Int. Conf. on Simulation, Automation & Smart Manufacturing (SASM 2021)* 20–1
[36] Wen L, Li Z and Li X 2013 Single-ended robust 8T SRAM cell for low voltage operations *Microelectron. J.* **44** 718–28
[37] Lin S, Kim Y B and Lombardi V 2008 A highly-stable nanometer memory for low power design *Proc. of IEEE Int. Workshop Design Test of Nano Devices, Circuits Systems 2008* 17–20
[38] Mansore S R, Gamad R S and Mishra D K 2020 A 32 nm read disturb-free 11T SRAM cell with improved write ability *Journal of Circuits Systems and Computing* **29** 2050067
[39] Rawat B and Mittal P 2022 A reliable and temperature variation tolerant 7T SRAM cell with single bitline configuration for low voltage application *Circuits Syst. Signal Process.* **41** 2779–801
[40] Ahlawat S, Siddharth, Rawat B and Mittal P 2021 *A Comparative Performance Analysis of Varied 10T SRAM Cell Topologies at 32 nm Technology Node Int. Conf. on Modeling, Simulation and Optimization 2021*
[41] Ahmad S, Gupta M K, Alam N and Hasan M 2016 Single-ended schmitt trigger based robust low power SRAM cell *IEEE Trans. Very Large Scale Integr. VLSI Syst.* **20** 2634–42
[42] Nelms R M and Johnson R W 1997 200 °C Operation of a 500-WDC-DC Converter Utilizing Power MOSFETs *IEEE Trans. Ind. Appl.* **33** 1267–72
[43] Uy K J, Abu P A and Chung W Y A 2009 High precision temperature insensitive current and voltage reference generator *World Academy of Science, Engineering, and Technology* **3** 72–732

[44] Chang I J, Kim J J, Park S and Roy K 2009 A 32 kb 10T sub-threshold SRAM array with bit-interleaving and differential read scheme in 90 nm CMOS *IEEE Journal of Solid State Circuits* **44** 650
[45] Ensan S S, Moaiyeri M H, Moghaddam M and Hessabi S 2019 A low-power single-ended SRAM in FinFET technology *AEU—International Journal of Electronics and Communication* **99** 361–8
[46] Rawat B and Mittal P 2022 *A Comprehensive Analysis of Different 7T SRAM Topologies to Design a 1R1W Bit Interleaving Enabled and Half Select Free Cell for 32 nm Technology Node Proc. of Royal Society A - Mathematical, Physical and Engineering Sciences* 478(2259) 20210745
[47] Wen L, Cheng X, Zhou K T and Zeng X 2016 Bit-interleaving-enabled 8T SRAM with shared data aware-write and reference-based sense amplifier *IEEE Transactions on Circuits Systems: Express Briefs* **63** 643–7
[48] Doorn T, Maten E T, Croon J, Bucchianico A D and Wittich O 2008 Importance sampling Monte Carlo simulations for accurate estimation of SRAM yield *ESSCIRC—34th European solid-State Circuits Conf.*
[49] Lin S, Kim Y B and Lombardi F 2010 Design and analysis of a 32 nm PVT tolerant CMOS SRAM cell for low leakage and high stability *Integration* **43** 176–87

A Stable Deep Learning Model for Recognizing Human Actions in Noisy Videos

Ishwar Chand, Sanjay Kumar, Jatin Popli*, Jitesh Kumar
Computer Engineering
Delhi Technological University
New Delhi, India
ishwarchand_2k19co171@dtu.ac.in, sanjay.kumar@dtu.ac.in, jatinpopli_2k19co176@dtu.ac.in,
jiteshkumar_2k19co179@dtu.ac.in

2023 IEEE World Conference on Applied Intelligence and Computing (AIC) | 979-8-3503-1006-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/AIC57670.2023.10263859

Abstract – The widespread application and development of technology have significantly improved people's lives. In recent years, deep learning has emerged as a powerful branch of artificial intelligence, enabling the creation of intelligent software and robots that assist with various daily tasks. Deep Learning, a subtype of Complex Artificial Neural Networks (ANNs) based on machine learning, is commonly used for accurate human behavior detection from smartphone data. In this paper, we propose a model to identify human activity using ConvLSTM, DenseNet, and LRCN. The statistical model is built on a substantial dataset of videos sourced from publicly accessible datasets UCF50 and HMDB51, which cover a wide range of activity variables. Our results demonstrate that the DenseNet-based model outperforms other models in terms of accuracy.

Keywords - ConvLSTM; LRCN; DenseNet; LSTM; Deep Learning;

I. INTRODUCTION

The evaluation and analysis of human behavior have become a popular research field, with deep learning emerging as a promising method to address this challenge. Researchers have utilized various devices, such as CCTV cameras, computers, and mobile applications, to collect surveillance data on people's activities [1].

Machine learning and deep learning models, including DenseNet, CNNs, LSTMs, ConvLSTMs, and LRCNs, are frequently employed in studying human behavior and have shown significant promise in identifying different tasks from videos. These models are trained using surveillance information gathered from various sources, such as CCTV cameras and mobile applications. Deep learning-based techniques have proven to be effective in solving real-life problems, including various computer vision tasks [2-8].

Our contributions are summarized as follows:

1. Development of a Human Activity Recognition Model Using CNN and LSTM, specifically ConvLSTM, LRCN, and DenseNet architectures.
2. Conducting several experiments to compare the model's performance using publicly accessible datasets (UCF50 and HMDB51) of videos. These datasets contain thousands of videos and images, illustrating fifty distinct human motions, including possible interactions.
3. Using accuracy as an evaluation metric to demonstrate that the DenseNet model outperforms other deep learning techniques.

The rest of the paper outlines our proposed study, including a literature review, methodology and dataset description, experiment setup and evaluation, and finally, a conclusion.

II. LITERATURE REVIEW

The detection of human actions using computer vision and pattern recognition has received considerable attention from academics due to its diverse range of potential applications.

A. Machine Learning for Recognizing Human Activities

In their work, Sowmya et al. [9] developed a machine learning system for detecting and classifying human activities using CNN along with spatio-temporal 3D kernels. The CNN model was trained on a dataset comprising three classes of human activities in daily life and work, represented by images. The model achieved high accuracy in identifying nearly three human activities.

Sandeep [10] proposed a machine learning-based approach for recognizing human movements using sensors such as gyroscopes, accelerometers, GPS sensors, and compass sensors in smartphones. The author developed an Activity Recognition (AR) framework to monitor the user's health by examining unprocessed sensor information from tiny sensors. The model demonstrated an accuracy of 93.75% with a 6.25% error rate and a preparation time of 10.56 seconds.

In a different investigation, Subasi et al. [11] created a system for identifying human activity in smart healthcare settings. The authors demonstrated how these techniques can achieve robust and accurate Human Activity Recognition (HAR) on two distinct datasets.

979-8-3503-1006-1/23/$31.00 ©2023 IEEE
DOI: 10.1109/AIC.2023.96

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 06:09:57 UTC from IEEE Xplore. Restrictions apply.

B. *Deep Learning for Human Activity Recognition*

Scholars have explored deep learning (DL) and introduced numerous DL-based algorithms for various applications, including computer vision, facial recognition, and speech recognition.

Wan et al.[12] conducted research on creating deep learning models for real-time smartphone activity recognition of human activity. For the UCI and Pamap2 datasets, they used mobile edge computing (MEC) to assess how well the CNN, LSTM, BLSTM, MLP, and SVM models performed. According to the authors' tests, CNN, as a traditional neural network technique, remains highly relevant in recognizing human behavior and works as a powerful classification and recognition system.

To recognize human action in movies, Orozco et al. [13] suggested a reliable method based on LSTM and CNN. Using the KTH, UCF-11, and HMDB-51 datasets, the authors evaluated the CNN-LSTM architectures being proposed. This model uses pre-trained VGG16 to extract features with a size of 40×250 from video data that had been standardized to 40 frames. The relevant LSTM units were used to process each vector after that, and the output stage comprised a dense layer with nodes that represented each. According to the study, the accuracy rates for the KTH dataset were 93.86%, the UCF-11 dataset 91.94%, and the HMDB-51 dataset was 47.36%.

Alexandros et al. [14] emphasize the importance of video input processing for recognizing temporally varying events, such as human actions. The Fractioned Adjacent Spatial and Temporal (FAST) 3D convolutions are a newly introduced convolution block for CNN architectures specifically designed to handle video input, as stated by the authors. In each FAST 3D convolution block, there are three consecutive convolution operations implemented to specifically capture the unique temporal features of actions, either in the horizontal or vertical direction. Experimental results on benchmark action recognition datasets indicate that FAST 3D convolution blocks outperform traditional 3D convolutions, especially for deeper networks. Notably, DenseNet-121 with FAST 3D convolutions achieves the best performance. The authors suggest that the decoupled spatio-temporal convolutions have merit for video processing tasks.

Xiaofang and colleagues' [15] study focuses on the challenge of efficiently extracting spatio-temporal features from videos to enhance the accuracy of action recognition. The authors propose a dense 3D convolutional block as the fundamental unit for building a dense convolutional 3D network that simultaneously extracts spatio-temporal features and reinforces the transmission and reuse of features within the network. To improve the discriminative power of the network, the authors also suggest a joint loss function that includes Fisher discriminant regularization. According to experimental results using the UCF-101 dataset, the proposed network attains a recognition accuracy rate of 92.4%, surpassing the C3D network's accuracy rate of 85.2%. These findings demonstrate the efficacy of the proposed method

III. METHODOLOGY

Figure 1 depicts the workflow of our activity recognition model. Firstly, video data is collected from a publicly available source. We then apply preprocessing techniques like filtering, shrinking frames, and normalizing data to the dataset before partitioning it into training and testing data. Several machine learning models, such as ConvLSTM, LRCN, and DenseNet121+LSTM are trained on the processed data. In our model, DenseNet121+LSTM is used. Lastly, performance analysis is conducted. For activity recognition, the model can analyze testing video data from various sources, including YouTube. Our experiments demonstrate that DenseNet121+LSTM performs better than all other models. The subsequent sections of the paper detail our proposed study, including a literature review, methodology, description of datasets, experiment setup and evaluation, and conclusion.

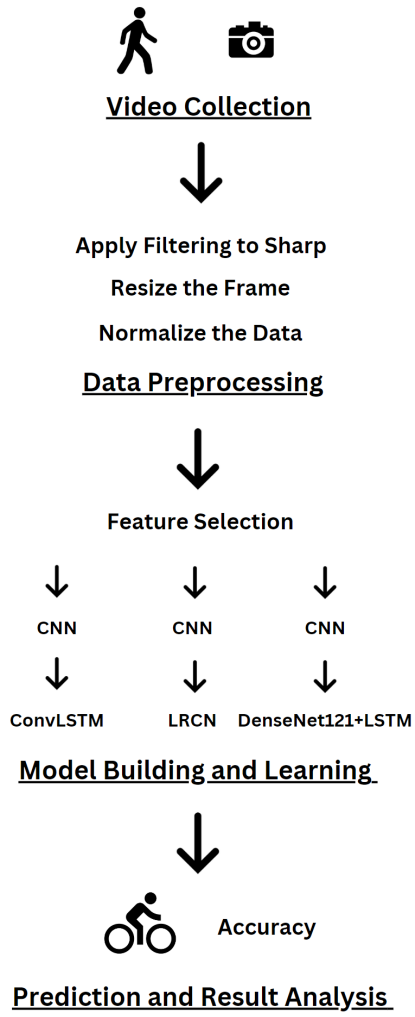


Fig. 1. The overall architecture of the proposed human activity recognition models.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 06:09:57 UTC from IEEE Xplore. Restrictions apply.

1. *Input Dataset:* Our model is trained on two publicly available datasets, UCF-50 and HMDB-51, which contains images and videos of many human activities, including running and walking, etc. Using machine learning or deep learning models to find these activities, image or video data, which can be captured using on-board computer systems or external cameras.

2. *Pre-processing the Dataset:* Preprocessing is a crucial step before feeding the datasets into the learning module with the intention of enhancing the data quality and getting better results. To hasten convergence during network training, we first read the video streams from the dataset before resizing the video frames to a particular length and width.

3. *Human Action Classification:* The next steps are taken to build the human activity recognition model:
    1. Collect video clip datasets for training the model on human activity.
    2. Apply processing on videos, such as filters to sharpen and resize frames to a given width and height, in order to normalize the data.
3. To facilitate training and testing, divide the normalized data into 75–25 groups, with 25% of the videos in the dataset used for validation.
4. Learn human activities by teaching Deep Learning models like LRCN, DenseNet121, and ConvLSTM.

4. Models for Deep Learning

A. *Approach for Long-Term Recurrent Convolutional Network (LRCN):* We combine convolutional and LSTM layers into a unified framework. Another approach is to train a CNN model and an LSTM model separately, with the pre-trained CNN model extracting spatial data from video frames and the LSTM model refining it for the application of activity recognition. In contrast, the LRCN model combines CNN and LSTM layers and clubs them into a single model. The CNN layers' collected spatial features are fed into an LSTM layer for temporal sequence modeling, yielding a robust model that learns spatio-temporal properties during end-to-end training.

B. *Approach for Convolutional Long Short-Term Memory (ConvLSTM) Approach:* This paper also uses ConvLSTM cells in the human activity recognition model. These cells incorporate convolution operations within the LSTM network, enabling the detection of spatial aspects in the data while also considering the temporal relationships between frames. Keras ConvLSTM2D recurrent layers are used to construct the model, which considers the size of the kernel necessary for convolutional operations and the number of filters.

C. *DenseNet121 Approach*: We first load the pre-trained DenseNet-121 model on ImageNet Dataset and freeze its layers so that they don't get updated during training. Next, We attach the model which is pre-trained as a TimeDistributed layer in the Keras Sequential model. This lets us put in the same DenseNet model to process every frame within the video sequence. After the TimeDistributed layer, we make use of the Flatten layer to convert the output from the pre-trained model to a 2D tensor that can be passed to the LSTM layer. In the end, we include an LSTM layer that has 64 units and a dropout rate of 0.5. This is then followed by a Dense layer which makes use of a softmax activation function to produce the predicted probabilities for each class. Figure 2 depicts the approach.

In this diagram, the input shape is (30, 64, 64, 3), where 30 is the number of frames in a video sequence, 64 is the height of each video frame, 64 is the width of each video frame, and 3 is the number of color channels (RGB).
The DenseNet121 model is used as a feature extractor, which is wrapped inside a TimeDistributed layer to process each frame of the video sequence. The flattened output is then fed to an LSTM layer with 64 units and a dropout rate of 0.5 to capture the temporal dependencies between frames. Finally, a Dense layer with 50 units (one for each class in the UCF50 dataset) and to generate the output probabilities for each class, a softmax activation function is employed.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 06:09:57 UTC from IEEE Xplore. Restrictions apply.

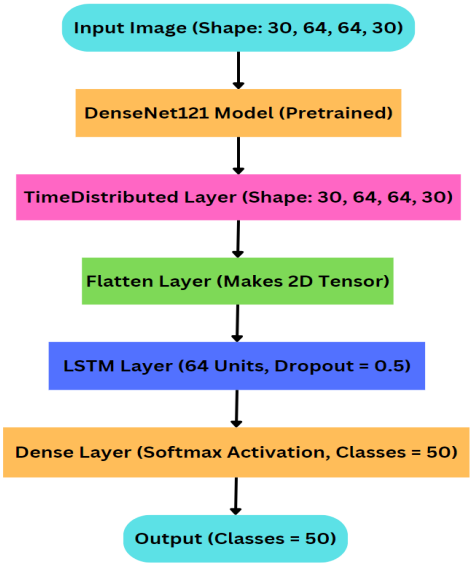


Fig. 2. DenseNet121+LSTM Approach.

IV. EXPERIMENTATION SETUP AND EVALUATIONS

In this section, we will go over the dataset that we used to train various deep learning models in our experiments. Accuracy and Training Loss metrics will be used to evaluate the performance of these models. We have performed comprehensive tests by using different quantities of videos to evaluate how well these models perform.

1. Dataset

To train a human activity recognition model, a large collection of videos of people performing various activities is required. The UCF50 action recognition dataset, which includes 50 different types of actions, was used in this study. The purpose of collecting this data is to provide a realistic action recognition dataset derived from YouTube videos to the computer vision community. Due to the various factors such as camera movements, item appearance, and posture, cluttered background, scale, illumination conditions, viewpoint, etc., the dataset is very difficult and opportunistic to work with. Some action categories in the dataset include shooting baskets in basketball, bench pressing, baseball pitch, diving, golf swing, fencing, and other activities. Another dataset used in this paper is HMDB51, which contains a diverse range of sources, including films and online videos, contributing to a sizable library of true-to-life videos. It contains at least 101 videos for each of the 51 categories of actions, such as laugh, jump etc. for a total of 6,849 video clips. Figure 3 shows some examples of clips from different categories.

Fig. 3. Examples Of Dataset Clips.

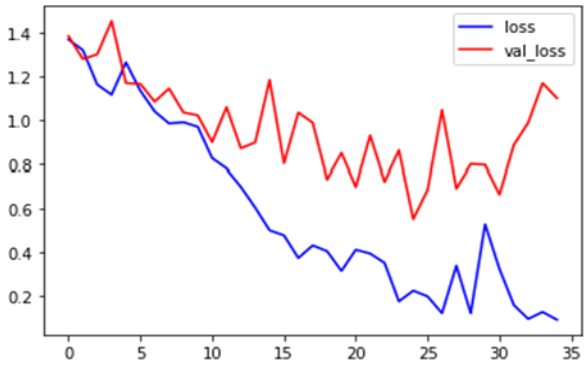
To ensure the fairness of our experiments, we split the dataset into two categories: testing and training. We divided the videos into 75 percent for training and 25 percent for testing. Furthermore, before splitting the dataset, we shuffled it to reduce bias and ensure that the splits accurately represent the overall distribution of the data.

2. Experiments

We ran experiments on two video datasets, converting them into individual frames for training and evaluation. The datasets were randomly divided into 25% for validation or testing and 75 percent for training purposes. We compared the results of three models namely, DenseNet, ConvLSTM, and LRCN.

A. Result of ConvLSTM Approach

In this case, we use more than 200 videos from the dataset. Figure 4 depicts the Convolutional LSTM's Training and Validation Accuracy and Loss. Table 1 examines the models' average accuracy.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 06:09:57 UTC from IEEE Xplore. Restrictions apply.

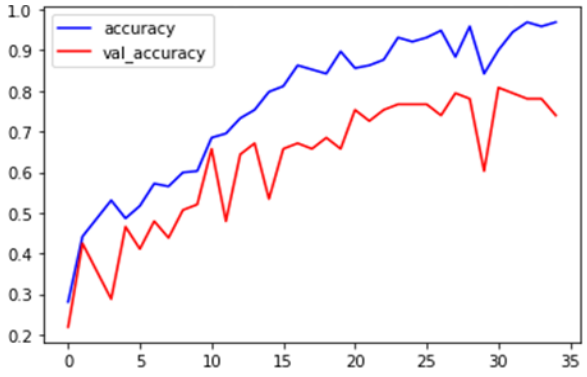


Fig. 4. ConvLstm Model Loss vs. Epoch

B. Result for LRCN Approach

As previously mentioned, we divided several videos into photos to assess the proposal's performance. In Figure 5, the Training and Validation Accuracy, along with the model loss, are illustrated for the LRCN approach. Our findings revealed that with an increase in the number of epochs, there is a gradual improvement in accuracy, but the model loss stabilizes after a few epochs, and it's shown in Fig 6.

The training and validation accuracy, and the training and validation loss for DenseNet121+LSTM, are shown in Figure 6. The accuracy of our model for the UCF50 dataset was 98.8% and for the HMDB51 dataset was 90.8%.

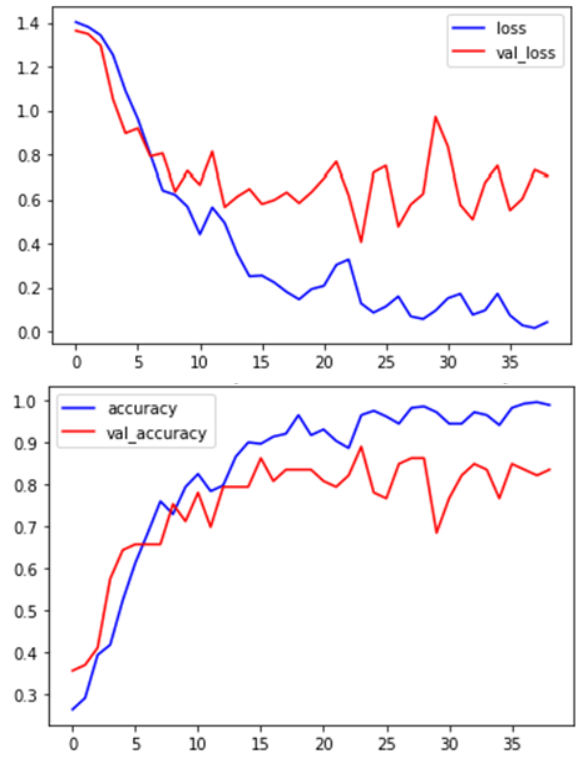


Fig. 5. LRCN Model and Loss vs. Epoch.

C. Result for DenseNet121+LSTM Approach

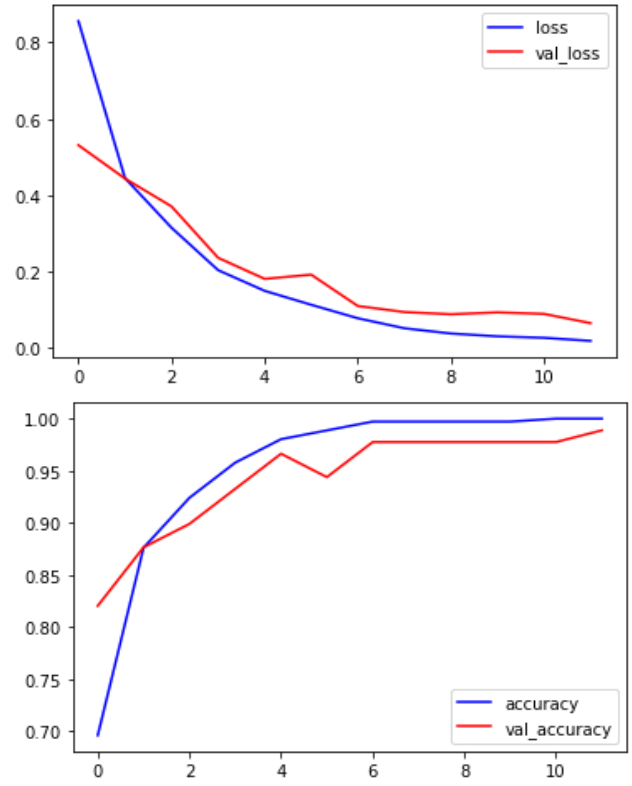


Fig. 6. DenseNet121+LSTM Loss vs. Epoch.

A. Result Comparison

This section discusses the outcomes of our experiments as well as the accuracy of our models. We trained our models on both UCF50 and HMDB51 datasets and evaluated their performance. To determine the best activation functions, we tested several options and analyzed their accuracy.

Our study focuses on comparing the accuracy of three models, ConvLSTM, LRCN, and DenseNet121+LSTM. Through our experiments, we found that the accuracy outcomes of the DenseNet121+LSTM model indicated its superiority over other models. Therefore, we recommend using the DenseNet121+LSTM model as a video action recognizer for human activity in video footage.

TABLE I. MODELS ACCURACY COMPARISON

Model	UCF50	HMDB51
ConvLSTM	77.87	59.13
LRCN	91.80	75.47
DenseNet121+LSTM	98.8	90.38

The accuracies for all three models on the UCF50 and HMDB51 datasets are presented in Table 1.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 06:09:57 UTC from IEEE Xplore. Restrictions apply.

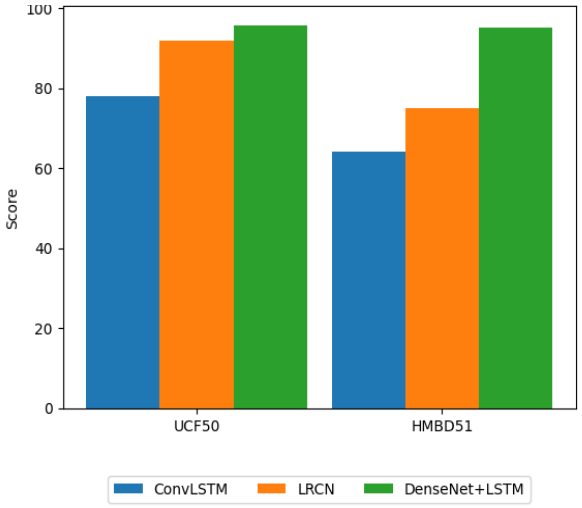


Fig. 7. Performance Across All Models for UCF50 and HMDB51.

Figure 7 demonstrates how our chosen models perform on two separate datasets.

The Confusion Matrix is used to report the accuracy for BaseballPitch, Basketball, BenchPress, and Biking classes to assess the performance of the DenseNet121+LSTM model. The Confusion Matrix offers a summary of the accurate and inaccurate predictions made for each class. The accuracy of classification for the best class is presented in Figure 8 as a matrix of percentages. These results provide insight into the strengths and weaknesses of the models.

Our model uses DenseNet121 + LSTM and achieves an accuracy of 98.8% on UCF50 and 90.38% on HMDB51. We only use 20 datasets from UCF50 and HMDB51 and apply simple data processing steps. This can produce better results than SOTA T3D + TCN[16] which uses DenseNet and reports an accuracy of 93.5% on UCF101 and 82.1% on HMDB51. Our model has a simpler architecture and requires less computational cost.

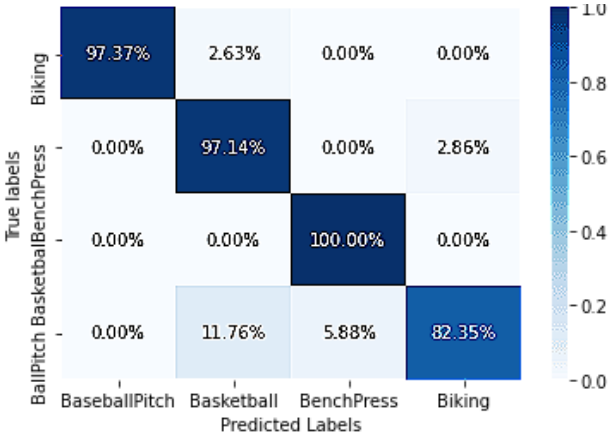


Fig. 8. Confusion Matrix for DenseNet121 Model.

In Figure 9, we present a comparison of the accuracy of our proposed DenseNet51+LSTM model with the performance of existing models on the UCF50 and HMDB51 datasets. Our model shows higher accuracy than the other models for both datasets.

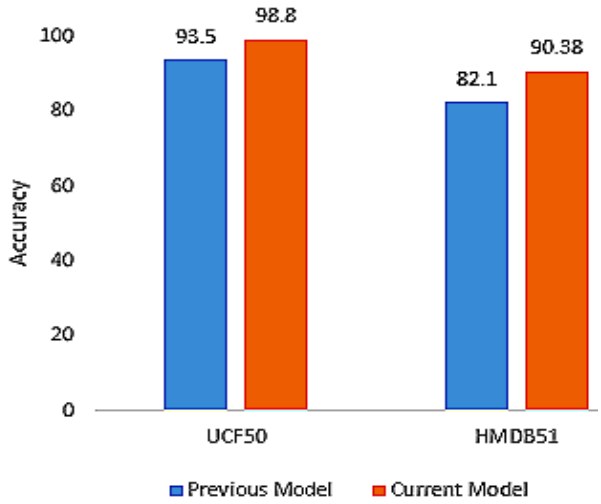


Fig. 9. Accuracy comparison between the suggested model and existing models.

V. CONCLUSION

Human Activity Recognition is an interesting research topic in the domain of computer vision. This paper introduced a Convolutional Neural Network-based architecture that recognizes human behavior and activities. Many deep learning architectures, such as Simple CNN, CNN+LSTM Hybrid Models like ConvLSTM, LRCN, and DenseNet121+LSTM are utilized as training models. We conducted comprehensive tests using two publicly available datasets. According to our tests conducted, DenseNet121+LSTM has an accuracy of 98.8 percent for UCF50 and 90.8 percent for HMDB51. Following that, LRCN reports an accuracy of 91.8 percent for UCF50 and 75.47 percent for HMDB51, respectively. The basic ConvLSTM shows accuracy for the UCF50 dataset as 77.87 percent and for the HMDB51 dataset as 59.13 percent. However, considering time complexity, we have suggested the DenseNet121+LSTM and LRCN model to construct a real-time Human Activity Recognition System. We tested and assessed our proposal in this research utilizing videos from two separate datasets. Although the design provides a good level of accuracy, implementing a real-time prototype system must be developed.

REFERENCES

[1] V. Bianchi, M. Bassoli, G. Lombardo, P. Fornacciari, M. Mordonini, and I. De Munari, "IoT Wearable Sensor and Deep Learning: An Integrated Approach for Personalized Human Activity Recognition in a Smart Home Environment," IEEE Internet of Things Journal, vol. 6, no. 5, pp. 8553-8562, Oct. 2019.

[2] D. Liciotti, M. Bernardini, L. Romeo, and E. Frontoni, "A Sequential Deep Learning Application for Recognising Human Activities in Smart Homes," Neurocomputing, vol. 396, pp. 501–513, 2020.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 06:09:57 UTC from IEEE Xplore. Restrictions apply.

[3] V. Jacquot, Z. Ying, and G. Kreiman, "Can Deep Learning Recognize Subtle Human Activities?," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 14232-14241.

[4] S. Kumar and A. Mallik, "COVID-19 Detection from Chest X-rays Using Trained Output Based Transfer Learning Approach," Neural Process Lett, vol. 1, pp. 1-24, Oct. 2022.

[5] S. Kumar, A. Mallik, A. Kumar, J. Del Ser, G. Yang, "Fuzz-ClustNet: Coupled Fuzzy Clustering and Deep Neural Networks for Arrhythmia Detection from ECG Signals," Computers in Biology and Medicine, vol. 153, pp. 106511.

[6] S. Kumar, A. Mallik, B. S. Panda, "Influence Maximization in Social Networks Using Transfer Learning via Graph-based LSTM," Expert Systems with Applications, vol. 212, p. 118770, 2023.

[7] S. Anand, A. Mallik, and S. Kumar, "Integrating Node Centralities, Similarity Measures, and Machine Learning Classifiers for Link Prediction," Multimedia Tools and Applications, vol. 81, no. 27, pp. 38593-38621, Nov. 2022.

[8] S. Kumar, A. Mallik, and S. S. Sengar, "Community Detection in Complex Networks Using Stacked Autoencoders and Crow Search Algorithm," The Journal of Supercomputing, vol. 79, no. 3, pp. 3329-3356, Feb. 2023.

[9] Puneeth, Sowmya, S. Ziyan, and M. R. Manu, "Human Activity Recognition Using Machine Learning," International Journal of Research in Engineering, Science and Management, vol. 4, no. 7, pp. 253-255, Jul. 2021.

[10] S. K. Polu, "Human Activity Recognition on Smartphones Using Machine Learning Algorithms," International Journal for Innovative Research in Science & Technology, vol. 5, no. 6, Nov. 2018.

[11] A. Subasi, K. Khateeb, T. Brahimi, and A. Sarirete, "Human Activity Recognition Using Machine Learning Methods in a Smart Healthcare Environment," in Innovation in Health Informatics, pp. 123-144, 2020.

[12] S. Wan, L. Qi, X. Xu, C. Tong, and Z. Gu, "Deep Learning Models for Real-time Human Activity Recognition with Smartphones," Mobile Networks and Applications, vol. 25, no. 2, pp. 743-755, Apr. 2020.

[13] C. I. Orozco, E. Xamena, M. E. Buemi, and J. J. Berlles, "Human Action Recognition in Videos Using a Robust CNN LSTM Approach," Ciencia y tecnología, no. 20, pp. 23–36, 2020.

[14] A. Stergiou and R. Poppe, "Spatio-Temporal FAST 3D Convolutions for Human Action Recognition," 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), Boca Raton, FL, USA, 2019, pp. 183-190.

[15] X. Mu, Z. Liu, J. Liu, H. Li, Y. Li, and Y. Li, "DC3D: A Video Action Recognition Network Based on Dense Connection," 2022 Tenth International Conference on Advanced Cloud and Big Data (CBD), Guilin, China, 2022, pp. 133-138.

[16] A. Diba, M. Fayyaz, V. Sharma, A. H. Karami, M. M. Arzani, R. Yousefzadeh, and L. Van Gool, "Temporal 3D ConvNets: New Architecture and Transfer Learning for Video Classification," arXiv preprint arXiv:1711.08200v1 [cs.CV], Nov. 2017.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 06:09:57 UTC from IEEE Xplore. Restrictions apply.

A Study of Recent Advancements in Deep Learning for Natural Language Processing

Aditya Raj, Rajni Jindal, Aditya Kumar Singh, Ajay Pal
Department of Computer Science and Engineering
Delhi Technological University
New Delhi, India
adityaraj_2k19co032@dtu.ac.in, rajnijindal@dce.ac.in, adityakumarsingh_2k19co031@dtu.ac.in, ajaypal_2k19co036@dtu.ac.in

2023 IEEE World Conference on Applied Intelligence and Computing (AIC) | 979-8-3503-1006-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/AIC57670.2023.10263979

Abstract*—Natural Language Processing has gained interests and progress over the last decade. Its success can be attributed to the recent advances made in the field of Deep Learning. This review provides a detailed overview of recent advancements in deep learning techniques specially in the last decade (2013-2023) used for Natural Language Processing, including the architectural changes and capability enhancements. In this review we discussed the advancements in application areas like language modeling, machine translation and dependency parsing, while also comparing the models presented in each application areas. We compared the language models against their size, parameters, deep learning technique, features and drawbacks; dependency parsing models against their unlabeled and labeled accuracy scores, and machine translation models against their BLEU scores for WMT'14 English-French and English-German tasks. We also explored the potential challenges and limitations in these areas. Finally, we reviewed the current state-of-the-art multimodal model like GPT-4 and discussed how these models are evolving towards a general form of artificial intelligence.

Keywords*—Natural Language Processing, Deep Learning, Feed-Forward Neural Network, Recurrent Neural Network, Long Short-Term Memory, Attention, Transformer, Word representations, Language Modeling, Machine Translation.

I. INTRODUCTION

The deep learning field has experienced tremendous growth in recent times due to revival of old techniques like Markov models, recurrent neural networks (RNNs) [3], long short-term memory (LSTM) [2], as data and computation power improved. It has widened the scope of NLP with its ability to learn more complex representations than shallow neural networks [1]. This literature review aims to study how the advancements in deep learning techniques have led to the current state-of-the-art models. For conducting this survey we referred to well-known archives like Google Scholar , arXiv and ACL anthology. We used keywords such as DL for NLP, LLM, NMT, Dependency Parsing, etc. to identify relevant papers. Finally, we selected the important papers after screening their titles, abstracts and findings. We reviewed the papers belonging to three main application areas: Dependency Parsing, Machine Translation, and Language Modeling. Advancements made in these areas were discussed and their models were compared. We discussed the emerging challenges and limitations of LLMs. The goal of our work is that by the end of this review, our readers would have an idea as to how deep learning-based research in NLP is moving towards a long-term goal of AI, a single AI model capable of doing many particular tasks simply by being instructed upon.

Language modeling has experienced progress from n-gram models to FNN [5], RNN [17] based models, while Dependency Parsing has advanced from simple neural networks [28,29] to neural graph [30] and encoder-decoder [23] based methods. Machine translation has experienced tremendous growth mainly due to the arrival of new techniques like attention [31] and transformers [25]. These models have paved the way for a more advanced technique pre-training, which is capable of learning general level knowledge of a language, simply by being trained upon an extremely large amount of texts. Large Pre-Trained Language Models[38,39,41,42,43] are capable of performing multiple tasks and have shown to learn new information with just a few examples. Finally, we will see how all this research has led to the current state-of-the-art multimodal models, which are pre-trained on different modalities like texts, images.

II. OVERVIEW OF DEEP LEARNING TECHNIQUES

A. Feed-Forward NN (FNN)

A simple FNN consists of input, hidden and output layer. Given an input vector (e), output is given by $y = g(Uh + b_2)$, where hidden layer (h) is $h = f(We + b_1)$. The model is trained using backpropagation[11], and its complexity depends on the number of layers and inputs.

B. RNN

RNN (as depicted in Figure 1) is similar to a FNN except that at each step there are two inputs to hidden layer: the output of the previous layer (h_{t-1}) , and the input at that step (e_t). The hidden layer (h_t) is calculated as: $h_t = f(W_h h_{t-1} + W_e e_t + b_1)$, and the output is given by $y = g(Uh_t + b_2)$. RNNs unlike FNNs may encode the information present in the sequence itself because they have memory of the prior output. This feature makes it useful for tasks like language modeling, sentiment classification, summarization, machine translation and question answering.

RNNs suffer from vanishing and exploding gradient problem. Since they are trained using backpropagation it involves computing gradient of loss function, and if we propagate further back gradients become smaller and approaches a value close to zero. This is called vanishing gradient and it is a problem because our parameters are not

979-8-3503-1006-1/23/$31.00 ©2023 IEEE
DOI: 10.1109/AIC.2023.49

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 06:09:31 UTC from IEEE Xplore. Restrictions apply.

updated properly. The reverse process where the product of gradients becomes larger is called exploding gradient, and it can cause overflows (e.g., Inf, NaN) and bad updates to the parameters. Bengio et al. showed that it is difficult to learn long-term dependencies for any gradient based learning algorithm [36].

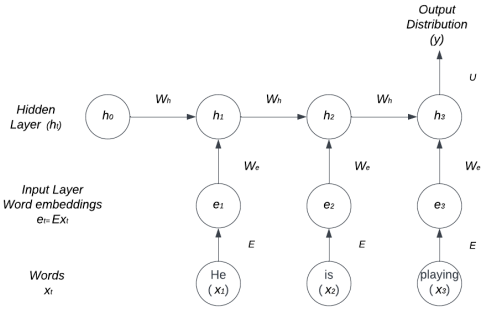


Fig. 1. A simple RNN based Language Model for a sentence "He is playing _____", where the model finds the probability distribution of the next word.

C. LSTM

Long Short-Term Memory proposed by Hochreiter and Schmidhuber in [2], is a special type of RNN. In addition to the hidden state it also has a cell state (c_t), which has the capacity to retain data for extended periods of time. It has three gates to control the flow of data into and out of the cell. Output Gate ($o_t = f(W_o h_{t-1} + U_o e_t + b_o)$) determines what part of cell are output to the hidden state, while the Input Gate ($i_t = f(W_i h_{t-1} + U_i e_t + b_i)$) determines which information should enter the cell state, and the Forget Gate ($f_t = f(W_f h_{t-1} + U_f e_t + b_f)$) decides the information from the previous state cell that will be memorized and the information that will be removed which is no longer useful. The hidden state is calculated using the element-wise multiplication of the output gate and the cell state. Cell state is computed with the help of new cell state ($\tilde{c_t}$),

$$\tilde{c_t} = g(W_c h_{t-1} + U_c e_t + b_c)$$
$$c_t = f_t o\, c_{t-1} + i_t o\, \tilde{c_t}$$
$$h_t = o_t o\, g(c_t)$$

where o represent element wise multiplication. LSTMs retain far-away information more easily than simple RNNs, making it a powerful technique for NLP tasks.

D. Bi-RNNs

Bi-Directional RNNs are a special type of RNNs which maintains two hidden layers: backward ($\overleftarrow{h_t} = RNN_{\leftarrow}(\overleftarrow{h_{t+1}}, e_t)$) and forward ($\overrightarrow{h_t} = RNN_{\rightarrow}(\overrightarrow{h_{t-1}}, e_t)$). These two hidden layers are concatenated to get the hidden layer for Bi-RNNs ($h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$). These are only applicable when the entire input sequence is present, and so it is not applicable for language modeling as it only has backward context.

E. Multi-layer RNNs

Simple RNNs have linear dimension as we move only in one direction, however we can increase the dimension of these RNNs by stacking multiple simple RNNs in layers. These are Multi-layer RNNs and are used to compute complex representations.

F. Encoder-Decoder

An encoder-decoder model is made up of two RNNs, usually LSTMs. The encoder represents the input text into a fixed size context vector and then the decoder generates the output using the context from the input. When the input and the output text are sequences it is called a seq2seq model. These models are trained using backpropagation, and the updates are made for the whole system end-to-end. Figure 2 represents a multilayer seq2seq model. Seq2seq models can be used for various other NLP tasks like- summarization, parsing, question answering and even code generation from pseudo algorithms written in natural language.

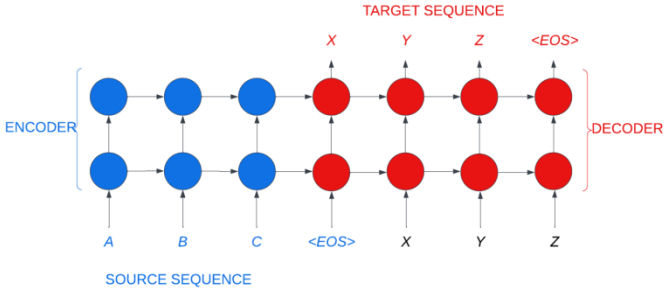


Fig. 2. A multilayer recurrent encoder-decoder network for neural machine translation of source sequence ABC into target sequence XYZ. <EOS> represents the end-of-sentence token.

G. Transformers

In the subsections discussed above, each hidden state of a recurrent network depends on the previous hidden states. This means that state h_t can't be computed before state h_{t-1}. Since modern GPUs can perform many parallel tasks there is not much parallel computation it can do for recurrent networks. This inhibits training on large corpuses. Another problem of RNNs is that it is hard to learn long-distance dependencies due to the vanishing-gradient problem. Bahdanau et al. proposed a seq2seq model based on attention [31]. Attention mechanism allowed direct interact interaction between two elements. Transformer [25] by Vaswani et al. is similar to seq2seq, but composed of a building block based entirely on self-attention. Self-attention is an attention mechanism that attends to words within a single sequence. The techniques that are essential for the building block of Transformers are:

Self-attention: The key-query-value self-attention mechanism is where we look up the information in key-value pair and select the value of the key most like the query. Consider an input sequence ($\boldsymbol{x_1, x_2, \ldots \ldots x_n}$), the key, query and value for each input is given as: $\boldsymbol{k_i = Kx_i}$, $\boldsymbol{q_i = Qx_i}$, and $\boldsymbol{v_i = Vx_i}$. The output is computed as: $\boldsymbol{output = softmax(XQ(XK)^T)}$, where X = $[\boldsymbol{x_1; x_2; \ldots \ldots; x_n}]$.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 06:09:31 UTC from IEEE Xplore. Restrictions apply.

Vaswani et al. found that it useful to perform multi-head attention instead of single attention, as it allowed the model to jointly attend to information from different representation subspaces at different positions. He applied self-attention multiple times each with different key, query and value transformations of the same input. Finally, the output values are scaled and then concatenated to get an output matrix.

$$\text{output}_\text{l} = \text{softmax}\left(\frac{XQ_\text{l}(XK_\text{l})^\text{T}}{\sqrt{d/h}}\right)XV_\text{l}$$

$$\text{output} = [\text{output}_1,\ldots\ldots,\text{output}_\text{h}]$$

where h is the number of attention heads, and d is the dimensionality.

Cross-attention: In this keys and values are used from a different sequence than the queries. Here, decoders use cross attention on encoder outputs. Keys and Values are defined using output vectors of the encoder ($\mathbf{h_1}, \mathbf{h_2}, \ldots\ldots \mathbf{h_n}$): $\mathbf{k_i} = \mathbf{Kh_i}$, $\mathbf{v_i} = \mathbf{Vh_i}$. Queries are defined using input vectors of the decoder ($\mathbf{s_1}, \mathbf{s_2}, \ldots\ldots \mathbf{s_n}$): $\mathbf{q_i} = \mathbf{Qs_i}$.

Residual Connection: Introduced by He et al. in [26], it is a technique to improve the training of models, simply by adding the input of a layer to the output of that layer: $\mathbf{y} = \mathbf{x} + \mathbf{Layer}(\mathbf{x})$, where x is the input of the layer, and y is the output of the layer.

Layer normalization: It is a technique to reduce the training time of the models and was introduced by Ba et al. in [34]. The idea is to cut down on uninformative variation in hidden vector values by normalizing to unit mean and standard deviation within each layer.

The building blocks represent a single layer of the encoder and decoder. Each encoder layer has two sublayers, consisting of multi-head self attention mechanism and FNN. Each sublayer follows a combination of residual connection and layer normalization. Decoder has one additional sublayer to the encoder. It inserts a sublayer performing multi-headed cross attention over the output of the encoder stack, between the masked multi-head self attention and feed-forward network sublayers. Transformers have shown to not just be useful for machine translation but question answering and summarization tasks as well.

III. Deep Learning for NLP

Here we will review in detail the main application areas where deep learning techniques are used for NLP. We will also review the challenges and limitations in these areas, and discuss the potential future directions.

A. Dependency Parsing

Dependency Parsing analyses the dependency structure and outputs a dependency tree, by creating a mapping from the input sentence to its dependency graph. Neural Based approach learns dense and compact feature representations, and uses DL models to predict a transition sequence. Chen and Manning [28], and Andor et. al. [29] performed dependency parsing using FNNs. A Neural graph-based dependency parser [30] was presented by Dozat and Manning, which achieved even better accuracy than Andor et. al., but was slower($O(n^2)$). Li et. al. presented a seq2seq dependency parser using attention and achieved on par results with state-of-the-art parsers [23]. Dependency parsing is evaluated using mainly two metrics: UAS which is the ratio of correct unlabeled dependencies to total number of unlabeled dependencies, and LAS which is the ratio of correct labeled dependencies to total number of labeled dependencies. Table 1 compares the DL models performing Dependency parsing based on their UAS and LAS.

TABLE I. UAS and LAS for different models. Experiments conducted on English Penn Treebank (PTB).

MODEL	UAS	LAS	DL TECHNIQUE
Chen and Manning, 2014 [28]	92.0	89.7	FNN
Andor et. al, 2016 [29]	94.61	92.79	FNN
Dozat and Manning, 2017 [30]	95.74	94.08	Neural Graph based
Li et. al, 2018 [23]	94.11	92.08	Seq2seq with attention

B. Machine Translation

It is the quintessential application of NLP, and involves translating a text from one language to another using a machine. Neural Machine Translation (NMT) involves a seq2seq model with a single end-to-end neural network. First seq2seq model was presented by Sutskever et al. for English to French translation [4]. Bahdanau et al. showed that the use of a fixed-length vector is a bottleneck in improving the performance of the encoder–decoder architecture and proposed to extend the performance using attention mechanism [31]. Luong et al. further proposed two classes of attention [32]: a global approach attends to all source words and a local approach which attends to only a subset of source words at a time. Google Neural Machine Translation System (GNMT) [27] uses a deep 8 layer encoder and decoder LSTM network. A multilingual NMT based on GNMT enables "zero-shot translation[33]. Vaswani et al. achieved state-of-the-art results using transformers for standard WMT'14 tasks[25]. Transformer based models outperform recurrent networks mainly due to its ability to learn long range dependencies. A key factor affecting this ability is the length of the paths that the forward and backward signals take in the network. The shorter the path the easier it is to learn long range dependencies [35]. Introduction of LLMs like BERT and GPT led to various attempts at incorporating these for NMT. Zhu et. al. incorporated BERT into NMT [6], while Han et. al. used GPT models' zero-shot translation capability for NMT [7].

Machine translated texts were evaluated using BLEU [12] which evaluates the precision score of the machine translated text against some reference human translations, while newer methods like Comet [9] and BLEURT [8] trains machine

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 06:09:31 UTC from IEEE Xplore. Restrictions apply.

translation evaluation models. Table 2 provides a comparison of the BLEU scores of the NMT models discussed above.

C. Language Modeling

Language modeling involves the task of predicting the next word given the context. Formally, it aims to learn joint probability function of sequence of words in a sentence, which can be represented by conditional probability of the next word given the context. ANN language model proposed by Bengio et. al. used a simple FNN with fixed length context [5]. In this approach only a limited amount of preceding words were used to predict the next word.

TABLE II. BLEU SCORES FOR NMT MODELS FOR WMT'14 EN-FR AND EN-DE TASKS

model	en-fr	en-de	dl technique
sutskever et. al. [4], 2014	36.5		seq2seq
bahdanau et. al., 2014 [31]	36.15		seq2seq with attention
luong et. al, 2015 [32]		23.0	stacked lstm encoder-decoder
deep-att + rl ensemble, 2016 [15]	40.4	20.6	stacked lstm encoder-decoder
gnmt + rl ensemble, 2016 [27]	41.16	26.30	stacked lstm encoder-decoder
transformer,2017 [25]	41.8	28.4	transformer
zhu et. al, 2020 [6]	43.78	30.75	bert based
han et. al, 2021 [7]	42.1		gpt-3 based

This approach was able to solve both the sparsity and curse of dimensionality problem of n-grams. But still the context window can't be large enough, as the model size depends on the context length.

RNNs solved this problem as they do not use limited size of context. They can process inputs of any length without increase in model size. Even though RNNs performed better than FNNs, it also suffered from vanishing and exploding gradient problem, which affected its ability to properly access very far away information. LSTMs fared better than RNNs in preserving information from very far away contexts, which made it popular for language modeling. Dai et al. introduced pre-training for the LSTM model [37] , where the model is pre-trained on language modeling. They found that this pre-training step significantly stabilized the training of LSTMs, and also concluded that additional unlabeled data could improve the generalizability of the model. The introduction of transformers and its parallelizability has allowed for more efficient pre-training and less time-consuming, which has made it the de-facto strategy nowadays. First the entire model is pre-trained on some data rich task using unlabeled text and then the pre-trained model is used to further train the model on some supervised tasks (fine-tuning). This type of learning is called transfer learning. Unsupervised pre-training is attractive because of the abundance of unlabeled text, which makes it possible for the model to learn general purpose knowledge that can then be transferred to downstream tasks having small amounts of data. Pre-training using transformers has been done mainly in three ways:

Pre-training on transformer decoder: Radford et al. introduced generative pre-training(GPT) [38] of a language model on a large corpus of unlabeled text to learn initial parameters of the model, followed by fine-tuning on supervised tasks. A key emergent ability of GPT-2 [43] was zero-shot learning, the ability to perform downstream tasks without any gradient updates or context examples. With the introduction of the GPT-3 [42] model by Brown et al. , it was shown that their model performed few-shot learning without any gradient updates or fine-tuning. The model simply learned from the examples provided within their contexts.

Pre-training on transformer encoder: Since GPT was pre-trained in only a single direction, it could give inconsistent results in tasks where it is important to incorporate contexts from both directions. Devlin et al. introduced BERT [39] to pre-train bidirectional representations from unlabeled text. It can then be fine-tuned using only one additional output layer. For pre-training, it used a masked language model (MLM) to randomly mask some of the words from the input and predict these words based on its context. Liu et al. showed that training a BERT model with more data, and longer sequences can significantly improve their performance[40].

Pre-training on transformer encoder-decoder: T5 [41] proposed by Raffel et al. used a transformer encoder-decoder to pre-train it on a text-to-text problem. The encoder learns bidirectional context and the decoder trains the whole model through language modeling. They found encoder-decoders to perform better than decoders and span corruption to work better than language modeling for their tasks. LMs are important because it is usually a part of many NLP tasks, e.g., in tasks like machine translation, question answering, etc., where multiple alternative sequences are produced. LMs are then used to score the sequences and choose the most likely sequence.

D. Limitations and Challenges

Even though LLMs have been shown to generate coherent texts in natural language exceptionally well, several recent papers have argued that LLMs are limited in capturing important aspect of linguistic meaning. Asher et. al. proved that these limitations occur because LLMs can't learn certain fundamental semantic properties [18]. As a result outputs may contain inaccuracies and errors. Yenduri et al. discussed the emerging challenges with LLMs like the GPT series [22]. Some of the challenges discussed were Biases, Ethical and Privacy Concerns, Multimodal and Multilingual capabilities. LLMs are trained on vast amounts of internet data which might be faulty, have biases, contain personal views or sensitive information. Therefore, LLMs might generate biased texts imitating the training data, raise ethical and privacy concerns. Multilingual capabilities are a challenge due to the differences in syntax and vocabulary across languages. LLMs could also be used for harmful purposes like creating fake news or deep fakes. LLMs also face the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 06:09:31 UTC from IEEE Xplore. Restrictions apply.

challenge of reducing their carbon footprints as they are run on GPUs, TPUs which consume lot of energy.

A. Moving towards general form of AI

Despite the challenges in developing multilingual models, Kondratyuk et al. presented UDify [17], a multilingual multi-task model, which is capable of predicting dependency trees for all 124 Universal Dependencies treebanks and across 75 languages using multilingual BERT (mBERT) [44].

TABLE III. COMPARISON OF VARIOUS LLMS

Model	Size	Parameters	Year	Input	DL Technique	Features	Drawbacks
T-5 [41]		11 B	2020	text	Transformer encoder-decoder	Introduced unified framework to convert all text based language problems into a text-to-text format	Large number of parameters, computationally expensive
GPT-1 [38]	12 layer decoder	117 M	2018	text	Transformer decoder	Used mostly for simple NLP tasks	Limited data, can't perform complex tasks, limited applications
GPT-2 [43]	10 times GPT-1	1.5 B	2019	text	Transformer decoder	Zero-shot learning capabilities	Limited data diversity, expensive computation
GPT-3 [42]	100 times GPT-2	175 B	2020	text, images and tables	Transformer decoder	Improved Zero-shot and one-shot learning capabilities, Useful for multiple NLP tasks	Limited control, limited data diversity, ethical concerns
GPT-4 [21]		100 T	2023	text, images, tables and audio	Transformer decoder	Improved Creative and technical writing capabilities	
BERT [39]	24 layer encoder	340 M	2018	text	Transformer encoder	Bi-directionality allows it to be fine-tuned with just one additional output layer	Expensive computation, limited data diversity, ethical concerns and biases

LLMs have been shown to perform in-context learning with the help of prompts (prompt-learning), but struggle to learn complex tasks like arithmetic. Chain-of-thought prompting which involves generating a chain of thought has been shown to significantly improve the ability of LPLMs to perform complex reasoning [10]. Instruction fine-tuning is another method to improve the in-context learning capabilities of LPLMs. It collects examples of (instruction, output) pairs across many tasks and fine-tunes LMs. Chung et al. explored instruction fine-tuning by performing fine-tuning on chain-of-thought data [13,14]. Wei et. al. showed that fine-tuning language models can significantly improve zero-shot performances on unseen tasks [16]. It is expensive to collect ground truth data for different tasks, and also there is a misalignment between the objective of a language model and the objective of satisfying human preferences. This problem was improved upon by fine-tuning language models from human feedback using reward models. A Reward Model (RM) is trained using supervised learning to maximize human preferred results from huge datasets of human preferences for a given task. Reinforcement learning is used to train a policy to maximize the RM's score. Stiennon et al. showed that using Reinforcement Learning from Human Feedback (RLHF) for summarization tasks significantly improved model's performance and its generalizability to new domains [20]. RLHF was further employed in InstructGPT [45] and ChatGPT. Current state-of-the-art model GPT-4 [21] is a multimodal LLM which can handle inputs with multiple modalities like images, texts, and audios.

IV. CONCLUSION

Deep Learning techniques have truly revolutionized the field of NLP. Even though ANN for language modeling was introduced way back in 2003 [5], it was the efforts of Collobert et al. in 2011 [24] that sparked DL research in NLP. 2014 marked a significant year in machine translation with the introduction of seq2seq and seq2seq with attention models. 2017 was the year of breakthroughs with the introduction of neural graph based dependency parser [30] which is still the state-of-the-art model for dependency parsing, and Transformer [25]. Table 2 helps us to understand that even though transformer models have greater time and space complexity than seq2seq models they outperform them mainly due to their shorter path lengths and parallelizability. The first large pre-trained language model was introduced in 2018. GPT [38] accepted only texts as inputs and also had limited applications. However within a span of just 5 years it evolved into the current state-of-the-art multimodal model GPT-4 [21], as LLMs were trained on more data and became larger in size over the years (Table 3). GPT-4 is able to perform multiple tasks ranging from complex reasonings, arithmetics, human-like conversations, image captionings, etc. with just being instructed upon.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 06:09:31 UTC from IEEE Xplore. Restrictions apply.

It represent a different class of models than we saw earlier with their ability to obtain a holistic understanding of the content. While the eventual possibilities for these models are only dimly understood, they have been proved to lack a human-level ability for careful logical or causal reasoning [18]. But the ability of these models to deploy knowledge gained from a huge amount of training data to many different runtime tasks will make these models more powerful, and they will give people in the coming decade their first glimpses of a more general form of artificial intelligence [19].

REFERENCES

[1] Y. Bengio, Y. LeCun. "Scaling learning algorithms towards AI. In: Large-Scale Kernel Machines", MIT Press, 2007

[2] S. Hochreiter, J. Schmidhuber. "Long Short-term Memory".Neural Computation,(1997) 9 (8): 1735–1780, Nov 1997.

[3] D. E. Rumelhart, G. E. Hinton, R. J. Williams. "Learning internal representations by error propagation". Technical report, DTIC Document, 1985.

[4] I. Sutskever, O. Vinyals, Quoc V. Le. "Sequence to Sequence Learning with Neural Networks". https://arxiv.org/pdf/1409.3215.pdf

[5] Y. Bengio, R. Ducharme and P. Vincent. 2003. "A neural probabilistic language model". Journal of Machine Learning Research, 3:1137-1155

[6] Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, Tie-Yan Liu. "INCORPORATING BERT INTO NEURAL MACHINE TRANSLATION". arXiv:2002.06823v1

[7] Jesse Michael Han, Igor Babuschkin, Harrison Edwards, Arvind Neelakantan, Tao Xu, Stanislas Polu, Alex Ray, Pranav Shyam, Aditya Ramesh, Alec Radford, Ilya Sutskever. "UNSUPERVISED NEURAL MACHINE TRANSLATION WITH GENERATIVE LANGUAGE MODELS ONLY". arXiv:2110.05448v1

[8] T.Sellam, D. Das, A. P. Parikh. "BLEURT: Learning Robust Metrics for Text Generation". arXiv:2004.04696v5

[9] R. Rei, C. Stewart, A. C Farinha, A. Lavie. "COMET: A Neural Framework for MT Evaluation". https://aclanthology.org/2020.emnlp-main.213.pdf

[10] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E.H. Chi, Q.V. Le, D. Zhou. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models". arXiv:2201.11903v6

[11] D. E. Rumelhart, G. E. Hinton, R. J. Williams. "Learning internal representations by backpropagating errors". Nature, 323:533.536, 1986.

[12] Kishore Papineni, Salim Roukos, Todd Ward,Wei-Jing Zhu. "BLEU: a Method for Automatic Evaluation of Machine Translation". https://aclanthology.org/P02-1040.pdf.

[13] Dharma KC, Clayton T. Morrison. "Neural Machine Translation for Code Generation". arXiv:2305.13504v1

[14] H.W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S.S.Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, J. Wei. "Scaling Instruction-Finetuned Language Models". arXiv:2210.11416v5

[15] Jie Zhou, Ying Cao, Xuguang Wang, Peng Li, Wei Xu. "Deep Recurrent Models with Fast-Forward Connections for Neural Machine Translation". https://aclanthology.org/Q16-1027.pdf

[16] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. "FINETUNED LANGUAGE MODELS ARE ZERO-SHOT LEARNERS". arXiv:2109.01652v5

[17] D. Kondratyuk, M. Straka. "75 Languages, 1 Model: Parsing Universal Dependencies Universally". arXiv:1904.02099v3.

[18] Nicholas Asher, Swarnadeep Bhar, Akshay Chaturvedi, Julie Hunter, Soumya Paul. "Limits for learning with Language Models". arXiv:2306.12213v1

[19] C.D. Manning."Human Language Understanding & Reasoning". Daedalus, 151(2):127–138.

[20] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, Paul Christiano. "Learning to summarize from human feedback". arXiv:2009.01325v3

[21] OpenAI. 2023. GPT-4 Technical report. arXiv preprint arXiv:2303.08774

[22] Gokul Yenduri, Ramalingam M, Chemmalar Selvi G, Supriya Y, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, Deepti Raj G, Rutvij H Jhaveri, Prabadevi B, Weizheng Wang, Athanasios V. Vasilakos, and Thippa Reddy Gadekallu. "GPT (Generative Pre-trained Transformer) – A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions". arXiv:2305.10435v2

[23] Z. Li, J. Cai, S. He, H. Zhao. "Seq2seq Dependency Parsing". https://aclanthology.org/C18-1271.pdf

[24] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa. "Natural Language Processing (Almost) from Scratch". Journal of Machine Learning Research, 12:2493- 2537, 2011

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Łukasz Kaiser, Illia Polosukhin. "Attention Is All You Need". arXiv:1706.03762v5

[26] K. He, X. Zhang, S. Ren, J. Sun. "Deep Residual Learning for Image Recognition". arXiv:1512.03385v1

[27] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Łukasz Kaiser, Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. "Google's neural machine translation system: Bridging the gap between human and machine translation". arXiv preprint arXiv:1609.08144 (2016)

[28] D. Chen, C. D. Manning. "A Fast and Accurate Dependency Parser using Neural Networks". https://www.emnlp2014.org/papers/pdf/EMNLP2014082.pdf.

[29] D. Andor, C. Alberti, D. Weiss, A. Severyn , A. Presta, K. Ganchev, S. Petrov and M. Collins. "Globally Normalized Transition-Based Neural Networks". https://arxiv.org/pdf/1603.06042.pdf

[30] T. Dozat, C. D. Manning."Deep Biaffine Attention For Neural Dependency Parsing". https://arxiv.org/pdf/1611.01734.pdf

[31] D. Bahdanau, K. H. Cho, Y. Bengio. "Neural Machine Translation By Jointly Learning To Align And Translate". https://arxiv.org/pdf/1409.0473.pdf

[32] Minh-Thang Luong, H. Pham, C. D. Manning. "Effective Approaches to Attention-based Neural Machine Translation". https://arxiv.org/pdf/1508.04025.pdf

[33] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, J. Dean. "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation". https://arxiv.org/pdf/1611.04558.pdf

[34] J. L. Ba, J. R. Kiros, G. E. Hinton."Layer Normalization". arXiv:1607.06450v1

[35] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber. "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies" https://www.bioinf.jku.at/publications/older/ch7.pdf.

[36] Y. Bengio, P. Simard, P. Frasconi. "Learning Long-Term Dependencies with Gradient Descent is Difficult".

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 06:09:31 UTC from IEEE Xplore. Restrictions apply.

http://www.comp.hkbu.edu.hk/~markus/teaching/comp7650/tnn-94-gradient.pdf

[37] A. M. Dai, Q. V. Le. “Semi-supervised Sequence Learning”. arXiv: 1511.01432v1.

[38] A. Radford, K. Narasimhan, T. Salimans, T. Salimans. “Improving Language Understanding by Generative Pre-Training”. https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf

[39] J. Devlin, M. W. Chang, K. Lee, K. Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. arXiv:1810.04805v2

[40] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov. “RoBERTa: A Robustly Optimized BERT Pretraining Approach”. arXiv: 1907.11692v1

[41] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li , P. J. Liu. “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”. arXiv:1910.10683v3

[42] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei. “Language Models are Few-Shot Learners”. arXiv:2005.14165v4

[43] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever. “Language Models are Unsupervised Multitask Learners”.

[44] https://github.com/google-research/bert/blob/master/multilingual.md

[45] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, R. Lowe. “Training language models to follow instructions with human feedback”. arXiv:2203.02155v1.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 06:09:31 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Power Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPEL.2023.3319589

A Variable-DC-Offset Modulation Scheme for Improved Performance of a Differential-Mode Inverter

[1]Kartik Tank, [2]Shantanu Gupta, *Student Member IEEE*, [3]Man Mohan Garg, *Senior Member IEEE*, [4]Sudip K. Mazumder, *Fellow IEEE*, and [5]Moien Mohamadi, *Student Member IEEE*
[1,3]Department of Electrical Engineering, Malaviya National Institute of Technology Jaipur, India
[2,4,5]Department of Electrical and Computer Engineering, University of Illinois Chicago, USA
Email: [1,3]{2020ppd5492, mmgarg.ee}@mnit.ac.in, [2,4,5]{sgupt57, mazumder, mmoham65}@uic.edu

***Abstract*—In this paper, a novel continuous modulation scheme with variable-dc-offset (CMS-VDO) is proposed for a single-phase differential-mode inverter (DMI). The output voltage of the DMI is the difference in output voltages of two differentially connected dc-dc converters. CMS-VDO is so designed that the output voltage of the individual dc/dc converter is reduced without affecting the output voltage of the DMI. This reduced voltage yields reductions in circulating current, peak switch voltage, total harmonic distortion (THD), and improved efficiency compared to conventional modulation scheme with fixed dc offset (CMS-FDO). The experimental results compare and validate the improved inverter performance using CMS-VDO over CMS-FDO.**

***Index Terms*—Differential mode, efficiency, harmonic distortion, inverter, modulation, switch stress**

I. Introduction

SINGLE phase inverters are widely used to convert the dc electrical energy to an ac electrical energy compatible with household and industrial appliances. Further, inverters are also widely used to harvest the renewable (e.g., photovoltaic (PV)) energy [1]–[3] and support battery energy storage [4], where such energy sources have varying dc output voltage. Performance of an inverter is generally measured by its capability to efficiently generate a low-THD ac sinusoid. Widely used implementation of inverters are generally realized using full-bridge voltage-source inverters (VSIs) which are buck dc/ac converters [5]. The VSIs typically require four switches per phase, and the dc-source voltage should always be higher than the peak of the output ac voltage. The dc-voltage constraint of the buck converter is addressed by adding another dc/dc converter stage. For low-power inverter implementation, the two-stage inverter solution results in a higher component count, complex control, and lower overall efficiency [5].

Single-stage inverter solutions with a wide dc swing necessitates buck-boost capability. A z-source dc/ac converter [6] is a single-stage solution and typically needs two active switches. However, the input and output currents of inverter are discontinuous and require additional filter circuit. Quasi-z-source inverter [7], [8] allows continuous input and output current; however, the average- and ripple-current stresses on the inductor and the switches are significant resulting in a reduced inverter efficiency.

A differential-mode converter [9]–[11] consists ideally of two identical dc/dc converter modules. In a differential-mode inverter (DMI [9]), the two modules are connected in parallel at the dc side while at the ac side, the modules are differentially connected. The two modules are operated to generate a bipolar (i.e., an ac) output voltage. Such DMIs, based on buck, boost, and buck-boost converters, are discussed in [10]. The Ćuk-based implementation of such a DMI, referred in this paper as a differential-mode Ćuk inverter (DMCI) has step-up/-down and continuous input- and output-current capabilities [9]. As shown in Fig. 1, the DMCI topology requires 4 active switches and 8 passive components. The input and output inductors of the dc/dc modules can be integrated to reduce the magnetic component counts [12]–[15]. In addition, the DMCI also allows isolation feature as presented in [9], [15]–[17].

In [16], [18]–[20], a conventional continuous modulation scheme with fixed dc offset (CMS-FDO) for the DMI is discussed which governs the operation of the two dc/dc modules. The continuous operation of the dc/dc modules in CMS-FDO results in a lower total harmonic distortion (THD) as the even harmonics generated by non-linearities in the modules get largely cancelled [16]. However, the CMS-FDO-based DMI suffers from higher circulating power resulting in reduced efficiency and higher peak switch voltages due to the higher dc-offset voltage of the individual dc/dc converter module.

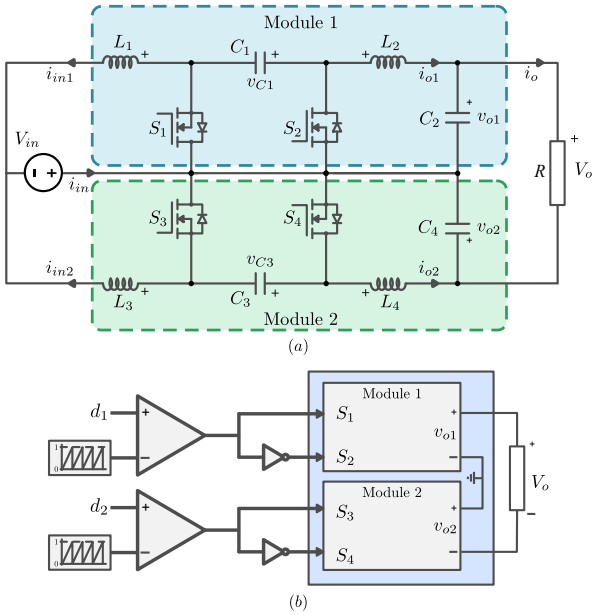


Fig. 1. (a) An illustration of a DMCI with two dc-dc converter modules, and (b) modulation mechanism of the DMCI.

In [16], a discontinuous modulation scheme (DMS) for an isolated DMI is outlined which results in higher inverter

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 03,2023 at 11:05:16 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Power Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPEL.2023.3319589

efficiency due to reduced circulating power and also lower switch voltage stress compared to that obtained using CMS-FDO. As the two modules are not operating continuously, the THD of the DMS-based open-loop DMI is higher and requires a close-loop control mechanism to limit the THD below 5%.

In [21], a novel continuous modulation scheme with variable-dc-offset modulation scheme (CMS-VDO) for DMI is outlined by the authors. CMS-VDO modulation for buck-, boost-, and buck-boost- derived DMI is discussed. A variable-dc voltage offset for both the dc/dc modules output is used to overcome the drawbacks of CMS-FDO yielding lower circulating power and lower switch voltage stress. Reference [10] presents simulation results to validate the operation of CMS-VDO. A comparison of CMS-FDO and CMS-VDO is also presented.

Unlike [21], in this article, detailed analyses of DMCI's circulating power and switch voltage stress are conducted for CMS-VDO to understand the dependencies of these parameters with varying DMCI operating conditions. Further, detailed experimental evaluations of the CMS-VDO- and CMS-FDO-based DMCI are also presented to support the analyses. Additionally, THD, efficiency, and unity- and non-unity-power-factor-load performances of the DMCI using the two modulation schemes are also demonstrated.

Overall, in Section II, an outline on CMS-FDO and CMS-VDO is provided. In Section III, comparison between CMS-FDO and CMS-VDO regarding voltage stress and circulating current is carried out. Section IV provides detailed experimental results associated with the two modulation schemes for the DMCI. Finally, Section V captures the conclusions.

II. Modulation Schemes for DMCI

The single-phase DMCI topology is shown in Fig. 1(a). DMCI consists of ideally two identical dc/dc Ćuk modules (Modules 1 and 2) connected in parallel at the dc side of the inverter while the outputs of the two modules are connected differentially at the ac side. The modulation mechanisms of the two dc/dc converter modules are shown in Fig. 1(b).

As shown in the figure, the duty cycle, d_1 is applied to Module 1 while the duty cycle, d_2 is applied to Module 2. Switch S_1 of Module 1 is operated using pulse-width modulation (PWM) associated with d_1 while S_2 is operated complementarily with the duty cycle $d_1' = 1 - d_1$. Similarly, switching of S_3 in Module 2 is governed by PWM based on the duty cycle d_2 while S_4 is switched complementarily based on the duty cycle, $d_2' = 1 - d_2$. As both the modules are supplied with the input voltage, V_{in}, the output voltages of the two modules in terms of respective duty cycles of the modules are the same as of Ćuk dc/dc converter [22] and given by the following expressions:

$$v_{o1} = \frac{d_1}{1-d_1} V_{in} \tag{1a}$$

$$v_{o2} = \frac{d_2}{1-d_2} V_{in}. \tag{1b}$$

The ac output voltage (v_o) of the DMCI is the voltage difference between the output voltages (v_{o1} and v_{o2}) of Modules 1 and 2 and is given by

$$v_o = v_{o1} - v_{o2} = \left(\frac{d_1}{1-d_1} - \frac{d_2}{1-d_2}\right) V_{in}. \tag{2}$$

Equation (2) demonstrates that the modulation scheme for the DMCI dictates the output of individual dc/dc converter modules to generate the desired sinusoid ac output voltage. The description of the CMS-VDO and CMS-FDO is discussed in the following subsection.

A. CMS-FDO

In CMS-FDO, the two dc/dc converter modules of the DMCI are so operated that $d_1 + d_2 = 1$ [9]. Applying this duty-cycle relationship of CMS-FDO and using (2), the following relation for the inverter output voltage is obtained:

$$v_o = V_m \sin(\omega t) = \left(\frac{d_1}{1-d_1} - \frac{1-d_1}{d_1}\right) V_{in} \tag{3}$$

where V_m is the required peak of the desired output ac (sinusoidal) voltage and ω is the angular frequency of the output voltage. Solving the quadratic equation in (3), d_1 (and hence $d_2 = 1 - d_1$) is found to be

$$d_1 = 0.5 + \frac{\sqrt{4 + \left(g\sin(\omega t)\right)^2} - 2}{2g\sin(\omega t)} \tag{4a}$$

$$d_2 = 0.5 - \frac{\sqrt{4 + \left(g\sin(\omega t)\right)^2} - 2}{2g\sin(\omega t)} \tag{4b}$$

where g is defined by V_m/V_{in}. Following (4), Fig. 2 captures the variation in d_1 and d_2 for CMS-FDO.

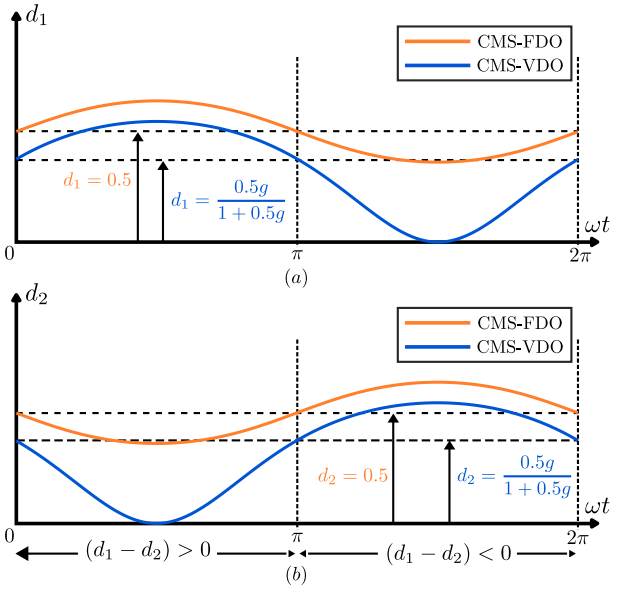


Fig. 2. Variation in (a) d_1 and (b) d_2 for CMS-FDO and CMS-VDO modulated DMCI over a line cycle.

Voltages, v_{o1} and v_{o1} are obtained using (1) and (4):

$$v_{o1} = \left(\frac{g\sin(\omega t)}{2} + \sqrt{1 + \left(\frac{g\sin(\omega t)}{2}\right)^2}\right) V_{in} \tag{5a}$$

$$v_{o2} = \left(-\frac{g\sin(\omega t)}{2} + \sqrt{1 + \left(\frac{g\sin(\omega t)}{2}\right)^2}\right) V_{in} \tag{5a}$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 03,2023 at 11:05:16 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Power Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPEL.2023.3319589

The output voltages of Modules 1 and 2 are shown in Fig. 3(a) and Fig. 3(b), respectively, using the duty-cycle variation shown in Fig. 2. Fig. 3(c) shows the desired sinusoidal output voltage, v_o, which is obtained by taking the difference of the output voltages of the two dc/dc converter modules. For the CMS-FDO, the fixed dc bias of 0.5 in the duty cycles results in an offset of V_{in} on the output of the dc/dc converter modules regardless of the normalized gain g.

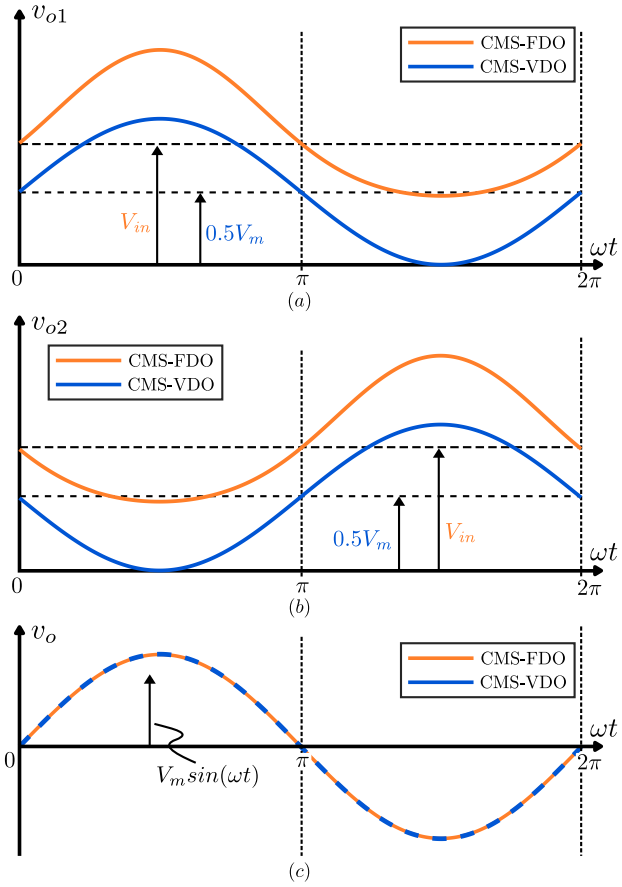


Fig. 3. Output voltages of (a) Module 1, (b) Module 2, and (c) the DMCI over a line cycle operated using CMD-FDO and CMS-VDO.

In CMS-FDO, the module dc-offset voltage is not varied instead, it is set at V_{in}, which generally yields higher module output voltage. The reduction in the dc-offset voltage at the output of each module leads to a reduction in the circulating power and switch voltage stress as detailed in Section III.

B. CMS-VDO

In the CMS-VDO, d_1 and d_2 are so determined that the inequality $V_{in} - 0.5V_m \geq 0$. In addition, because the output voltage of the inverter is determined by the differences in the output voltages of the two dc/dc converter modules, the minimum output voltages of the modules also contribute more to building the desired inverter output voltage. The CMS-VDO has been proposed considering these aspects. In a CMS-VDO-based DMCI, each dc/dc converter module produces a sinusoidal output voltage with dc-offset volage $= 0.5gV_{in}$ yielding the following:

$$v_{o1} = (0.5g + 0.5g\sin(\omega t))V_{in} \quad (6a)$$

$$v_{o2} = (0.5g - 0.5g\sin(\omega t))V_{in.} \quad (6b)$$

It is evident that $0.5V_m$ is not fixed but a function of the peak output voltage. For instance, when the voltage of the second module is at its minimum, the first module needs to generate less voltage (compared to that needed for CMS-FDO) to have the same inverter output voltage. Duty cycles d_1 and d_2 are obtained using (1) and (6) and are given by the following equations:

$$d_1 = \frac{0.5g + 0.5g\sin(\omega t)}{1 + 0.5g + 0.5g\sin(\omega t)} \quad (7a)$$

$$d_2 = \frac{0.5g - 0.5g\sin(\omega t)}{1 + 0.5g - 0.5g\sin(\omega t)}. \quad (7b)$$

Fig. 2 shows the duty cycles for CMS-VDO (along with that for CMS-FDO) of the two dc/dc converter modules for the same inverter output voltage. Fig. 3 shows the module voltages and inverter output voltage for the CMS-VDO along with the CMS-FDO obtained using the duty cycles shown in Fig. 2.

Fig. 4(a) and Fig. 4(b) show the variation in the desired DMCI output voltage (i.e., $V_m\sin(\omega t)$) normalized to the input voltage (i.e., V_{in}) with varying duty cycle d_1 for CMS-FDO and CMS-VDO $g = 1.2$. Fig. 4(a) shows that, for CMS-FDO, even though the normalized dc/dc converter module output voltages are nonlinear functions of the duty cycle, the inverter normalized output voltage appears almost a linear function of the duty cycle. Fig. 4(b) shows that, for CMS-VDO, the inverter and module normalized output voltages are nonlinearly related to the duty cycle. As the minimum gain of the individual module is zero, the gain trace of the modules

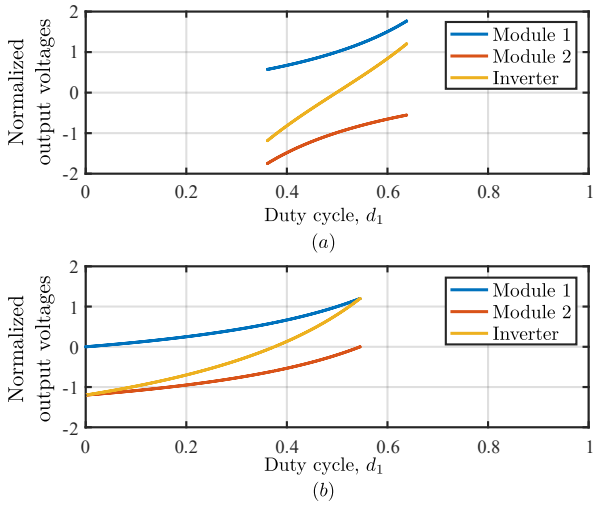


Fig. 4. Inverter and module output voltages (normalized to the input voltage) with varying duty cycle for (a) CMS-FDO and (b) CMS-VDO.

shifts with varying g. Finally, for the same, normalized inverter output voltage, CMS-FDO normalized module output voltages are found to be higher indicating a higher switch voltage stress.

III. Comparison of CMS-VDO and CMS-FDO

In this section, a comparative analysis of CMS-VDO and CMS-FDO is provided to outline the advantage of CMS-VDO. The following analysis assumes an ideal DMCI where the switches and the passive components are assumed to be ideal.

A. Switch Voltage Stress

The peak voltage across the switches in a Ćuk derived converter is dictated by the peak voltage across the blocking

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 03,2023 at 11:05:16 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Power Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPEL.2023.3319589

capacitor (e.g., C_1 in Module 1). The voltage across the blocking capacitor is the sum of the instantaneous input and output voltages of the Ćuk module. Hence, the peak voltage stress across the switches for Module 1 is given by

$$V_{peak} = v_{o1,peak} + V_{in}. \tag{8}$$

Using (5), (6), and (8), peak voltage stress for CMS-FDO and CMS-VDO are given by (9) and (10), respectively:

$$V_{peak,CMS-FDO} = V_{in}\left(1 + \frac{g}{2} + \sqrt{1 + \left(\frac{g}{2}\right)^2}\right) \tag{9}$$

$$V_{peak,CMS-VDO} = V_{in}(1 + g). \tag{10}$$

Fig. 5 illustrates the peak switch voltage for the two modulation schemes for varying g. It indicates that, the peak switch voltage stress increases with an increase in the magnitude of the output voltage of the DMCI for both CMS-FDO and CMS-VDO. However, the peak switch voltage stress is significantly higher for CMS-FDO compared to CMS-VDO, especially for lower g.

B. Circulating Power

As the two dc/dc converter modules of the DMCI operate in an identical manner except for the out-of-phase operation, analysis of only the first module is provided. Due to a cascaded

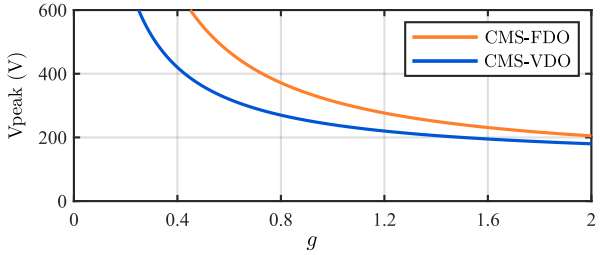


Fig. 5. Peak switch voltage as a function of g for V_m = 120 V.

connection of the modules at the output side, the load current is also the output current of each of the two modules. Hence, the instantaneous output power of the DMCI module, $p_{o1}(t)$, is given by

$$p_{o1} = v_{o1} i_o \tag{11}$$

where $i_o(t) = i_{o1}(t) = -i_{o2}(t)$ ignoring the switching ripple. Using (5), (6), and (11), the instantaneous output power for CMS-FDO and CMS-VDO for a unity-power-factor load are determined to be the following:

$$p_{o1,CMS-FDO} = \frac{V_{in}^2}{R} g_s \left(\frac{g_s}{2} + \sqrt{1 + \left(\frac{g_s}{2}\right)^2}\right) \tag{12}$$

$$p_{o1,CMS-VDO} = \frac{V_{in}^2}{R} g_s \frac{(g + g_s)}{2} \tag{13}$$

where $g_s = g\sin(\omega t)$ and R is the resistive load.

Using (12) and (13), and for the same input and output voltages, the instantaneous power of the DMCI module operating with CMS-FDO and CMS-VDO are plotted in Fig. 6 for $g = 1.2$. Fig. 6 implies that, even though the output power of the inverter is the same (given the same output voltage and resistive load) for both the modulation schemes, the individual module of the inverter handles more power when it is operated using CMS-FDO rather than CMS-VDO. This indicates the presence of higher circulating power for CMS-FDO-based DMCI.

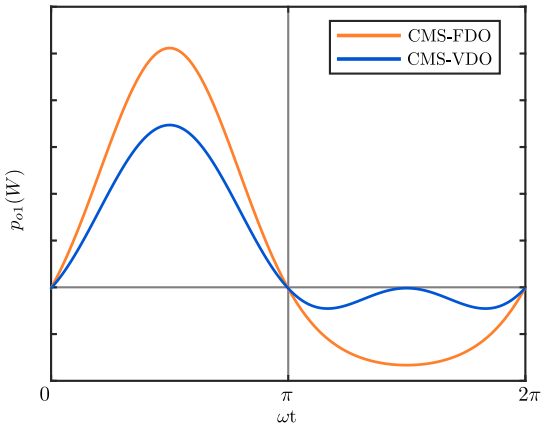


Fig. 6. Instantaneous output power of a DMCI module over a line cycle for CMS-FDO and CMS-VDO.

The circulating power reflects the additional power, beyond the power that needs to be delivered to the load, which circulates in a dc/dc converter module of the DMCI [2]. As such, such a circulating power is not desired in a DMCI since it results in additional dc/dc converter module (and hence inverter) losses. Additionally, a lower circulating power also yields reduced current and/or voltage stress for the same power delivered to the load.

In this paper, the ratio of the circulating (reactive) power to the active power for the DMCI module is used to assess the extent of the circulating power. The reactive power (Q_{o1}) through a DMCI module can be derived from apparent power, S_{o1}, and active power, P_{o1}, over a line cycle period, T, as captured by the following relationship:

$$S_{o1}^2 = P_{o1}^2 + Q_{o1}^2 = V_{o1}^2 I_o^2 \tag{14}$$

where

$$V_{o1}^2 = \frac{1}{T}\int_0^T v_{o1}^2 dt$$

$$I_o^2 = \int_0^T i_{o1}^2 dt$$

V_{o1} and I_o represent the root-mean-square (RMS) values of v_{o1} and i_{o1}, respectively. The ratio of the reactive power to the active power is given by the following:

$$\frac{Q_{o1}}{P_{o1}} = \sqrt{\frac{\frac{1}{T}\int_0^T v_{o1}^2 dt \frac{1}{T}\int_0^T i_o^2 dt}{\left[\frac{1}{T}\int_0^T v_{o1} i_o dt\right]^2} - 1}. \tag{15}$$

Using (5), (6), and (12), the ratio of Q_{o1}/P_{o1} for the CMS-FDO and CMS-VDO operated DMCI is determined to be

$$\left(\frac{Q_{o1}}{P_{o1}}\right)_{CMS-FDO} = \sqrt{\frac{g^2 + 8}{g^2}} \tag{16a}$$

$$\left(\frac{Q_{o1}}{P_{o1}}\right)_{CMS-VDO} = \sqrt{\frac{\frac{1}{T}\int_0^T \left(0.5V_m(1 + \sin(\omega t))\right)^2 dt \frac{1}{T}\int_0^T \left(\frac{V_m}{R}\sin(\omega t)\right)^2 dt}{\left[\frac{1}{T}\int_0^T \frac{V_m^2}{2R}(1 + \sin(\omega t))\sin(\omega t)\, dt\right]^2}} - 1$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 03,2023 at 11:05:16 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Power Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPEL.2023.3319589

$$\left(\frac{Q_{o1}}{P_{o1}}\right)_{CMS-VDO} = \sqrt{2} \quad (16b)$$

which indicates that unlike the CMS-FDO, the ratio of reactive to active power in CMS-VDO is independent of g.

Fig. 7 shows the comparison of the ratio of the reactive (circulating) power to the active power for varying g. It shows that, the reactive (circulating) power for the CMS-FDO-based DMCI is significantly higher than that obtained using CMS-VDO-based DMCI at lower g. With increasing g, this difference between the two modulation schemes narrows.

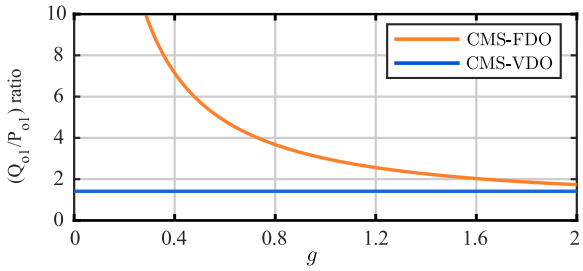


Fig. 7. Ratio of reactive (circulating) to active power for varying peak-normalized dc-voltage gain (g).

C. Current stress

The current through the switches in a Ćuk derived converter is sum of the instantaneous input and output current of the Ćuk module. As the modules are assumed to be identical, the current stress on both the modules are equal, hence, only Module 1 is used to analyze the current stress. For an ideal Ćuk module, the peak current through the switches over a switching period is given by

$$I_{sw} = \langle I_{sw} \rangle + \Delta I_{sw}. \quad (17a)$$

In (17a), $\langle I_{sw} \rangle$ is the average current through the switches over a switching period and is given by the following equation:

$$\langle I_{sw} \rangle = \left({}^{v_{o1}}/_{V_{in}} + 1\right) i_{o1} \quad (17b)$$

and ΔI_{sw} is the ripple current through the switches. By using volt-second balance on the input and output inductor currents, ΔI_{sw} is given by the following:

$$\Delta I_{sw} = \frac{V_{in} d_1}{2 f_{sw}} \left(\frac{1}{L_1} + \frac{1}{L_2}\right). \quad (17c)$$

where L_1 and L_2 is the input and output inductors, respectively. From (17c), ΔI_{sw} for Ćuk module is proportional to the duty cycle of the module which is shown in Fig. 2 for a constant V_{in}. Using (5)-(7), the average current through the switches for the CMS-FDO and CMS-VDO operated DMCI is determined to be the following:

$$\langle I_{sw} \rangle_{CMS-FDO} = \left(\frac{g_s}{2} + \sqrt{1 + \left(\frac{g_s}{2}\right)^2} + 1\right)\left(\frac{g_s V_{in}}{R}\right) \quad (18)$$

$$\langle I_{sw} \rangle_{CMS-VDO} = (0.5g + 0.5g_s + 1)\left(\frac{g_s V_{in}}{R}\right) \quad (19)$$

The peak value of ripple current through the switches for the CMS-FDO and CMS-VDO operated DMCI occurs at the peak of output voltage and is obtained by substituting (4) and (7) into (17c). The peak value of the current ripple through the switch is given by

$$\Delta I_{sw,pk,CMS-FDO} = \frac{V_{in}}{2 f_{sw}} \left(0.5 + \frac{\sqrt{4 + g^2} - 2}{2g}\right)\left(\frac{1}{L_1} + \frac{1}{L_2}\right) \quad (20)$$

$$\Delta I_{sw,pk,CMS-VDO} = \frac{g V_{in}}{2 f_{sw}} \left(\frac{1}{1+g}\right)\left(\frac{1}{L_1} + \frac{1}{L_2}\right) \quad (21)$$

Fig. 8(a) shows the comparison of the peak value of the average current through the switches for CMS-FDO and CMS-VDO operated DMCI using (18) and (19). As shown in Fig. 8(a), the peak value of the average current through the switches is found to be higher using CMS-FDO compared to CMS-VDO, which results in a higher condution loss.

In Fig. 8(b), the peak value of ripple current through the switches is shown for CMS-FDO and CMS-VDO operated DMCI. The peak value of ripple across the swtiches is higher for CMS-FDO compared to CMS-VDO. The difference is significantly higher for lower g due to higher duty cycle (higher module voltage) in case of CMS-FDO compared to CMS-VDO. The high ripple at lower g results in higher conduction losses in the switch.

As the switch current is the sum of the input and output inductor currents, the higher switch current also indicates higher current through the inductors which increases magnetic conduction losses. The ripple current through an inductor produces varying magnetic flux field in the magnetic core. Higher ripple current in the case of CMS-FDO compared to CMS-VDO causes higher nonlinear magnetic core losses [12], [23] especially for lower g.

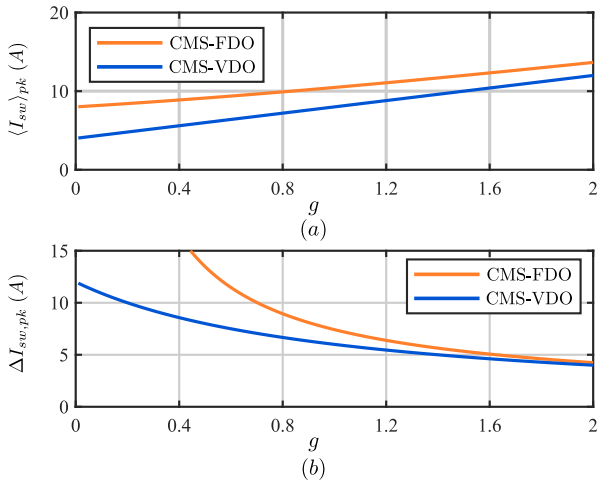


Fig. 8. (a) Peak value of average current, and (b) peak value of ripple current through the switches for varying peak-normalized dc-voltage gain (g) with V_m = 120 V and R = 30 Ω.

IV. Experimental Results

An experimental prototype of the DMCI is developed to validate the analysis regarding CMS-VDO and provide comparison with CMS-FDO. Fig. 9 shows a hardware prototype of the DMCI with the specifications provided in Table I. A TMS320F28335 digital signal processor (DSP) based controller manufactured by Texas Instruments is used to generate the desired PWM signals for the CMS-VDO- and CMS-FDO-based DMCI. A PA3000 power analyzer manufactured by Tektronix is used to measure the efficiency, and the output-voltage total harmonic distortion (THD) of the inverter. The same hardware prototype is used for both CMS-FDO and CMS-VDO-operated DMCI for experimental comparison and validation.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 03,2023 at 11:05:16 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Power Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPEL.2023.3319589

Fig. 9. Experimental prototype of the DMCI.

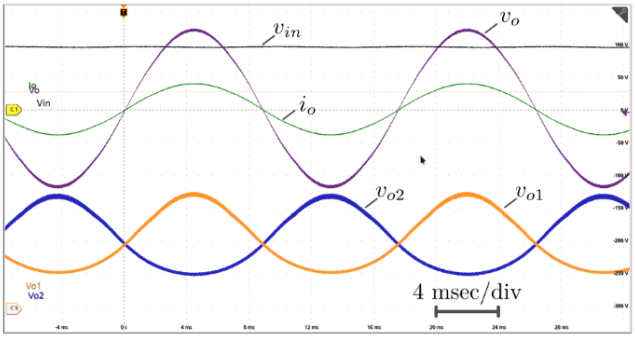
The inverter waveform for a unity-power-factor load of 30 Ω and $g = 1.2$ is shown in Fig. 10. Fig. 10(a) shows the waveform for CMS-FDO-based DMCI. The modulation scheme is realized using the duty cycle relation given by (4) for the desired output voltage. As outlined earlier and illustrated in Fig. 3, the output voltages of the dc/dc converter modules of the DMCI have a dc offset equalling the input voltage. The difference of the output voltages of the modules results in the sinusoidal output voltage of the DMCI.

TABLE I. PARAMETERS OF THE EXPERIMENTAL DMCI PROTOTYPE

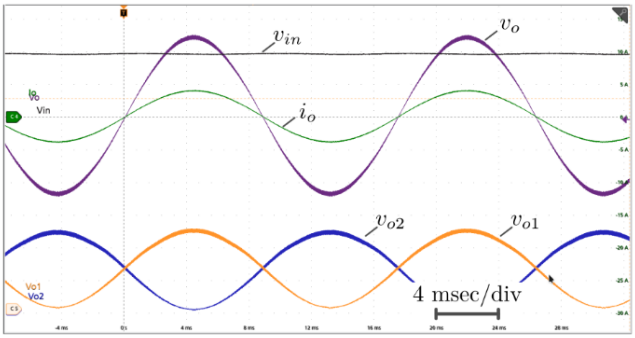
Input voltage, V_{in}	70 – 150 Vdc
Gain, g	0.8 – 1.7
Output frequency, f	50-60 Hz
Full load resistance, R	30 Ω
DC side filter inductance L_1, L_3	145 µH
AC side filter inductance L_2, L_4	161 µH
Blocking capacitance C_1, C_3	1.5 µF
AC side capacitance C_2, C_4	3.3 µF
Switching frequency, f_{sw}	50 kHz

Fig. 10(b), shows waveforms for the CMS-VDO modulated DMCI; the modulation scheme is realized following (7). The minimum value of the output voltage of a DMCI module operated with CMS-VDO is 0 V, which results in a lower circulating current. The peak output voltage of a DMCI module operated with CMS-FDO is found to be 179 V while the peak output voltage of a DMCI module operated with CMS-VDO is found to be 122 V for the same inverter RMS output voltage (V_o) of 85 V. The higher module output voltage for the CMS-FDO-based DMCI results in a higher switch-voltage stress.

In Fig. 11, the waveforms for the DMCI operated with a load of power factor 0.9 are shown to demonstrate the abilities of the CMS-VDO and CMS-FDO to support active as well as reactive power. Equation (4) and (7) show that for the ideal DMCI, operated, respectively, using CMS-FDO and CMS-VDO, the duty cycles are not dependent on the load current or the load power factor. As such, (4) and (7) are also used to operate the DMCI for the non-unity-power-factor load with a resistance of 30 Ω and an inductance of 40 mH. The results for both the modulation schemes are obtained for $g = 1.2$ and the same RMS output voltage of 85 V.

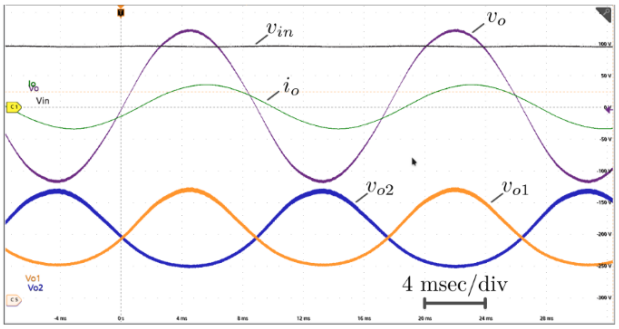


(a)

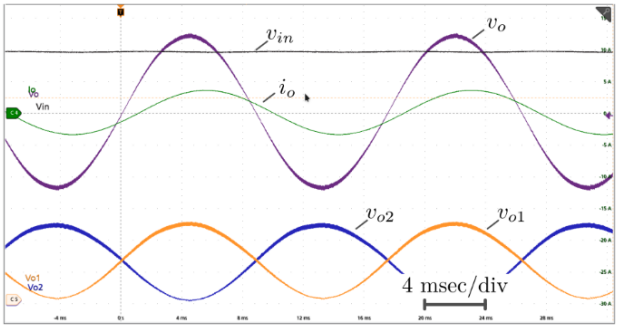


(b)

Fig. 10. Experimental waveforms of the DMCI with (a) CMS-FDO, and (b) CMS-VDO at $R = 30$ Ω and $g = 1.2$. The output voltage of the DMCI is the purple trace (50 V/div), the output current is the green trace (5 A/div), the output voltage of Module 1 is the orange trace (50 V/div), and the output voltage of Module 2 is the blue trace (100 V/div). The time scale is 4 msec/div.



(a)



(b)

Fig. 11. Experimental waveforms of the DMCI with (a) CMS-FDO, and (b) CMS-VDO at $R = 30$ Ω, $L = 40$ mH, and $g = 1.2$. The output voltage of DMCI is the purple trace (50 V/div), the output current is the green trace (5 A/div), the output voltage of Module 1 is the orange trace (50 V/div), and the output voltage of Module 2 is the blue trace (100 V/div). The time scale is 4 msec/div.

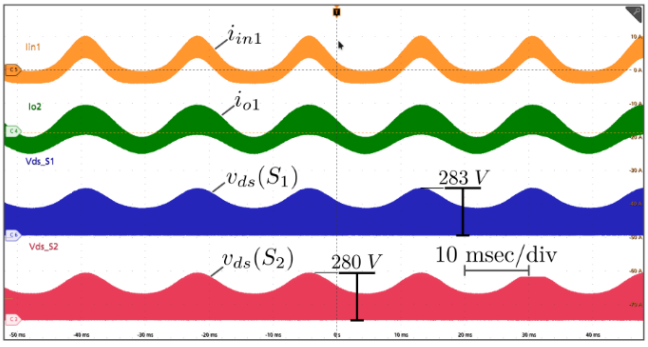
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 03,2023 at 11:05:16 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Power Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPEL.2023.3319589

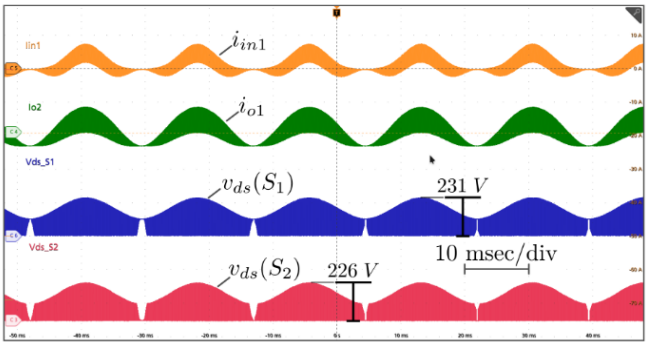
Fig. 12 shows the waveforms of Module 1 while the DMCI is operated with CMS-FDO and CMS-VDO and feeds a load of 30 Ω with $g = 1.2$. The figure shows that, the peak of the input inductor current for the CMS-FDO-based DMCI is higher than that obtained using the CMS-VDO-based DMCI. Since the input voltage for both the modulation schemes are kept the same, a higher input current for the CMS-FDO-based DMCI implies that the inverter for this case draws a higher circulating (reactive) power.

Further, the RMS input (output) inductor current for the CMS-FDO- and CMS-VDO-operated DMCI, following Fig. 12, are found to be 3.5 A (3.5 A) and 2.2 A (3.2 A), respectively. The higher current ripples of the inductors for the CMS-FDO-based DMCI results in higher copper and core magnetic losses.

Fig. 12 also shows that the peak switch voltages for a CMS-FDO-based DMCI module are about 21% higher compared to that obtained using CMS-VDO-based DMCI. This is because the switch voltage, which is nominally dictated by the blocking capacitor voltage (that in turn, is the sum of the input and output voltages of a module), is higher for the CMS-FDO-based DMCI given its higher output voltage. The higher voltages across the switches coupled with higher current ripples (alluded to in the earlier paragraph) also yield higher losses for the switches of the CMS-FDO-based DMCI.



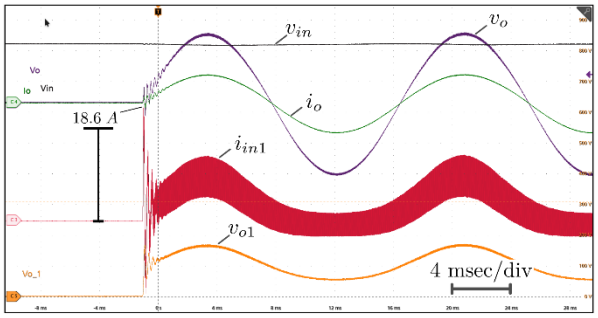
(a)



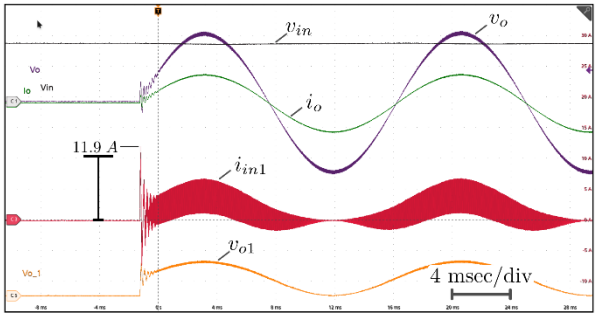
(b)

Fig. 12. Input and output currents and switch voltages for one of the modules (Module 1) of the DMCI, which is operated using (a) CMS-FDO and (b) CMS-VDO. The load resistance (R) is set at 30 Ω while g is set at 1.2. The orange trace is the input inductor current (10 A/div), the green trace is the output inductor current (10 A/div), the blue trace is the drain-to-source voltage of S_1 (200 V/div), and the red trace is the drain-to-source voltage of S_2 (200 V/div). The time resolution for the horizontal axis is set at 10 msec/div.

In Fig. 13, dynamic response of the DMCI operated with CMS-FDO and CMS-VDO is shown by starting the inverter abruptly at rated voltage. In case of CMS-FDO operated DMCI as shown in Fig. 13(a), the output voltage of the modules jumps from 0 V to the bias voltage, V_{in} as given by (5) for $\omega t = 0$. The startup results in LC excitation between the input inductor and capacitor producing peak input current of 18.6 A. In Fig. 13(b), dynamic response of DMCI operated with CMS-VDO is shown. As the required module output voltages given by (6) is $0.5V_m$ for $\omega t = 0$ which is lower for CMS-VDO operated DMCI compared with CMS-FDO operated DMCI, the LC excitation is lower resulting in peak inductor current is 11.9 A.



(a)



(b)

Fig. 13. Transient response of the DMCI using (a) CMS-FDO and (b) CMS-VDO. The load resistance (R) is set at 60 Ω while g is set at 1.2. The input voltage of DMCI is the black trace (50 V/div), the output voltage of DMCI is the purple trace (50 V/div), the green trace is the output current (2 A/div), the red trace is the input inductor current (5 A/div), and the output voltage of Module 1 is the orange trace (100 V/div). The time resolution for the horizontal axis is set at 4 msec/div.

Fig. 14 shows the efficiency of the DMCI operated with CMS-FDO and CMS-VDO for various test conditions. Fig. 14(a) provides the results of the DMCI for $g = 1.2$ with varying load. At full load ($R = 30\ \Omega$), the efficiency of the DMCI, operated with CMS-VDO, is found to be 94.77% compared to 91.50% obtained using CMS-FDO. The higher efficiency of the CMS-VDO-based DMCI is due to its lower circulating (reactive) power that yields in reduced conduction, switching, and core losses. For both of the modulation schemes, the diminished efficiency of the DMCI at reduced load is attributed to the constant voltage dependent losses. However, the difference in the efficiencies of the DMCI operated with CMS-FDO and CMS-VDO amplifies further for lower load. Higher drop in the efficiency of the DMCI operated with CMS-FDO are attributed to the higher current ripple of the input and output inductors resulting in a higher magnetic core losses. At low load ($R = 180\ \Omega$), the measured efficiency is 85.03% and 76.11% for DMCI operated with CMS-VDO and CMS-FDO, respectively.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 03,2023 at 11:05:16 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Power Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPEL.2023.3319589

Inverters are typically expected to maintain a constant output voltage with variable input voltage [3], [5], [24]. Hence, the performances of the two modulation schemes for the DMCI are also compared with variable normalized gain (input voltage) for a set output voltage, as captured in Fig. 14(b). The difference in efficiencies of the DMCI operated with CMS-FDO and CMS-VDO at high g (low input voltage) is found to be small due to similar reactive power generation and current stress yielded by the two modulation schemes as discussed in Section III and shown in Fig. 7. In contrast, the efficiencies of the DMCI operated with CMS-FDO and CMS-VDO is found to be higher at low g (high input voltage), as shown in Fig. 14(b).

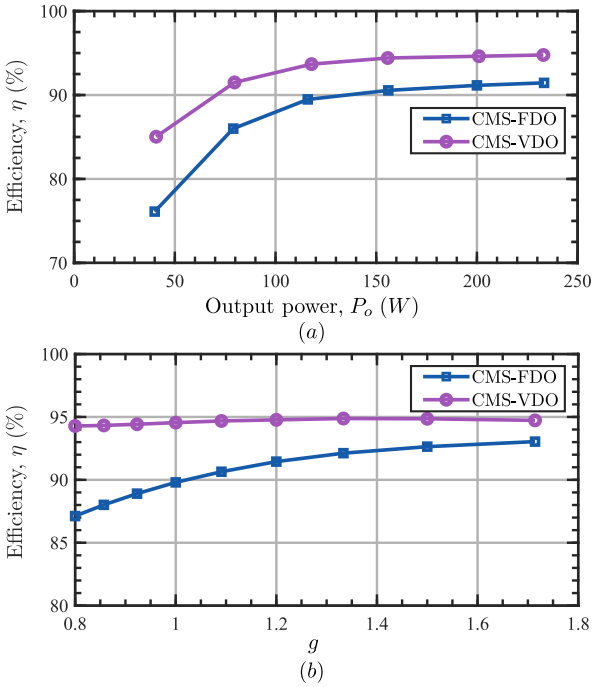


Fig. 14. Comparison of efficiencies of CMS-FDO, and CMS-VDO-based DMCI operated with constant $V_o = 85$ VRMS for varying (a) output power, and (b) normalized gain.

Fig. 15 provides the estimated loss distribution of the DMCI operated with CMS-FDO and CMS-VDO for various normalized gain which follows the experimental data shown in Fig. 14(b). In case of the CMS-VDO, the total switch loss increases slightly with g due to increase in overall RMS current through the switches. In contrast, the total switch loss for CMS-FDO decreases with g due to significant drop in ripple current as given by (20) resulting in drop in the RMS current through the switches. The magnetic winding and core losses [12] which have high dependency on the ripple current decreases with g for both CMS-FDO and CMS-VDO operated DMCI.

In Fig. 15, the winding loss is obtained by considering the skin and proximity effect on the wire while the magnetic core loss is estimated using the Steinmetz's equation. The aggregate loss of the DMCI operated with CMS-VDO remains nearly the same even with wide g variation. For the CMS-FDO-based DMCI, the aggregate loss reduces significantly with g (higher input voltage) due to progressively lower reactive-power generation and ripple current.

In Fig. 16, the output-voltage THDs of the DMCI operated with CMS-FDO and CMS-VDO are shown. The THD of the DMCI operated with CMS-FDO and CMS-VDO for wide load variation is shown in Fig. 16(a). It is found that the THD is lower than 2% for DMCI operated with CMS-FDO while the THD is lower than 1% for DMCI operated with CMS-VDO. While the trend of THD is similar in both the modulation schemes for varying load, the offset can be accounted by higher ripple current which results in non-linear losses [12]. The THD for varying g is shown in Fig. 16(b) for both the modulation schemes. The THD trend is similar to the ripple current dependency shown in Fig. 8(b).

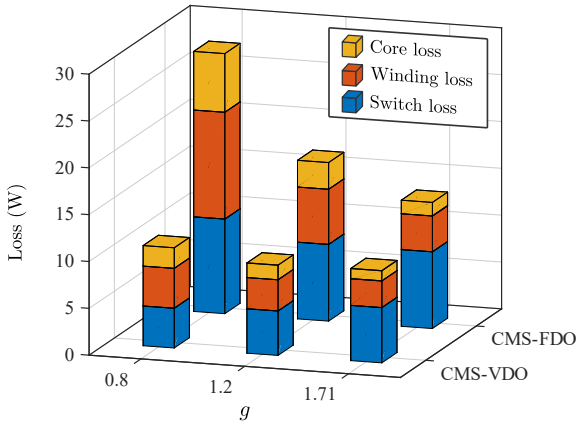


Fig. 15. Comparison of estimated loss CMS-FDO, and CMS-VDO-based DMCI operated with constant $V_o = 85$ VRMS and $R = 30\ \Omega$.

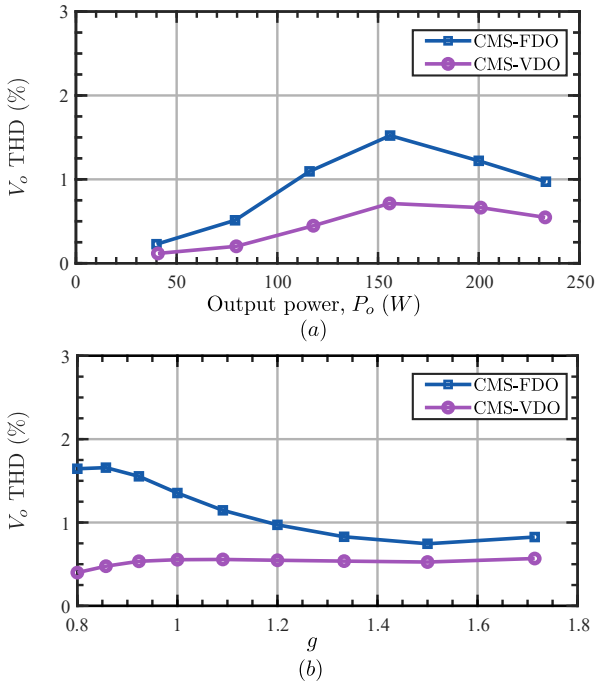


Fig. 16. Comparison of output voltage THD of CMS-FDO, and CMS-VDO based DMCI operated with a constant $V_o = 85$ V RMS against (a) varying output power for $V_{in} = 100$ V and (b) varying normalized gain with a load of $R = 30\ \Omega$.

The output-voltage THD of the DMCI operated with CMS-VDO is found to be lower than that obtained with CMS-FDO as corroborated with Fig. 17. The THD originates in the DMCI due to the nonlinearity and parasitic elements of the inverter

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 03,2023 at 11:05:16 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Power Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPEL.2023.3319589

resulting in deviation from the ideal relation developed for the module output voltages captured by (5) and (6) for the DMCI operated with CMS-FDO and CMS-VDO, respectively. The major contributors to the output-voltage THD are the 3rd, 5th, and 7th harmonics for both the modulation schemes. In Fig. 17, the odd-harmonic content compared with CMS-VDO modulated DMCI is nearly half of CMS-FDO modulated DMCI resulting in an overall better THD.

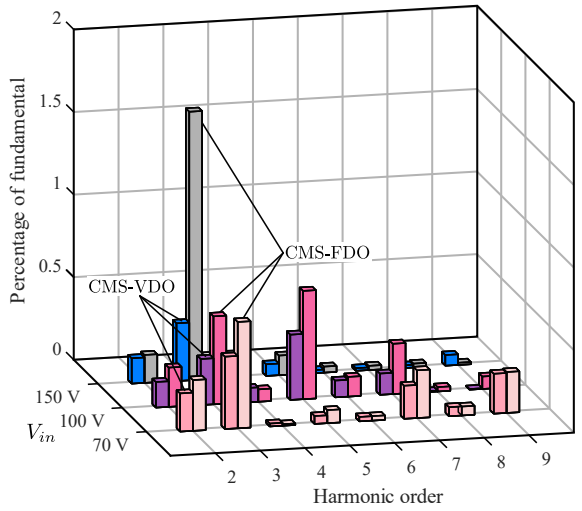


Fig. 17. Harmonic analyses of the output voltage of the CMS-VDO- and CMS-FDO-based DMCI relative to the fundamental harmonic: $V_{in} = 70$ V, 100 V and 150 V, $V_o = 85$ V RMS ($g = 1.2$), and $R = 30\ \Omega$.

V. Conclusion

This paper provides a parametric and experimental comparison between CMS-VDO- and CMS-FDO-based single-phase DMCI. Unlike the CMS-FDO, in CMS-VDO, the dc-offset voltage is varied as a function of the inverter peak output voltage for duty-cycle generation of each of the two dc/dc converter modules of the DMCI. Compared to the CMS-FDO-based DMCI, CMS-VDO-based DMCI yields a lower module output voltage, which results in lower voltage stress on the switches and lower input and output inductor-current ripples. CMS-VDO also reduces the reactive (circulating) power flowing between the two modules, which enhances the inverter efficiency over a wide operating range. The analyses described in this paper are ratified with experimental validations. For the DMCI hardware prototype, overall, the CMS-VDO-based DMCI yields higher efficiency, lower switch voltage stress, and lower THD compared to the CMS-FDO-based DMCI.

VI. References

[1] S. D'Silva, M. Shadmand, S. Bayhan, and H. Abu-Rub, "Towards grid of microgrids: Seamless transition between grid-connected and islanded modes of operation," *IEEE Open J. Ind. Electron. Soc.*, vol. 1, no. 1, pp. 66–81, 2020, doi: 10.1109/OJIES.2020.2988618.

[2] S. Gupta, M. F. Umar, M. B. Shadmand, and S. K. Mazumder, "Self-Synchronization Scheme for Network of Grid-following and Grid-forming Photovoltaic Inverters," *IECON Proc. (Industrial Electron. Conf.*, vol. 2021-October, Oct. 2021, doi: 10.1109/IECON48115.2021.9589752.

[3] J. M. A. Myrzik and M. Calais, "String and Module Integrated Inverters for Single-Phase Grid Connected Photovoltaic Systems - A Review," *2003 IEEE Bol. PowerTech - Conf. Proc.*, vol. 2, pp. 8–15, 2003, doi: 10.1109/PTC.2003.1304589.

[4] G. Oriti, A. L. Julian, and N. J. Peck, "Power-Electronics-Based Energy Management System With Storage," *IEEE Trans. Power Electron.*, vol. 31, no. 1, pp. 452–460, Jan. 2016, doi: 10.1109/TPEL.2015.2407693.

[5] Y. M. Chen, Y. C. Liu, S. C. Hung, and C. S. Cheng, "Multi-Input Inverter for Grid-Connected Hybrid PV/Wind Power System," *IEEE Trans. Power Electron.*, vol. 22, no. 3, pp. 1070–1077, May 2007, doi: 10.1109/TPEL.2007.897117.

[6] Y. Tang, S. Xie, and C. Zhang, "Single-Phase Z-Source Inverter," *IEEE Trans. Power Electron.*, vol. 26, no. 12, pp. 3869–3873, 2011, doi: 10.1109/TPEL.2009.2039955.

[7] T. N. Gautham, R. Reddivari, and D. Jena, "A cost-effective single-phase semi flipped gamma type magnetically coupled impedance source inverters," *Int. J. Circuit Theory Appl.*, vol. 49, no. 4, pp. 1078–1102, 2021, doi: 10.1002/cta.2865.

[8] D. Cao, S. Jiang, X. Yu, and F. Z. Peng, "Low-cost semi-Z-source inverter for single-phase photovoltaic systems," *IEEE Trans. Power Electron.*, vol. 26, no. 12, pp. 3514–3523, 2011, doi: 10.1109/TPEL.2011.2148728.

[9] S. Ćuk and R. D. Middlebrook, "Advances in Switched-Mode Power Conversion Part I," *IEEE Trans. Ind. Electron.*, vol. IE-30, no. 1, pp. 10–19, 1983, doi: 10.1109/TIE.1983.356697.

[10] R. O. Cáceres and I. Barbi, "A Boost DC-AC Converter: Analysis, Design, and Experimentation," *IEEE Trans. Power Electron.*, vol. 14, no. 1, pp. 134–141, 1999, doi: 10.1109/63.737601.

[11] S. Gupta and S. K. Mazumder, "A Differential-Mode Isolated AC/AC Converter," *IEEE Trans. Power Electron.*, vol. 38, no. 2, pp. 12846–12858, 2023, doi: 10.1109/TPEL.2023.3292983.

[12] M. Mohamadi, S. K. Mazumder, and N. Kumar, "Integrated Magnetics Design for a Three-Phase Differential-Mode Rectifier," *IEEE Trans. Power Electron.*, vol. 36, no. 9, pp. 10561–10570, Sep. 2021, doi: 10.1109/TPEL.2021.3066506.

[13] A. A. Aboulnaga and A. Emadi, "Performance Evaluation of The Isolated Bidirectional Ćuk Converter with Integrated Magnetics," *PESC Rec. - IEEE Annu. Power Electron. Spec. Conf.*, vol. 2, pp. 1557–1562, 2004, doi: 10.1109/PESC.2004.1355657.

[14] S. Cuk, "A New Zero-Ripple Switching DC-to-DC Converter and Integrated Magnetics," *IEEE Trans. Magn.*, vol. 19, no. 2, pp. 57–75, 1983, doi: 10.1109/TMAG.1983.1062238.

[15] D. P. Fernandez, R. A. Fantino, J. C. Balda, and S. K. Mazumder, "Inductor and Transformer-Coupled Magnetic Structure for Zero-Ripple DC-DC Ćuk Converter," *Conf. Proc. - IEEE Appl. Power Electron. Conf. Expo. - APEC*, pp. 654–659, Jun. 2021, doi: 10.1109/APEC42165.2021.9487405.

[16] S. Mehrnami and S. K. Mazumder, "Discontinuous Modulation Scheme for a Differential-Mode Ćuk Inverter," *IEEE Trans. Power Electron.*, vol. 30, no. 3, pp. 1242–1254, 2015, doi: 10.1109/TPEL.2014.2314688.

[17] S. Gupta and S. K. Mazumder, "A Novel Modulation Scheme for Isolated PWM Active-Clamp Ćuk DC/DC Converter," *IEEE Trans. Power Electron.*, 2022, doi: 10.1109/TPEL.2022.3189672.

[18] F. Barzegar and S. Cuk, "Solid-State Drives for Induction Motors: Early Technology to Current Research," in *IEEE Region 6 Conf.*, 1982, p. Feb. 15–18.

[19] B. Han, J. S. Lai, and M. Kim, "Bridgeless Cuk-Derived Single Power Conversion Inverter with Reactive-Power Capability," *IEEE Trans. Power Electron.*, vol. 35, no. 3, pp. 2629–2645, Mar. 2020, doi: 10.1109/TPEL.2019.2927645.

[20] C. Bao, S. Gupta, and S. K. Mazumder, "Modeling and Analysis of Peak-Current-Controlled Differential Mode Ćuk Inverter," *Proc. 2021 IEEE 12th Int. Symp. Power Electron. Distrib. Gener. Syst. PEDG 2021*, Jun. 2021, doi: 10.1109/PEDG51384.2021.9494167.

[21] K. Tank, M. M. Garg, and S. K. Mazumder, "A New Continuous Modulation Scheme for Differential Mode Boost Inverter," *2022 IEEE 13th Int. Symp. Power Electron. Distrib. Gener. Syst. PEDG 2022*, 2022, doi: 10.1109/PEDG54999.2022.9923272.

[22] R. W. Erickson and D. Maksimović, "Fundamentals of Power Electronics," *Fundam. Power Electron.*, 2001, doi: 10.1007/B100747.

[23] J. Mühlethaler, J. Biela, J. W. Kolar, and A. Ecklebe, "Core losses under DC bias condition based on Steinmetz parameters," *2010 Int. Power Electron. Conf. - ECCE Asia -, IPEC 2010*, pp. 2430–2437, 2010, doi: 10.1109/IPEC.2010.5542385.

[24] M. G. Varzaneh, A. Rajaei, M. Forouzesh, Y. P. Siwakoti, and F. Blaabjerg, "A Single-Stage Multi-Port Buck-Boost Inverter," *IEEE*

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 03,2023 at 11:05:16 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Power Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPEL.2023.3319589

Trans. Power Electron., vol. 36, no. 7, pp. 7769–7782, Jul. 2021, doi: 10.1109/TPEL.2020.3042338.

Kartik Tank his B.Tech. Degree in Electrical Engineering from Rajasthan Technical University in year 2020 and M.Tech. Degree in Power Electronics and Drives specialization at Department of Electrical Engineering, Malaviya National Institute of Technology (MNIT) Jaipur, India in year 2022. His current research interest includes modelling and control of dc-dc converters, differential mode inverters, electric vehicle chargers, etc.

Shantanu Gupta (Graduate Student Member, IEEE) received the B.Tech. degree in electrical engineering from Delhi Technological University, Delhi, India, in 2016. He is currently working toward the Ph.D. degree in electrical and computer engineering at the University of Illinois Chicago, Chicago, IL, USA.

From 2016 to 2019, he was a Research Engineer with TEX E.G., Japan. His work involved working on embedded and power electronics systems. Among the projects he worked on were the development of a semiconductor wafer alignment machine and a motor driver for a refrigerator compressor. His research interests include soft-switching techniques and power electronics topologies. Mr. Gupta serves as a reviewer for IEEE TRANSACTIONS ON POWER ELECTRONICS and IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS.

Man Mohan Garg is currently serving in the Department of Electrical Engineering, Malaviya National Institute of Technology (MNIT) Jaipur, India as an Assistant Professor. Prior to joining MNIT Jaipur in June 2020, he has worked as an Assistant Professor in the Department of Electrical Engineering, National Institute of Technology (NIT) Rourkela, India. He has received the B.E. degree in Electrical Engineering from the M.B.M. University Jodhpur, India; M.Tech. and Ph.D. degrees in Electrical Engineering from Indian Institute of Technology (IIT) Roorkee, India. He completed his post-doctoral with the Department of Electrical Engineering, IIT Kanpur, India. His current research interest includes design, modeling, and control of power electronic converters, grid integration of distributed renewable energy sources, dc microgrid, electric vehicles, cyber-physical energy systems, LED driver circuits, etc. His research work has been published in the IEEE Transactions/Journals and other reputed international journals, book chapters and conferences. He is an active editorial board member in several reputed journals and conferences. He is Senior Member, IEEE.

Sudip K. Mazumder (S'97–M'01–SM'03–F'16) received the Ph.D. degree in electrical and computer engineering from Virginia Tech, in 2001. He is the President of NextWatt LLC since 2008 and a professor at UIC since 2001. In June 2023, he was chosen as a recipient of the honorary title of UIC Distinguished Professor to be effective in AY'23 pending Board of Trustees approval. He has over 30 years of professional experience and has held research and development and design positions in leading industrial organizations and has served as a technical consultant for several industries.

He was named a Fellow of the American Association for the Advancement of Science (AAAS), in 2020, a Fellow of the Institute of Electrical and Electronics Engineers (IEEE), in 2016, and a Fellow of the Asia-Pacific Artificial Intelligence Association (AAIA) in 2022. He serves as the Editor in Large for the IEEE Transactions on Power Electronics (TPEL) since 2019 and served as an IEEE Distinguished Lecturer between 2016-2019. He has been serving as an Administrative Committee Member for IEEE PELS, since 2015. He has also been serving as a Member-at-Large for IEEE PELS, since 2020. He has served as the Chair for the IEEE PELS Technical Committee on Sustainable Energy Systems, from 2015 to 2020. He serves as the General Chair for IEEE PEDG Conference in 2023.

He is the recipient of the 2023 IEEE Power & Energy Society's Ramakumar Family Renewable Energy Excellence Award. He received several IEEE awards/honors, including IEEE Transactions on Power Electronics Prize Paper Awards (2022, 2002) and Highlighted Papers (2023, 2022, 2018), Featured Article for IEEE Transactions on Biomedical Engineering (2023), IEEE Conference Best Paper Award (2013), and IEEE International Future Energy Challenge Award (2005).

Moien Mohamadi (S'14) received Ph.D. degree in 2022, from University of Illinois at Chicago, Chicago, IL, USA, where he worked at Laboratory for Energy and Switching-Electronics Systems (LESES) to improve efficiency and power density of off-board chargers. He is currently working at Tesla.

He has been working in power electronics research and development since 2014 and has been a member of IEEE and PES community. His research interests are optimal design and hardware realization of highly efficient and compact power electronics converter, high-frequency magnetic design, optimization, and hardware realization, and soft-switching power electronics.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 03,2023 at 11:05:16 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

Review of Scientific Instruments

RESEARCH ARTICLE | OCTOBER 09 2023

A versatile 4 K insert for characterization of the superconducting joints

Ajit Nandawadekar ; Mukhtiar Singh; Soumen Kar

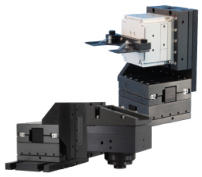
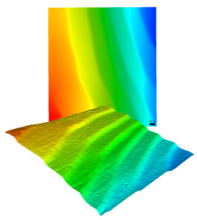
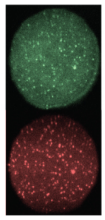
Check for updates

Rev. Sci. Instrum. 94, 105108 (2023)
https://doi.org/10.1063/5.0161148

View Online
Export Citation
CrossMark

03 November 2023 05:51:45

MCL
MAD CITY LABS INC.
www.madcitylabs.com
Nanopositioning Systems

Modular Motion Control

AFM and NSOM Instruments

Single Molecule Microscopes


AIP Publishing

Review of Scientific Instruments ARTICLE pubs.aip.org/aip/rsi

A versatile 4 K insert for characterization of the superconducting joints

Cite as: Rev. Sci. Instrum. 94, 105108 (2023); doi: 10.1063/5.0161148
Submitted: 8 June 2023 • Accepted: 20 September 2023 •
Published Online: 9 October 2023

View Online Export Citation CrossMark

Ajit Nandawadekar,[1,2] Mukhtiar Singh,[2] and Soumen Kar[1,a)]

AFFILIATIONS

[1] Inter-University Accelerator Centre, New Delhi 110067, India
[2] Department of Electrical Engineering, Delhi Technological University, Shahbad Daulatpur, Delhi 110042, India

[a)] **Author to whom correspondence should be addressed:** kar.soumen@gmail.com

ABSTRACT

A multi-coil superconducting magnet is the heart of any modern Magnetic Resonance Imaging (MRI) scanner. The superconducting magnet in MRI produces a highly homogeneous field (±5 ppm) and stable field (<0.1 ppm/h) in its imaging volume. The temporal field stability depends on the electrical resistance of the inter-coil joints and the electrical resistance of the joint between the coil and the superconducting switch. To establish the joint technology, a 4 K test rig is required for testing the sample joints. This paper briefly discusses the details of the versatile 4 K insert recently developed at IUAC for the characterization of the superconducting joints using the field-decay measurement technique. It can measure the joint resistance up to ~10^{-15} Ω in a straight or spiral joint configuration with lengths up to 150 mm. The maximum current that can be induced on the test coil is more than 600 A for the field-decay measurement. The behavior of the joint resistance can also be studied in the presence of a 1.5 T background field. We have characterized many superconducting joints using this versatile 4 K insert. The joints between two different conductors can also be studied using this 4 K insert. The lowest joint resistance measured is 8 × 10^{-15} Ω at zero field. The joint resistance is also measured in the presence of the background field.

Published under an exclusive license by AIP Publishing. https://doi.org/10.1063/5.0161148

03 November 2023 05:51:45

I. INTRODUCTION

The superconducting magnets generate a strong magnetic field, which would otherwise be impossible with conventional, normal electromagnets. The superconducting magnets are thereby widely used in applications such as nuclear magnetic resonance (NMR), Magnetic Resonance Imaging (MRI), and high-energy particle accelerators.

Whole-body clinical MRI scanners need a highly homogeneous magnetic field and a highly stable field in the imaging volume. The magnet needs to provide a temporal field stability better than 0.1 ppm/h, i.e., 15 μT/h for a 1.5 T magnet, which is only possible through a persistent circulation of the current through the coils of the magnet shorted through a superconducting switch or the persistent current switch (PCS). In the MRI magnet, the superconducting coils are initially energized to the target current using a magnet ramping unit (MRU). Once the magnet is parked at the target field, the MRU is disconnected from the magnet after establishing the persistent current through the multi-coil magnet and the PCS. In the persistent mode, the magnet circuit consisting of the multiple superconducting coils and the PCS behaves as the L–R circuit. Hence, the temporal field stability is governed by the standard current-decay equation of the L–R circuit as given by the following equation:

$$I(t) = I_0 e^{\frac{-tR}{L}}, \tag{1}$$

where L is the self-inductance of the magnet, I_0 and I(t) are correspondingly the initial currents and the current at time t, and R is the effective electrical resistance of the circulating current path.

In most cases, the electrical resistance (R) of such magnet circuits is dominated by the electrical resistances of the inter-coil joints or the coil-to-PCS joints. The magnetic field is proportional to the current. Hence, Eq. (1) can be written as

$$B(t) = B_0 e^{\frac{-tR}{L}}, \tag{2}$$

where B_0 and B(t) correspond to the initial magnetic field and the magnetic field at time t.

Published under an exclusive license by AIP Publishing

Using Eq. (2), the field decay equation can be derived as given by the following equation:

$$\Delta B = B_0 - B(t) = B_0\left(1 - e^{\frac{-tR}{L}}\right). \tag{3}$$

Using Eq. (3), the total resistance of the circuit can be expressed in terms of total energy, $\left(E = \frac{1}{2}LI^2\right)$, which is given by the following equation:

$$R < \frac{2E * Field\ Decay}{3600I^2}, \tag{4}$$

where $\frac{\Delta B}{B_0}$ is the field decay in ppm, the total energy E is in MJ.

Currently, the Inter-University Accelerator Centre, New Delhi, is developing an actively shielded 1.5 T superconducting magnet for a whole-body clinical scanner.[1] The operating current of the magnet is ~450 A. The total stored energy of the magnet is about 3.7 MJ. To achieve field stability better than 0.1 ppm/h, the effective resistances should be less than 1×10^{-9} Ω. A multi-coil magnet typically has a few inter-coil joints and two coil-to-PCS joints. Hence, to achieve very high temporal stability, the electrical resistance of each such joint has to be ~10^{-13} Ω or better. Each individual electrical joint needs to be made *in situ* directly onto the magnet after the completion of the winding of the coils. There is no possibility to characterize the quality of such *in situ* joints until the MRI magnet is housed in a cryostat and cooled to 4.2 K. Hence, the jointing process needs to be established for their *in situ* implementation after rigorous testing of the sample joints. The development of superconducting joints with electrical resistance in the range 10^{-13}–10^{-15} Ω not only poses a big technological challenge, but precise measurements are also a challenging task. The superconducting joint in the MRI magnet experiences a significant amount of background field; therefore, it is necessary to characterize the joint in the presence of a background magnetic field.

The two most common methods of measuring the resistance of the superconducting joints are the four-probe method and the current or field-decay method. The four-probe technique is limited by the voltage measurement up to ~10^{-9} V, thereby limiting the measurement up to 10^{-11} Ω of the electrical resistance.[2,3] Whereas, the field-decay method has a good sensitivity until 10^{-15} Ω of the electrical resistance.[4–8] Although the field decay method cannot measure the resistance of an individual joint if two or more joints are connected in series; a longer duration is also needed to achieve a steady state. It is possible to reduce the duration of the measurement by reducing the self-inductance of the test loop, but it may reduce the overall accuracy. The overall measurement accuracy can be improved by increasing the induced current in the test coil; therefore, a higher induced field needs to be generated. This field decay technique measures the decaying magnetic field generated by an induced current into the sample coil or the test coil closed by the joints. The effective resistance is then calculated from the field-decay curve, thereby making it a direct measurement method. The superconducting quantum interference device (SQUID) based commercial magnetic property measurement system (MPMS) is also used for the measurement of electrical joint resistance.[9,10] The MPMS-based commercial measurement system, although it uses an indirect method to measure resistance, provides similar accuracy. Although the measurement time is substantially reduced in the SQUID-based system, it has a limitation in the size of the joint sample. In addition, all commercial measurement system, i.e., MPMS, does not allow for measuring the electrical resistance of the joints between two different conductors, which is necessary to characterize the coil-to-PCS joint, i.e., between the conductor (Cu–NbTi) of the magnet and the conductor (CuNi–NbTi) of the PCS.

We have developed a versatile 4 K insert-type test rig that does not need any specialized experimental dewar. It can be inserted into any helium dewar with a neck diameter of more than 90 mm. Our 4 K insert allows testing longer joint samples, i.e., up to 150 mm in length, either in straight configuration or even much longer in spiral configuration. It also allows for measuring the electrical resistance of joints between two different conductors. It has the capacity to induce a higher amplitude of current (>600 A) into the sample coil, thereby generating a higher amplitude of field. A superconducting solenoid magnet at the bottom of the insert allows us to characterize the joint at the background field up to 1.5 T. This paper briefly describes the details of the versatile 4 K insert for characterizing the superconducting joints having a longer length at a very high current in the presence of a higher background field.

II. WORKING PRINCIPLE

Figure 1 shows the simple schematic of the 4 K insert. A large amount of the current is induced into a small superconducting coil acting as the secondary coil of a superconducting transformer by another superconducting coil acting as the primary coil of the same transformer. Any change in flux in the primary coil will be conserved by the Lenz's law of flux conservation, which is given by the following equation:

$$\int_s (B_p + B_s)dA = 0. \tag{5}$$

When a shorted superconducting coil, i.e., the secondary coil, experiences a varying magnetic field (either during ramping up or ramping down of the primary coil), the current will be induced in the secondary coil to maintain the total magnetic flux because of their inductive coupling. Therefore, an induced field is generated by the secondary coil, which will slowly decay over time following Eq. (2).

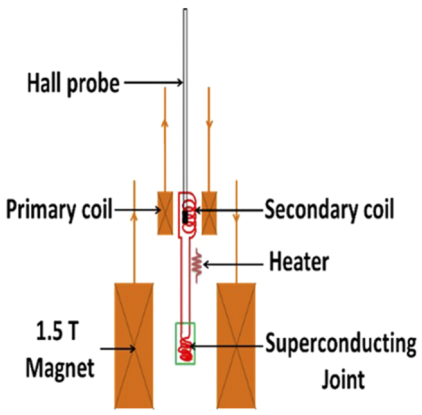


FIG. 1. Schematic representation of basic components of the 4 K insert.

03 November 2023 05:51:45

Published under an exclusive license by AIP Publishing

The decay of the shorted superconducting coil needs to be measured very precisely. The primary field needs to be zero during such decay measurements.

To achieve such a condition, the current into the secondary coil needs to be induced while ramping down the primary field to zero.

III. EXPERIMENTAL RIG

Figures 2(b) and 2(c), respectively, show the 3D model of the 4 K insert and the experimental dewar integrated with the 4 K insert. Figure 2(a) shows a photograph of the 4 K insert. It consists of the following major components:

a. superconducting primary coil,
b. superconducting secondary coil, i.e., test loop having joint(s),
c. 1.5 T superconducting solenoid magnet,
d. two pairs of vapor-cooled current leads (VCCLs).

The insert has a top lid which is used to mount the helium dewar. There is a 1.5 T superconducting solenoid magnet at the bottom of the insert, as shown in Fig. 2(b), which generates the background field for the superconducting joint. The magnet is supported by three GFRP rods that are fixed to the top lid of the insert. There is a small superconducting coil acting as the primary coil mounted at a certain height from the magnet, as shown in Fig. 2(b). The primary coil is positioned in the minimum background field. The secondary coil, which is part of the test coil, is placed concentrically and coaxially inside the primary coil. However, the test joint of the secondary coil is placed at the center of the superconducting magnet. There is a pair of vapor-cooled current leads connected to the leads of the primary coil. Another pair of vapor-cooled current leads is connected to the 1.5 T solenoid magnet. A manganin wire heater (36 AWG) with a resistance of 30 Ω is wound on the secondary coil to quench the test coil whenever necessary during the measurement. The GE-varnish is used to make thermal contact between the manganin wire heater and the secondary coil. The heater is fed 250 mA of excitation current to quench the secondary coil using a Keithley source meter (model 2400). It will dissipate 3.7 J of energy into the liquid helium bath in 2 s of quenching time. Both pairs of current leads have the capacity to carry 150 A of current. A liquid helium level sensor is fixed to the 4 K insert to measure the helium level during the experiment. However, the liquid helium level sensor is not essential, as most commercial helium dewars have differential-pressure-based helium level meters.

The design parameters of the primary coil are given in Table I. It is made up of a Cu (NbTi) conductor with a 1:1.35 superconductor to copper ratio. A total of 150 turns are wound on a glass fiber-reinforced plastic (GFRP) former of the primary coil, which is impregnated with Stycast 2850FT epoxy.[11] Figures 3(a) and 3(b), respectively, show the photograph and the schematic of the superconducting primary coil of the 4 K insert. The secondary coil is made up of the same superconducting wire that is used for the magnet, as the joint behavior needs to be studied for the same conductor using the 4 K insert. Typically, 12–15 turns are wound on a small GRRP or tufnol bobbin having a diameter less than the inner diameter of the former primary coil, thereby making it possible to place the secondary coil inside the primary coil (Table II).

The turns ratio between the primary and secondary coils is in the range of 10–12.5:1. The ratio of the cross-sectional areas between the primary and secondary coils is 3.4:1. Hence, the secondary coil will experience 66% of the flux of the primary coil. The secondary coil needs to be changed every time for the new test sample. Figure 4(a) shows the schematic representation of the

03 November 2023 05:51:45

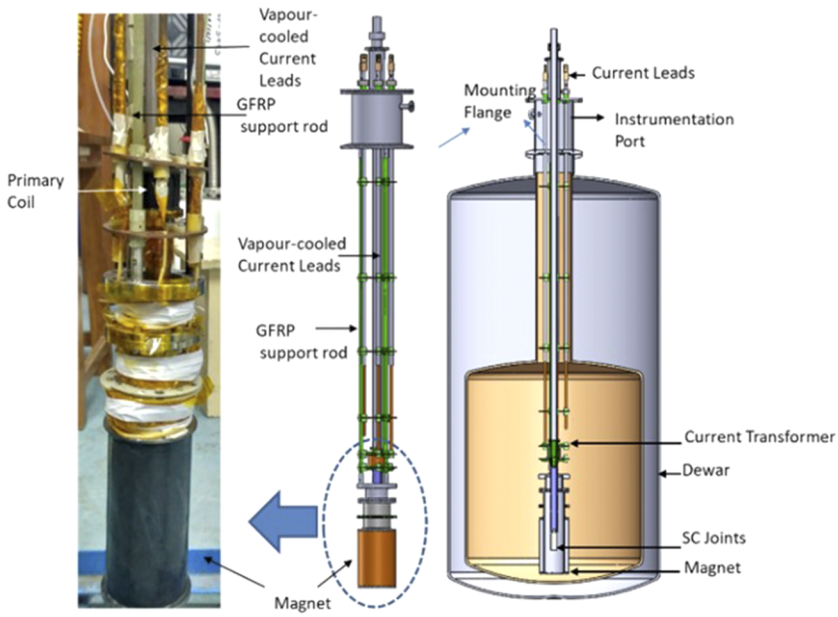


FIG. 2. (a) Photograph of the partial section of the 4 K insert, (b) 3D model of the 4 K insert, and (c) insert mounted into the 4 K dewar.

Published under an exclusive license by AIP Publishing

TABLE I. Parameters of the primary coil.

Description	Values
Winding length	30 mm
Total no. of turns	150
No. of layers	3
Inductance (L_p)	0.75 mH
Conductor	Cu (NbTi) (diameter 0.54 mm)
Field at 100 A	0.31 T

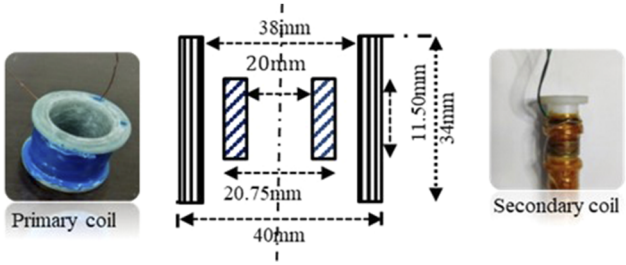


FIG. 3. (a) Photograph of the superconducting primary coil, (b) dimensional schematic of the superconducting primary and secondary coils, and (c) photograph of the superconducting secondary coil, i.e., the test coil of the 4 K insert.

straight single joint between two ends of a single superconducting wire (conductor-1). Figure 4(b) shows the schematic representation of the two straight joints between each pair of ends of two different superconducting wires (conductors-1 and 2). Figure 4(c) shows the schematic representation of the spiral joint between two ends of a single superconducting wire (conductor-1). The joint making process is not in the scope of this paper.

In the case of a single conductor, both configurations shown in Figs. 4(a) and 4(c) can be used. These configurations can be used for the inter-coil joining of the magnet. In the case of two conductors, the configuration shown in Fig. 4(b) can be used. The configuration shown in Fig. 4(c) is primarily for the coil-to-PCS joining. Any of the configurations shown in Fig. 4 can be used in the 4 K insert. For the straight joint as shown in Figs. 4(a) and 4(b), a copper or brass tube having an inner diameter in the range of 6–10 mm is used to hold the joint at the center of the magnet as shown in Fig. 5(c). Similarly, for the spiral joints as shown in Fig. 4(c), a copper or brass cup having an inner diameter in the range of 30–40 mm is used to hold the joint at the center of the magnet at the bottom of the 4 K

TABLE II. Parameters of the secondary coil.

Description	Values
Winding length	12 mm
Total no. of turns	12
No. of layer	1
Inductance (L_s)	5 μH
Conductor	Cu (NbTi)

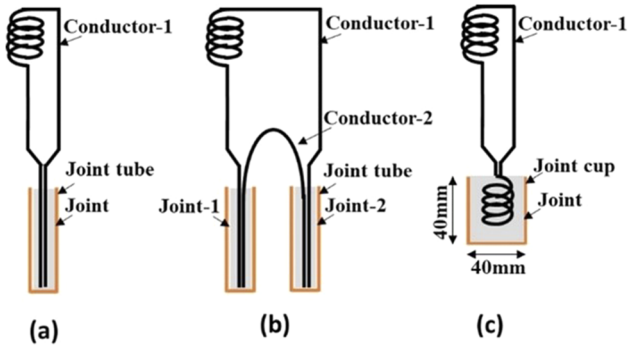


FIG. 4. Schematic representation of the configuration of (a) a single straight joint between two ends of the secondary coil or the test coil, (b) two straight joints between two pairs of ends of two different conductors, and (c) spiral joint between two ends of a single conductor of the secondary coil.

insert. The parameters of the solenoid magnet are given in Table III. The magnet will generate a 1.5 T magnet field at 53 A of the operating current. The diameter of the magnet bore is more than 60 mm, sufficient to place a maximum of two joint tubes side by side for straight joints [as shown in Fig. 4(b)] or the joint cup for the spiral joint [as shown in Fig. 4(c)]. The length of the magnet is also decided to accommodate the joint tube, which has a straight joint

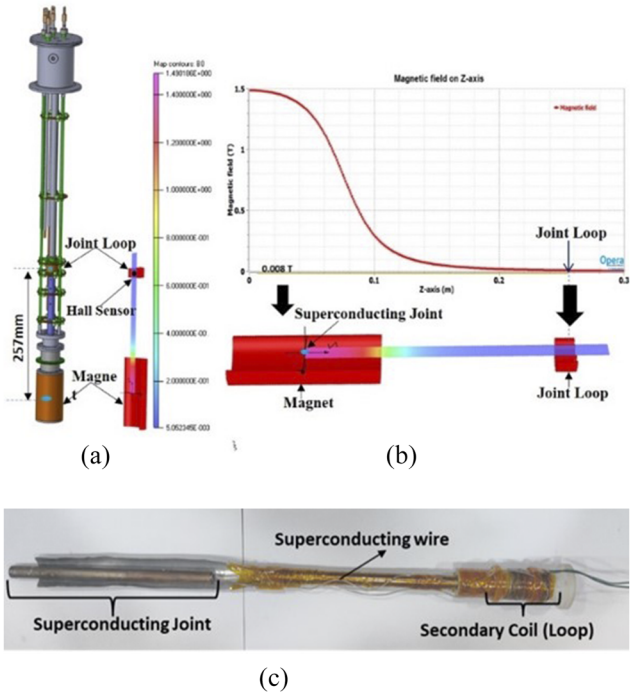


FIG. 5. (a) 3D model of the 4 K insert along with the relative position of the 1.5 T solenoid magnet with respect to the secondary coil, (b) axial field profile of the 1.5 T solenoid magnet of the 4 K insert, and (c) photograph of the sample joint having a straight configuration.

03 November 2023 05:51:45

Published under an exclusive license by AIP Publishing

TABLE III. Parameters of the superconducting magnet.

Items	Value
Total no. of turns	3704
Wire dia	0.55 mm
Cu:NbTi	2:1
Inductance (L)	0.32 H
Room temp. resistance	12 Ω
Magnetic field at 53 A	1.5 T
Diameter of the magnet bore	61 mm
Outer diameter of the magnet	76 mm
Length of the magnet	150 mm

with a maximum length of up to 150 mm. The outer diameter of the solenoid magnet is restricted by the inner diameter of the neck of the helium dewar. Figure 5(b) shows the position of the magnet in the 4 K insert and the field profile of the magnet along the vertical axis of the magnet. The center of the joint (straight or spiral) is placed at the center of the magnet to have the maximum background field during the testing, whereas the joint loop or the secondary loop is located far apart (~257 mm) from the center of the magnet, thereby minimizing the effect of the stray field at the center of the secondary coil or the test coil where the induced field is measured. At the 1.5 T background field, the stray field is 8 MT at the center of the test coil. The magnet generates the background field parallel to the axis of the magnet. Each joint is placed at the center of the magnet, axially parallel to the background field. The primary coil and the magnet are connected to two separate magnet power supplies having the capacity to feed 160 and 100 A current, respectively. However, a single power supply having a capacity of 160 A can also be used for both the secondary coil and the magnet. A cryogenic hall probe with a gaussmeter (Cryomagnetic, model LM-700) is placed at the center of the primary coil to measure the decay of the induced field.[12] The inner diameter of the former secondary coil is sufficient to place the axial hall probe at the center of the secondary coil. The gaussmeter is connected to a laptop through the RS232 interface for the recording of the magnetic field data.

Figure 6 shows a picture of the top lid of the helium dewar with the vapor-cooled current leads, the instrumentation port, and the liquid filling port. The cryogenic heat inleak, through the 4 K insert along with the heat flow through the neck tube of the dewar, will determine the rate of evaporation of the liquid helium, thereby determining the duration of the experimental measurement. Hence, the heat inleak needs to be minimized to have a longer duration of the experiment without any refilling during the measurement. There is a vent port on the top lid of the dewar to vent the evaporated helium gas. The evaporated helium gas through the VCCL is also connected to the same vent port as the dewar. To minimize the heat inleak to the liquid helium, the heat flow to the liquid helium has been analyzed. The solid conduction through various components of the 4 K insert is the major contributor to the heat inleak. The conduction heat transfer through the GFRP support rods, the SS tube to hold the secondary coil in position inside the insert, the SS tube of the hall sensor, and the instrumentation leads (for the heater and the level sensor) will primarily be responsible for the heat conduction to 4.2 K. The governing equation for the

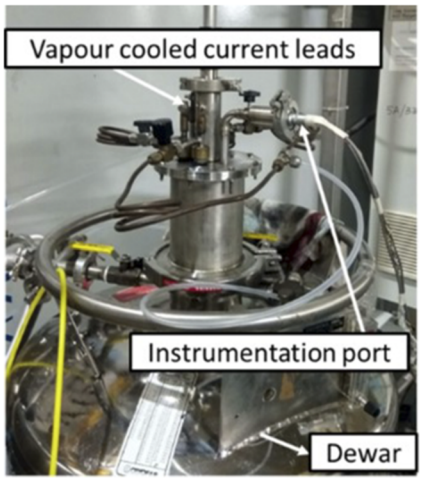


FIG. 6. Photograph of the top lid of the helium dewar with the 4 K insert.

heat conduction through each component is given in the following equation:

$$Q_i^c = \frac{A_i}{L_i}\int_{4.2}^{300} K_i(T)dT, \tag{6}$$

where Q (in W) is the heat conduction through the i-th component, A (in m^2) and L (in m) are, respectively, the cross-sectional area of the i-th component through which solid head conduction takes place and the length between 300 and 4.2 K of the i-th component. $\int_{4.2}^{300} K_i(T)dT$ is the thermal conductivity integral of the i-th component between 300 and 4 K. The cross-sectional area and the length of each component are calculated as per their dimensions summarized in Table IV. The thermal conductivity integral values of the GFRP, SS-304, and the sensor leads are taken from Refs. 13 and 14. The estimated heat inleak is summarized in Table IV.

The heat inleak through the vapor-cooled current lead is calculated based on the optimized value of the heat flow, i.e., 1 mW/A.[15,16] The radiation heat inleak through the neck is negligible as there are multiple baffles at different heights of the 4 K insert. Hence, the total heat inleak to 4.2 K is 0.8 W, which will evaporate 1.2 l of liquid helium per hour. The helium dewar has an inherent heat evaporation of 1% liquid per day, which is 0.042 l/h. Hence, the total rate of evaporation of the liquid is ~1.25 l/h. The heat inleak due to the energy dissipation through the heater during quenching of the secondary coil is not considered as it dissipates a momentary energy of 3.8 J. Such quenching is performed only once during a single set of measurements. The static heat flow through the leads of the heater has been considered in the overall heat inleak calculation. Figure 7 shows the complete test rig for the measurement of the electrical resistance of the superconducting joint. A safety relief valve is fixed at the neck of the helium dewar. During measurement, if the pressure of the helium bath increases beyond 5 psi due to the quenching of the magnet or the excessive energy dissipation from the current leads and heaters, the helium vapor will be released into the atmosphere through the safety valve.

03 November 2023 05:51:45

Published under an exclusive license by AIP Publishing

TABLE IV. Heat inleak to 4.2 K through the 4 K insert.

S/N	Item	Length (mm) and dia (mm)	Heat load (W)
1	GFRP rods (3)	606, ϕ8	0.067
2	Hall probe sensor tube	800, OD-12.5/ID-10	0.168
3	SS tube	907, OD-19/ID-16.5	0.204
4	VCCL (2) of the primary coil (2), 135 A	⋯	0.27
5	VCCL (2) of the magnet (2). 25 A	⋯	0.05
6	Instrumentation leads. Total heat inleak	⋯	0.05
		⋯	0.809

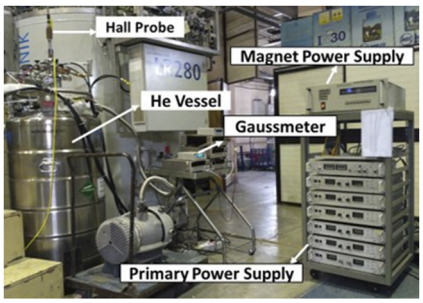


FIG. 7. Photograph of the test rig for the measurement of the joint resistance using a 4 K insert probe.

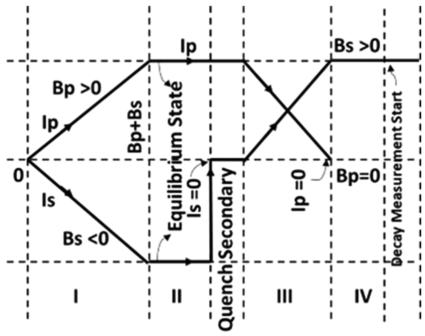


FIG. 8. Schematic representation of the major steps for the measurement of the joint resistance using a 4 K insert probe.

IV. JOINT SAMPLE

As described in Sec. III, the secondary coil is made of Cu (NbTi) wire. In the case of a single conductor, two ends of the conductor of the secondary coil are joined using Thornton's matrix replacement method.[17,18] The joint is then embedded in a copper tube or cup filled with the superconducting solder,[10] as shown in Figs. 4(a) and 4(c). In the case of two different conductors, two joints are made, which are then embedded in two separate copper tubes filled with superconducting solder, as shown in Fig. 4(b). The superconducting joints made by using any other standard techniques, such as ultrasonic welding, diffusion welding, cold-pressing, and spot welding,[19,20] can be characterized using this 4 K insert. The shorted secondary coil is then mounted on the 4 K insert by placing it concentrically and coaxially to the superconducting primary coil.

V. MEASUREMENT PROCESS

The measurement process is started once the 4 K insert with the test coils is cooled to 4.2 K using liquid helium inside the dewar. It has various intermediate steps prior to starting the measurement of the field-decay. Figure 8 shows the schematic representation of each step. There are five major steps described below:

I. Ramp up the superconducting primary coil at a certain ramp rate up to the desired current (I_p), which will generate the primary field (B_p). The change in magnetic flux in the primary coil will induce the current (I_s) into the secondary coil or the test coil in the opposite direction to generate a secondary field (B_s).
II. The cryogenic hall probe measures the resultant field ($B_p + B_s$) at the center of the secondary coil. A heat pulse of ~1.9 W is given to the secondary coil for 2 s through the heater wire (I_h = 0.25 A) to quench. Thereby, the induced current reduces to zero (I_s = 0), as shown in Fig. 8. The hall probe measures the primary field (B_p).
III. Ramp down the primary coil to zero (I_p = 0) at a certain ramp rate. The change in the magnetic flux by the primary coil will again induce the current (I_s) into the secondary coil or the test coil in the opposite direction to generate the secondary field (B_s).
IV. The primary field (B_p) is reduced to zero, whereas the secondary field (B_s) is generated. The hall probe measures the secondary field (B_s).
V. Measure the decay of the secondary field (B_s) using the hall probe. Initially, the decay is measured at the zero-field (ZF). Once the decay rate is stabilized at ZF, the magnet at the bottom is energized to measure the decay rate in the presence of the background field.

VI. RESULTS AND DISCUSSION

Figure 9 shows the variation of the magnetic field at the center of the primary and secondary coils during the process of current induction in one of the test coils or the secondary coil using the 4 K insert. Table V shows the various parameters during the testing of two separate joints. In set-1, the joints have been made between

03 November 2023 05:51:45

Published under an exclusive license by AIP Publishing

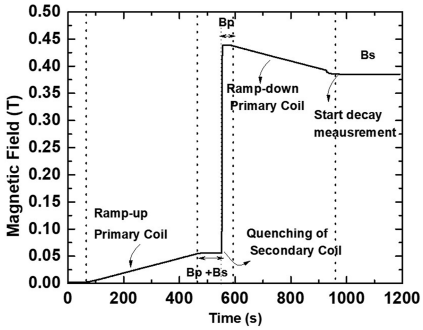


FIG. 9. Variation of the magnetic field with time during the process of current induction into the secondary coil using the 4 K insert.

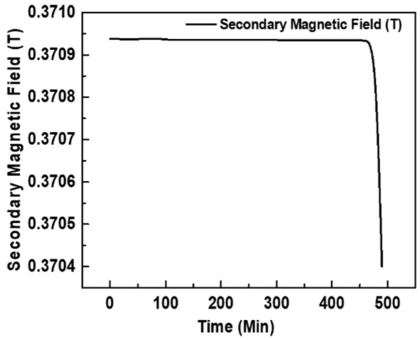


FIG. 10. Long-term measurement of the decay of the secondary field using the 4 K insert.

TABLE V. Parameters of the two sets of measurements.

Parameter	Set-1 Cu (NbTi)–CuNi (NbTi)	Set-2 Cu (NbTi)
I_P	135 A	135 A
$B_p + B_s$	54.72 mT	60.1 mT
I_H	250 mA	250 mA
B_p	0.438 72 T	0.441 43 T
B_s	0.385 10 T	0.381 43 T
I_s	~525 A	~520 A

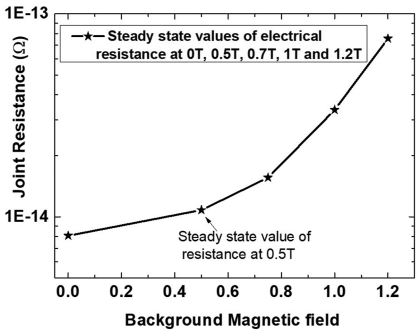


FIG. 11. Steady-state values of the joint resistance calculated from the field decay at various background fields using the 4 K insert.

two conductors, Cu(NbTi) and CuNi(NbTi), using the configuration shown in Fig. 4(b). In set-2, the joint has been made using a single conductor. The primary coil in set-1 is ramped up to 135 A of current. At the steady state, the resultant field ($B_p + B_s$) is measured to be ~54.5 mT. At the ideal condition, i.e., with a 100% coupling of the flux between the primary and secondary coils, the total field will be zero. The non-zero resultant field signifies the partial coupling, which is discussed in Sec. III. The secondary coil is then quenched by 0.25 A of current to the wire heater, as indicated in Fig. 9. After quenching the secondary coil, the field at the center is only due to the primary coil, which is 0.438 72 T. The current through the primary coil is then ramped down to zero to again induce the current into the secondary coil. When the primary current reaches zero, the field measured at the center is only due to the secondary coil, which is 0.385 10 T. The secondary field slowly decays over time, which is measured for calculating the joint resistance. The secondary field corresponds to the 525 A of the induced current in it, which signifies 3.9 times higher current induction than the primary coil. At 160 A of primary current, almost 620 A of current can be induced into the secondary coil. It almost takes ~1000 s to complete one cycle of current induction into the secondary coil. The decay measurement can be started after that.

In the beginning, the field decay is faster due to the transient settling phenomenon that redistributes the current over the superconducting filaments. The decay measurement is continued until the steady state value of joint resistance is reached. It takes almost an hour to reach the steady state value. However, any long-term measurement, if needed, can be done using this 4 K insert without any refilling of the liquid helium. Figure 10 shows one such long-term measurement of the secondary field for a different joint sample. The measurement was continued for 450 min until it was intentionally quenched using the heat pulse through the wire heater.

After reaching the steady state value of the joint resistance at the zero background field (ZF), the solenoid magnet is energized up to the desired current to generate the background field that is parallel to the joint. In the beginning, the field decay is again faster due to the transient settling phenomenon, which redistributes the current over the superconducting filaments in the presence of the background field. The background field is maintained at a constant value until the field decay reaches the steady state value. Once the joint resistance reaches the steady state value at that background field, the current to the magnet is increased to the next higher level. Figure 11 shows the steady state values of the resistance of a superconducting joint at different background magnetic fields. At zero field, the electrical resistance is ~8×10^{-15} Ω. The steady-state values of the electrical resistance have been measured at 0.5, 0.7, 1.0, and 1.2 T background fields, as shown in Fig. 11. It is observed that the resistance of the joint deteriorates at higher fields. The electrical resistances are 1×10^{-14} and 3.5×10^{-14} Ω, respectively, at 0.5 and 1 T of background field.

VII. CONCLUSION

A versatile 4 K insert has been developed for characterizing the superconducting joints. The lowest value of the electrical resistance

03 November 2023 05:51:45

Published under an exclusive license by AIP Publishing

measured for a superconducting joint between two different conductors, Cu (NbTi) and CuNi (NbTi), is 8×10^{-15} Ω at zero field. The maximum current induced in the secondary coil is 525 A during a measurement although it has the capacity to induce more than 600 A of current, thereby generating a higher amplitude of the secondary field. The repeatability of the measured values has also been checked for the same superconducting joint. A long-term (~8 h) measurement of field decay has also been carried out using the 4 K insert. The rate of liquid depletion in the experimental vessel closely matches the heat inleak to 4.2 K.

ACKNOWLEDGMENTS

The authors would like to thank the Cryogenic Group of Inter-University Accelerator Centre, New Delhi, for providing liquid helium for this experimental study. The authors would like to acknowledge the technical support provided by Dr. Navneet Suman, Sankar ram Theketthil, Vijay Soni, S. K. Saini, Rajesh Kumar, and Dr. Ram Gopal Sharma for providing technical support during the measurement. The authors would also like to convey thanks to SAMEER-Mumbai and MeitY for their financial support in carrying out this activity for the IMRI project.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Ajit Nandawadekar: Formal analysis (equal); Methodology (equal); Software (equal); Writing – original draft (equal). **Mukhtiar Singh**: Investigation (equal); Supervision (equal); Visualization (equal); Writing – review & editing (equal). **Soumen Kar**: Conceptualization (lead); Formal analysis (equal); Funding acquisition (lead); Methodology (equal); Supervision (equal); Writing – review & editing (lead).

DATA AVAILABILITY

The data that support the findings of this study are available on request from the corresponding author.

REFERENCES

[1] S. Kar, N. Suman, T. S. Ram, V. Soni, R. Kumar, S. K. Saini, J. Antony, S. Sahu, M. Kumar, S. Babu *et al.*, "Development of high homogeneity and high stability 1.5 T superconducting magnet for whole body MRI scanner," Indian J. Cryog. **44**(1), 193 (2019).
[2] J. Jaroszynski, "Race against time: Resistance of superconducting joints measurements," Supercond. Sci. Technol. **28**(1), 010501 (2014).
[3] D. T. Ryan, H. Jones, W. Timms, and N. Killoran, "Critical current measurements at electric fields in the pV m/sup −1/regime," IEEE Trans. Appl. Supercond. **7**(2), 1455–1458 (1997).
[4] C. Swenson and W. Markiewicz, "Persistent joint development for high field NMR," IEEE Trans. Appl. Supercond. **9**(2), 185–188 (1999).
[5] J. Liu, J. Cheng, F. Zhou, Q. Wang, K. Chang, and X. Li, "Electrical properties of cold-pressing welded NbTi persistent joints," Cryogenics **58**, 62–67 (2013).
[6] J. Cheng, J. Liu, Z. Ni, C. Cui, S. Chen, S. Song, L. Li, Q. Wang, and Y. Dai, "Fabrication of NbTi superconducting joints for 400-MHz NMR application," IEEE Trans. Appl. Supercond. **22**, 4300205 (2012).
[7] D. Ni, W. Yang, S. Han, L. Zhu, X. Zhang, B. Wu, and L. Ma, "A test system for superconducting joints based on GM cryocooler," IEEE Trans. Appl. Supercond. **24**(3), 9000505 (2014).
[8] T. Fukuzaki, H. Maeda, S. Matsumoto, S. Yokoyama, and T. Kiyoshi, "Study of joint resistance in Nb_3Al–NbTi superconducting joint for high field NMR," IEEE Trans. Appl. Supercond. **17**(2), 1435–1437 (2007).
[9] G. D. Brittles, P. Noonan, S. A. Keys, C. R. M. Grovenor, and S. Speller, "Rapid characterisation of persistent current joints by SQUID magnetometry," Supercond. Sci. Technol. **27**(12), 122002 (2014).
[10] G. D. Brittles, T. Mousavi, C. R. M. Grovenor, C. Aksoy, and S. C. Speller, "Persistent current joints between technological superconductors," Supercond. Sci. Technol. **28**(9), 093001 (2015).
[11] See https://datasheets.tdx.henkel.com/LOCTITE-STYCAST-2850FT-BL-en_US.pdf for more information about for Stycast 2850 FT epoxy.
[12] See https://cryomagnetics.com/wpontent/uploads/2022/05/GM_700_Rev0.pdf for more information about for the cryogenic hall probe.
[13] J. W. Ekin, *Experimental Techniques for Low-Temperature Measurements* (Oxford University Press, Oxford, 2006).
[14] See https://trc.nist.gov/cryogenics/materials/material properties.htm for more information about Cryogenics Material Properties, Trc.nist.gov, 2020; accessed 9 September 2022.
[15] K. R. Efferson, "Helium vapor cooled current leads," Rev. Sci. Instrum. **38**, 1776 (1967).
[16] M. N. Wilson, *Superconducting Magnets* (Clarendon Press, Oxford, 1983).
[17] R. Thornton, "Superconducting joint for superconducting wires and coils," US Patent 4584547 (18 April 1986).
[18] D. Patel, S.-H. Kim, W. Qiu, M. Maeda, and A. Matsumoto, G. Nishijima, H. Kumakura, S. Choi, and J. H. Kim, "Niobium–titanium (Nb–Ti) superconducting joints for persistent-mode operation," Sci. Rep. **9**, 14287 (2019).
[19] S. Phillip, J. V. Porto, and J. M. Parpia, "Two methods of fabricating reliable superconducting joints with multifilamentary Nb–Ti superconducting wire," J. Low Temp. Phys. **101**, 581–585 (1995).
[20] W. Tan, "Different techniques of producing superconducting joints and methods of joints characterization," J. Phys.: Conf. Ser. **1635**, 012008 (2020).

03 November 2023 05:51:45

Published under an exclusive license by AIP Publishing

Taylor & Francis
Taylor & Francis Group

Numerical Functional Analysis and Optimization

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/lnfa20

An approach to preserve functions with exponential growth by using modified Lupaş-Kantrovich operators

Neha Kajla & Naokant Deo

To cite this article: Neha Kajla & Naokant Deo (2023) An approach to preserve functions with exponential growth by using modified Lupaş-Kantrovich operators, Numerical Functional Analysis and Optimization, 44:14, 1510-1522, DOI: 10.1080/01630563.2023.2263977

To link to this article: https://doi.org/10.1080/01630563.2023.2263977

Published online: 30 Oct 2023.

Submit your article to this journal

Article views: 41

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=lnfa20

NUMERICAL FUNCTIONAL ANALYSIS AND OPTIMIZATION
2023, VOL. 44, NO. 14, 1510–1522
https://doi.org/10.1080/01630563.2023.2263977

Taylor & Francis
Taylor & Francis Group

Check for updates

An approach to preserve functions with exponential growth by using modified Lupaş-Kantrovich operators

Neha Kajla and Naokant Deo

Department of Applied Mathematics, Delhi Technological University, Delhi, India

ABSTRACT

As part of this study, we propose a modification of the so-known Lupaş-Kantrovich that preserve exponential function e^{-x}. To support this claim, we estimate the convergence rate of the operators in terms of both the usual and exponential modulus of continuity. Our analysis also includes a global estimate and quantitative Voronovskaya results. Demonstrating the effectiveness of modified operators, we provided a result and supporting graphs.

ARTICLE HISTORY
Received 7 June 2023
Accepted 28 June 2023

KEYWORDS
Exponential functions; rate of convergence; Voronovskaya theorem

MATHEMATICS SUBJECT CLASSIFICATION
41A25; 41A36

1. Introduction

King [1] received recognition in 2003 for his modification of Bernstein operators that preserve test functions e_0 and e_2 on $[0, 1]$. Later, King drew the attention of scholars to this topic, and they presented numerous related studies. Many researchers have done exceptional work in this area up to this point, defining operators that preserve e_0 and e_2, e_2+ae_1 for $a > 0$, exponential functions, linear functions, and so on. According to the overall purpose of this paper, we will limit the scope of our investigation to only preserving the exponential functions. We know, studying how to preserve exponential functions is still at its dawn. Let us provide some of the most recent sources for this subject here. Acar et al. [2], published the Szász-Mirakyan operator modification that preserves e^{2ax}, for $a \in (0, \infty]$. They discussed shape preservation properties and compared modified operators to Szász-Mirakyan operators. They also used a natural transformation to estimate the error in terms of the modulus of continuity of the first order. Further, these modified operators [2] were studied in detail by Aral et al. [3], who demonstrated their importance from a computational perspective. Acar et al. [4] calculated the approximation order using a new proposed weighted modulus of smoothness and introduced the Szász-Mirakyan operators that fix $e^{\alpha x}$ and $e^{2\alpha x}$ where $\alpha > 0$ concurrently. Furthermore, they presented some saturation results for their modified operators to illustrate the accuracy of their estimates. Reader can also refer to the article [5–7] and references therein.

CONTACT Neha Kajla neha_phd2k18@dtu.ac.in Department of Applied Mathematics, Delhi Technological University, Bawana Road, Delhi-110042, India.
© 2023 Taylor & Francis Group, LLC

Many operators have been modified during the last four years using the same idea, including Bernstein [3], Baskakov-Szász-Mirakyan [8], Stancu Szász-Mirakyan-Durrmeyer operators [9], Phillips operators [10, 11], Baskakov-Schurer-Szász-Stancu operators [12], and Baskakov-Schurer-Szász operators [13]. Also, Deo et al. [14, 15] suggested the operators based on King's technique that outperforms the Baskakov Durrmeyer operators and Szász-Mirakyan Durrmeyer in terms of convergence rate. Yilmaz et al. [16] altered the Baskakov-Kantorovich operators in such a way that we get a sequence of operators that preserve constant functions and e^x. The whole purpose of our paper is to give an alteration of the operators [1] that preserves constant functions and e^{-x}.

We have organized this manuscript as described below. Method for constructing the operators is covered in Section 2. The idea of moment-generating function is used in Section 3 to determine the moment and central moment for modified operators. In Section 4, we demonstrate uniform convergence and use a theorem of the Voronovskaya type to examine the asymptotic behavior. In Section 5, we will present a result and accompanying graphs to illustrate the efficiency of our modified operators.

2. Construction of operators

In the year 1995, Lupaş [17] proposed a discrete operators and established some direct results. For $f \in C[0,\infty)$, the Lupaş operators are defined as

$$\mathcal{G}_m\left(f;x\right) = \sum_{k=0}^{\infty} g_{m,k}\left(x\right) f\left(\frac{k}{m}\right),$$

where

$$g_{m,k}\left(x\right) = 2^{-mx}\frac{\left(mx\right)_k}{2^k k!}.$$

In [18, p. 210], the modified form of Lupaş Kantorovich operators for the objective of approximating integral functions is defined as:

$$G_m\left(f;x\right) = \left(m+1\right)\sum_{k=0}^{\infty} 2^{-mx}\frac{\left(mx\right)_k}{2^k k!}\int_{\frac{k}{m+1}}^{\frac{k+1}{m+1}} f\left(u\right)du. \tag{2.1}$$

We aim to construct operators that preserve 1 and e^{-x}. Let us assume operators (2.1) for some function $\lambda_m\left(x\right)$

$$G_m\left(e^{-u};\bar{\lambda}_m\left(x\right)\right) = \left(m+1\right)\sum_{k=0}^{\infty} 2^{-m\bar{\lambda}_m(x)}\frac{\left(m\bar{\lambda}(x)\right)_k}{2^k k!}\int_{\frac{k}{m+1}}^{\frac{k+1}{m+1}} e^{-u}du,$$

therefore

$$G_m\left(e^{-u};\bar{\lambda}_m(x)\right) = -(m+1)\,2^{-m\bar{\lambda}_m(x)}\left(e^{-\frac{1}{m+1}}-1\right)\sum_{k=0}^{\infty}\frac{(m\bar{\lambda}(x))_k}{k!}\left(\frac{e^{-\frac{1}{m+1}}}{2}\right)^k.$$

Now using the binomial series $\sum_{k=0}^{\infty}\frac{(a)_k}{k!}z^k = (1-z)^{-a}\ |z|<1$. we obtain

$$\begin{aligned} G_m\left(e^{-u};\bar{\lambda}_m(x)\right) &= -(m+1)\,2^{-m\bar{\lambda}_m(x)}\left(e^{-\frac{1}{m+1}}-1\right)\left(1-\frac{e^{-\frac{1}{m+1}}}{2}\right)^{-m\bar{\lambda}_m(x)} \\ &= -(m+1)\left(e^{-\frac{1}{m+1}}-1\right)\left(2-e^{-\frac{1}{m+1}}\right)^{-m\bar{\lambda}_m(x)}. \end{aligned}$$

Taking $G_m\left(e^{-u};\bar{\lambda}_m(x)\right) = e^{-x}$ into account, we can calculate

$$\bar{\lambda}_m(x) = \frac{x+\log\left((1+m)\left(1-e^{\frac{-1}{m+1}}\right)\right)}{m\log\left(2-e^{-\frac{1}{m+1}}\right)}. \tag{2.2}$$

We analyze the modified form of the operators (2.1)

$$\tilde{\mathcal{K}}_m(f;x) = (m+1)\sum_{k=0}^{\infty}2^{-m\bar{\lambda}_m(x)}\frac{(m\bar{\lambda}(x))_k}{2^k k!}\int_{\frac{k}{m+1}}^{\frac{k+1}{m+1}} f(u)du, \tag{2.3}$$

where $\bar{\lambda}_m(x)$ is given in (2.2).

3. Preliminaries

Performing basic calculations, the moment-generating function of the operators (2.3) can be expressed as:

$$\tilde{\mathcal{K}}_m\left(e^{\theta u};x\right) = \frac{1}{\theta}(m+1)\left(e^{\frac{\theta}{m+1}}-1\right)\left(2-e^{\frac{\theta}{m+1}}\right)^{-m\bar{\lambda}_m(x)}. \tag{3.1}$$

Since the moments are linked to the moment-generating function, the $r-$th moment $\tilde{\mathcal{K}}_m(e_r;x), e_r(u) = u^r\ (r\in\mathbb{N}\cup\{0\})$ can be calculated by employing the following relation:

$$\begin{aligned} \tilde{\mathcal{K}}_m(e_r;x) &= \left[\frac{\partial^r}{\partial\theta^r}\tilde{\mathcal{K}}_m\left(e^{\theta u};x\right)\right]_{\theta=0} \\ &= \left[\frac{\partial^r}{\partial\theta^r}\frac{1}{\theta}(m+1)\left(e^{\frac{\theta}{m+1}}-1\right)\left(2-e^{\frac{\theta}{m+1}}\right)^{-m\bar{\lambda}_m(x)}\right]_{\theta=0} \end{aligned}$$

Employing Mathematica software, the expansion of the above statement in powers of θ would be as follows:

$$\begin{aligned}
&\tilde{\mathcal{K}}_m\left(e^{\theta u};x\right)\\
&=\frac{1}{\theta}(m+1)\left(e^{\frac{\theta}{m+1}}-1\right)\left(2-e^{\frac{\theta}{m+1}}\right)^{-m\bar{\chi}_m(x)}\\
&=1+\frac{\theta(2m\bar{\chi}_m(x)+1)}{2(m+1)}+\frac{\theta^2\left(3m^2\left(\bar{\chi}_m(x)\right)^2+9m\bar{\chi}_m(x)+1\right)}{6(m+1)^2}\\
&+\frac{\theta^3\left(4m^3\left(\bar{\chi}_m(x)\right)^3+30m^2\left(\bar{\chi}_m(x)\right)^2+40m\bar{\chi}_m(x)+1\right)}{24(m+1)^3}\\
&+\frac{\theta^4\left(5m^4\left(\bar{\chi}_m(x)\right)^4+70m^3\left(\bar{\chi}_m(x)\right)^3+250m^2\left(\bar{\chi}_m(x)\right)^2+215m\bar{\chi}_m(x)+1\right)}{120(m+1)^4}+O\left(\theta^5\right). \quad (3.2)
\end{aligned}$$

Moreover, by changing the scale attribute of moment-generating function, if we can expand $e^{-\theta x}\tilde{\mathcal{K}}_m\left(e^{\theta u};x\right)$ in powers of θ, the r-th order central moment $\psi_r(x)=\tilde{\mathcal{K}}_m\left((u-x)^r;x\right)$ may be produced by accumulating the coefficient of $\frac{\theta^r}{r!}$.

$$\begin{aligned}
&e^{-\theta x}\tilde{\mathcal{K}}_m\left(e^{\theta u};x\right)\\
&=e^{-\theta x}\frac{1}{\theta}(m+1)\left(e^{\frac{\theta}{m+1}}-1\right)\left(2-e^{\frac{\theta}{m+1}}\right)^{-m\bar{\chi}_m(x)}\\
&=1+\frac{\theta(2m\bar{\chi}_m(x)-2mx-2x+1)}{2(m+1)}\\
&+\frac{\theta^2\left(3m^2\left(\bar{\chi}_m(x)\right)^2+3m^2x^2-6m^2\bar{\chi}_m(x)x+9m\bar{\chi}_m(x)+6mx^2-6m\bar{\chi}_m(x)x-3mx+3x^2-3x+1\right)}{6(m+1)^2}\\
&+\frac{\theta^3}{24(m+1)^3}(4m^3\left(\bar{\chi}_m(x)\right)^3-m^3x^3+12m^3\bar{\chi}_m(x)x^2-12m^3\left(\bar{\chi}_m(x)\right)^2x+30m^2\left(\bar{\chi}_m(x)\right)^2-12m^2x^3\\
&+24\bar{\chi}_m(x)m^2x^2+6m^2x^2-12m^2\left(\bar{\chi}_m(x)\right)^2x-36m^2\bar{\chi}_m(x)x+40m\bar{\chi}_m(x)-12mx^3+12m\bar{\chi}_m(x)x^2\\
&+12mx^2-36m\bar{\chi}_m(x)x-4mx-4x^3+6x^2-4x+1)\\
&+\frac{\theta^4}{120(m+1)^4}(5m^4\left(\bar{\chi}_m(x)\right)^4+5m^4x^4-20m^4\bar{\chi}_m(x)x^3+30m^4\left(\bar{\chi}_m(x)\right)^2x^2-20m^4\left(\bar{\chi}_m(x)\right)^3x\\
&+70m^3\left(\bar{\chi}_m(x)\right)^3+20m^3x^4-60m^3\bar{\chi}_m(x)x^3-10m^3x^3+60m^3\left(\bar{\chi}_m(x)\right)^2x^2+90m^3\bar{\chi}_m(x)x^2\\
&-20m^3\left(\bar{\chi}_m(x)\right)^3x-150m^3\left(\bar{\chi}_m(x)\right)^2x+250m^2\left(\bar{\chi}_m(x)\right)^2+30m^2x^4-60m^2\bar{\chi}_m(x)x^3-30m^2x^3\\
&+30m^2\left(\bar{\chi}_m(x)\right)^2x^2+180m^2\bar{\chi}_m(x)x^2+10m^2x^2-150m^2\left(\bar{\chi}_m(x)\right)^2x-200m^2\bar{\chi}_m(x)x+215m\bar{\chi}_m(x)\\
&+20mx^4-20m\bar{\chi}_m(x)x^3-30mx^3+90m\bar{\chi}_m(x)x^2+20mx^2-200m\bar{\chi}_m(x)x-5mx+5x^4-10x^3\\
&+10x^2-5x+1)+O\left(\theta^5\right).
\end{aligned}$$

Lemma 3.1. *We calculate the moments of the proposed operator (2.3) using (3.2) as described below:*

$$\begin{aligned}
&\tilde{\mathcal{K}}_m(e_0;x)=1\\
&\tilde{\mathcal{K}}_m(e_1;x)=\frac{(2m\bar{\chi}_m(x)+1)}{2(m+1)}\\
&\tilde{\mathcal{K}}_m(e_2;x)=\frac{\left(3m^2\left(\bar{\chi}_m(x)\right)^2+9m\bar{\chi}_m(x)+1\right)}{3(m+1)^2}
\end{aligned}$$

$$\tilde{\mathcal{K}}_m(e_3;x) = \frac{\left(4m^3(\tilde{\lambda}_m(x))^3 + 30m^2(\tilde{\lambda}_m(x))^2 + 40m\tilde{\lambda}_m(x) + 1\right)}{4(m+1)^3}$$

$$\tilde{\mathcal{K}}_m(e_4;x) = \frac{\left(5m^4(\tilde{\lambda}_m(x))^4 + 70m^3(\tilde{\lambda}_m(x))^3 + 250m^2(\tilde{\lambda}_m(x))^2 + 215m\tilde{\lambda}_m(x) + 1\right)}{5(m+1)^4}.$$

Lemma 3.2. *As a result of the above analysis, the first four moments are as follows:*

$$\psi_1(x) = \frac{(2m\tilde{\lambda}_m(x) - 2mx - 2x + 1)}{2(m+1)}$$

$$\psi_2(x) = \frac{\left(3m^2(\tilde{\lambda}_m(x))^2 + 3m^2x^2 - 6m^2\tilde{\lambda}_m(x)x + 9m\tilde{\lambda}_m(x) + 6mx^2 - 6m\tilde{\lambda}_m(x)x - 3mx + 3x^2 - 3x + 1\right)}{3(m+1)^2}$$

$$\begin{aligned}\psi_3(x) = {} & \frac{1}{4(m+1)^3}(4m^3(\tilde{\lambda}_m(x))^3 - 4m^3x^3 + 12m^3\tilde{\lambda}_m(x)x^2 - 12m^3(\tilde{\lambda}_m(x))^2x + 30m^2(\tilde{\lambda}_m(x))^2 - 12m^2x^3 \\ & + 24m^2\tilde{\lambda}_m(x)x^2 + 6m^2x^2 - 12m^2(\tilde{\lambda}_m(x))^2x - 36m^2\tilde{\lambda}_m(x)x + 40m\tilde{\lambda}_m(x) - 12mx^3 + 12m\tilde{\lambda}_m(x)x^2 \\ & + 12mx^2 - 36m\tilde{\lambda}_m(x)x - 4mx - 4x^3 + 6x^2 - 4x + 1)\end{aligned}$$

$$\begin{aligned}\psi_4(x) = {} & \frac{1}{5(m+1)^4}(5m^4(\tilde{\lambda}_m(x))^4 + 5m^4x^4 - 20m^4\tilde{\lambda}_m(x)x^3 + 30m^4(\tilde{\lambda}_m(x))^2x^2 - 20m^4(\tilde{\lambda}_m(x))^3x \\ & + 70m^3(\tilde{\lambda}_m(x))^3 + 20m^3x^4 - 60m^3\tilde{\lambda}_m(x)x^3 - 10m^3x^3 + 60m^3(\tilde{\lambda}_m(x))^2x^2 + 90m^3\tilde{\lambda}_m(x)x^2 \\ & - 20m^3(\tilde{\lambda}_m(x))^3x - 150m^3(\tilde{\lambda}_m(x))^2x + 250m^2\tilde{\lambda}_m(x)^2 + 30m^2x^4 - 60m^2\tilde{\lambda}_m(x)x^3 \\ & - 30m^2x^3 + 30m^2(\tilde{\lambda}_m(x))^2x^2 + 180m^2\tilde{\lambda}_m(x)x^2 + 10m^2x^2 - 150m^2(\tilde{\lambda}_m(x))^2x - 200m^2\tilde{\lambda}_m(x)x \\ & + 215n\tilde{\lambda}_m(x) + 20nx^4 - 20n\tilde{\lambda}_m(x)x^3 - 30nx^3 + 90n\tilde{\lambda}_m(x)x^2 + 20nx^2 - 200n\tilde{\lambda}_m(x)x - 5nx + 5x^4 \\ & - 10x^3 + 10x^2 - 5x + 1).\end{aligned}$$

Also,

$$\lim_{m\to\infty} m\psi_1(x) = \lim_{m\to\infty} m\left[\frac{(2m\tilde{\lambda}_m(x) - 2mx - 2x + 1)}{2(m+1)}\right] = x$$

and

$$\begin{aligned}\lim_{m\to\infty} m\psi_2(x) = \lim_{m\to\infty} \frac{m}{3(m+1)^2}\big[3m^2(\tilde{\lambda}_m(x))^2 + 3m^2x^2 - 6m^2\tilde{\lambda}_m(x)x + 9m\tilde{\lambda}_m(x) \\ +6mx^2 - 6m\tilde{\lambda}_m(x)x - 3mx + 3x^2 - 3x + 1\big] = 2x.\end{aligned}$$

4. Main results

The subspace of real-valued continuous functions with uniform norms that have finite limits at infinity can be represented as $C^*[0,\infty)$. For a function over an infinite interval, Boyanov [19] described its approximation properties. Holhoş [20] later proved the next theorem to figure out how fast a function will converge.

Theorem. *Let $Q : C^*[0,\infty) \to C^*[0,\infty)$ be the sequence of linear positive operators and*

$$\begin{aligned}\|Q(e_0) - 1\|_{[0,\infty)} &= \alpha_1(m), \\ \left\|Q\left(e^{-u}\right) - e^{-x}\right\|_{[0,\infty)} &= \alpha_2(m), \\ \left\|Q\left(e^{-2u}\right) - e^{-2x}\right\|_{[0,\infty)} &= \alpha_3(m).\end{aligned}$$

then

$$\left\|Q_m f - f\right\|_{[0,\infty)} \leq \alpha_1(m)\left\|f\right\|_{[0,\infty)} + (2+\alpha_1(m))\omega^*(f, \sqrt{\alpha_1(m)+2\alpha_2(m)+\alpha_3(m)}).$$

The modulus of continuity can be expressed as:

$$\omega^*(\tilde{\nu}, \delta) = \sup_{\substack{|e^{-x}-e^{-u}|\leq\delta \\ x,u>0}} \left|\tilde{\nu}(x) - \tilde{\nu}(u)\right|$$

with the property

$$\left|\tilde{\nu}(u) - \tilde{\nu}(x)\right| \leq \left(1 + \frac{\left(e^{-u}-e^{-x}\right)^2}{\delta^2}\right)\omega^*\left(\tilde{\nu}, \delta\right), \quad \delta > 0. \tag{4.1}$$

In the presented theorem, we present quantitative estimates for proposed operators as an application of the previous theory.

Theorem 4.1. *For $f \in C^*[0,\infty)$, the inequality*

$$\left\|\tilde{\mathcal{K}}_m f - f\right\|_{[0,\infty)} \leq 2\omega^*\left(f, \sqrt{\alpha_3(m)}\right),$$

holds true. Here $\tilde{\mathcal{K}}_m f$ converges to f uniformly and $\alpha_3(m) \to 0$ as $m \to \infty$.

Proof The operators preserve constant functions as well as e^{-x} so $\alpha_1(m) = \alpha_2(m) = 0$. we simply need to evaluate $\alpha_3(m)$. From (3.1), we obtain

$$\tilde{\mathcal{K}}_m\left(e^{-2u}; x\right) = \frac{-1}{2}(m+1)\left(e^{\frac{-2}{m+1}} - 1\right)\left(2 - e^{\frac{-2}{m+1}}\right)^{-\bar{\lambda}_m(x)}$$

where

$$\bar{\lambda}_m(x) = \frac{x + \log\left((1+m)\left(1 - e^{\frac{-1}{m+1}}\right)\right)}{m\log\left(2 - e^{-\frac{1}{m+1}}\right)}.$$

Employing Mathematica software, we obtain

$$\tilde{\mathcal{K}}_m\left(e^{-2u}; x\right) = e^{-2x} + \frac{2xe^{-2x}}{m} + \frac{\left(24x^2 - 72x - 11\right)e^{-2x}}{12m^2} + O\left(\left(\frac{1}{m}\right)^3\right).$$

Since

$$\sup_{x\in[0,\infty)} xe^{-2x} = \frac{1}{2e}, \quad \sup_{x\in[0,\infty)} x^2e^{-2x} = \frac{1}{e^2}.$$

So, we obtain

$$\begin{aligned}\alpha_3(m) &= \left\|\tilde{\mathcal{K}}_m\left(e^{-2u}\right) - e^{-2x}\right\|_{[0,\infty)} \\ &= \sup_{x\in[0,\infty)}\left|\tilde{\mathcal{K}}_m\left(e^{-2u}\right) - e^{-2x}\right| \\ &\leq \frac{1}{m}\left(\frac{1}{e}\right) + \frac{1}{m^2}\left(\frac{2}{e^2} + \frac{3}{e} + \frac{11}{12}\right) + O\left(\left(\frac{1}{m}\right)^3\right) \\ &\leq O\left(\frac{1}{m}\right) \to 0 \text{ as } m \to \infty.\end{aligned}$$

□

Remark 4.1. *With the aid of Mathematica and the Lemma 3.2, we find*

$$\begin{aligned}\lim_{n\to\infty} m^2\psi_4(x) = \lim_{n\to\infty} \frac{m^2}{5(n+1)^4}&(5m^4(\bar{\chi}_m(x))^4 + 5m^4x^4 - 20m^4\bar{\chi}_m(x)x^3 + 30m^4(\bar{\chi}_m(x))^2x^2 \\ &- 20m^4(\bar{\chi}_m(x))^3x + 70m^3(\bar{\chi}_m(x))^3 + 20m^3x^4 - 60m^3\bar{\chi}_m(x)x^3 - 10m^3x^3 \\ &+ 60m^3(\bar{\chi}_m(x))^2x^2 + 90m^3\bar{\chi}_m(x)x^2 - 20m^3(\bar{\chi}_m(x))^3x - 150m^3(\bar{\chi}_m(x))^2x \\ &+ 250m^2\bar{\chi}_m(x)^2 + 30m^2x^4 - 60m^2\bar{\chi}_m(x)x^3 - 30m^2x^3 + 30m^2(\bar{\chi}_m(x))^2x^2 \\ &+ 180m^2\bar{\chi}_m(x)x^2 + 10m^2x^2 - 150m^2(\bar{\chi}_m(x))^2x - 200m^2\bar{\chi}_m(x)x + 215m\bar{\chi}_m(x) \\ &+ 20mx^4 - 20m\bar{\chi}_m(x)x^3 - 30mx^3 + 90m\bar{\chi}_m(x)x^2 + 20mx^2 - 200m\bar{\chi}_m(x)x \\ &- 5mx + 5x^4 - 10x^3 + 10x^2 - 5x + 1) \\ &= 12x^2\end{aligned}$$

and

$$\lim_{m\to\infty} m^2\tilde{\mathcal{K}}_m\left(\left(e^{-x} - e^{-u}\right)^4; x\right) = 12e^{-4x}x^2.$$

Theorem 4.2. *For $x \in [0,\infty)$, and $f, f'' \in C^*[0,\infty)$, we obtain*

$$\begin{aligned}&\left|m\left[\tilde{\mathcal{K}}_m(f;x) - f(x)\right] - x\left[f'(x) + f''(x)\right]\right| \\ &\leq |\alpha_m(x)|\left|f'(x)\right| + |\beta_m(x)|\left|f''(x)\right| + 4\omega^*\left(f'', \frac{1}{\sqrt{m}}\right)((\beta_m(x) + x) + \gamma_m(x)).\end{aligned}$$

Proof Using Taylor's expansion, we obtain

$$f(u) = f(x) + (u-x)f'(x) + \frac{1}{2}(u-x)^2f''(x) + \tilde{\nu}(u,x)(u-x)^2,$$

where

$$\tilde{\nu}(u,x) = \frac{f''(\mu) - f''(x)}{2}, \quad x < \mu < u.$$

Applying the operators $\tilde{\mathcal{K}}_m$ to both sides of the aforementioned expression and the Lemma (3.2) results in the following expression:

$$\left|\tilde{\mathcal{K}}_m(f;x) - f(x) - \psi_1(x)f'(x) - \frac{1}{2}\psi_2(x)f''(x)\right| \leq \left|\tilde{\mathcal{K}}_m\left(\tilde{\nu}(u,x)(u-x)^2; x\right)\right|.$$

Using Lemma 3.2

$$\left|m\left[\tilde{\mathcal{K}}_m\left(f;x\right)-f\left(x\right)\right]-xf'\left(x\right)-\frac{1}{2}\left(2x\right)f''\left(x\right)\right|$$
$$\leq\left|m\left(\psi_1\left(x\right)\right)-x\right|\left|f'\left(x\right)\right|+\frac{1}{2}\left|m\left(\psi_2\left(x\right)\right)-2x\right|\left|f''\left(x\right)\right|$$
$$+\left|m\tilde{\mathcal{K}}_m\left(\tilde{\nu}\left(u,x\right)\left(u-x\right)^2;x\right)\right|.$$

Taking $\alpha_m\left(x\right)=m\left(\psi_1\left(x\right)\right)-x$ and $\beta_m\left(x\right)=\frac{1}{2}\left|m\left(\psi_2\left(x\right)\right)-2x\right|$, we get

$$\left|m\left[\tilde{\mathcal{K}}_m\left(f;x\right)-f\left(x\right)\right]-x\left[f'\left(x\right)+f''\left(x\right)\right]\right|$$
$$\leq\left|\alpha_m\left(x\right)\right|\left|f'\left(x\right)\right|+\left|\beta_m\left(x\right)\right|\left|f''\left(x\right)\right|+\left|m\tilde{\mathcal{K}}_m\left(\tilde{\nu}\left(u,x\right)\left(u-x\right)^2;x\right)\right|.$$

To conclude our proof, we need to find the estimate of $\left|n\tilde{\mathcal{K}}_m\left(\tilde{\nu}\left(u,x\right)\left(u-x\right)^2;x\right)\right|$. Employing inequality (4.1), we arrive at

$$\left|\tilde{\nu}\left(u,x\right)\right|\leq\left(1+\frac{\left(e^{-u}-e^{-x}\right)^2}{\delta^2}\right)\omega^*\left(f'',\delta\right).$$

For the cases $\left|e^{-u}-e^{-x}\right|\leq\delta$ and $\left|e^{-u}-e^{-x}\right|>\delta$ the two inequalities $\left|\tilde{\nu}\left(u,x\right)\right|\leq 2\omega^*\left(f'',\delta\right)$ and $\left|\tilde{\nu}\left(u,x\right)\right|\leq\frac{2\left(e^{-u}-e^{-x}\right)^2}{\delta^2}\omega^*\left(f'',\delta\right)$ holds true respectively.

Thus

$$\left|\tilde{\nu}\left(u,x\right)\right|\leq 2\left(1+\frac{\left(e^{-u}-e^{-x}\right)^2}{\delta^2}\omega^*\left(f'',\delta\right)\right).$$

Making use of preceding argument and Cauchy Schwarz inequality, we obtain

$$m\tilde{\mathcal{K}}_m\left(\tilde{\nu}\left(u,x\right)\left(u-x\right)^2;x\right)$$
$$\leq m\tilde{\mathcal{K}}_m\left(2\left(1+\frac{\left(e^{-u}-e^{-x}\right)^2}{\delta^2}\omega^*\left(f'',\delta\right)\right)\left(u-x\right)^2;x\right)$$
$$=2m\left(\psi_2\left(x\right)\right)\omega^*\left(f'',\delta\right)+\frac{2m}{\delta^2}\omega^*\left(f'',\delta\right)\tilde{\mathcal{K}}_m\left(\left(e^{-u}-e^{-x}\right)^2\left(u-x\right)^2;x\right)$$
$$=2\omega^*\left(f'',\delta\right)\left[m\left(\psi_2\left(x\right)\right)+\frac{m}{\delta^2}\left(\psi_4\left(x\right)\right)^{1/2}\left(\tilde{\mathcal{K}}_m\left(\left(e^{-u}-e^{-x}\right)^2;x\right)\right)^{1/2}\right].$$

Conclude the result by selecting $\delta=\frac{1}{\sqrt{m}}$ and $\gamma_m\left(x\right)=\left(\psi_4\left(x\right)\right)^{1/2}\left(\tilde{\mathcal{K}}_m\left(\left(e^{-u}-e^{-x}\right)^2;x\right)\right)^{1/2}$. □

Theorem 4.3. *For $f,f''\in C^*[0,\infty)$ and $x\in[0,\infty)$, the following equality holds*

$$\lim_{m\to\infty}m\left[\tilde{\mathcal{K}}_m\left(f;x\right)-f\left(x\right)\right]=x\left[f'\left(x\right)+f''\left(x\right)\right].$$

Proof With the help of Taylor's expansion of f, we obtain

$$f(u) = f(x) + (u-x)f'(x) + \frac{1}{2}(u-x)^2 f''(x) + \tilde{\nu}(u,x)(u-x)^2. \quad (4.2)$$

where

$$\lim_{u\to x} \tilde{\nu}(u,x) = 0.$$

Applying $\tilde{\mathcal{K}}_m$ to both sides of (4.2) and from Lemma 3.2, we obtain

$$\tilde{\mathcal{K}}_m(f;x) - f(x) = \psi_1(x) f'(x) + \frac{1}{2}\psi_2(x) f''(x) + \tilde{\mathcal{K}}_m\left(\tilde{\nu}(u,x)(u-x)^2; x\right).$$

The Cauchy Schwarz inequality can be used to determine

$$\tilde{\mathcal{K}}_m\left(\tilde{\nu}(u,x)(u-x)^2; x\right) \leq \sqrt{\tilde{\mathcal{K}}_m\left(\tilde{\nu}^2(u,x); x\right)\tilde{\mathcal{K}}_m\left((u-x)^4; x\right)} \quad (4.3)$$

Also, we have

$$\lim_{m\to\infty} \tilde{\mathcal{K}}_m\left(\tilde{\nu}^2(u,x); x\right) = 0. \quad (4.4)$$

From (4.3) and (4.4), we obtain

$$\lim_{m\to\infty} m\tilde{\mathcal{K}}_m\left(\tilde{\nu}(u,x)(u-x)^2; x\right) = 0.$$

Thus we achieve

$$\begin{aligned}
&\lim_{m\to\infty} m\left[\tilde{\mathcal{K}}_m(f;x) - f(x)\right] \\
&= \lim_{n\to\infty} m\left[\psi_1(x) f'(x) + \frac{1}{2}\psi_2(x) f''(x) + \tilde{\mathcal{K}}_m\left(\tilde{\nu}(u,x)(u-x)^2; x\right)\right] \\
&= x\left[f'(x) + f''(x)\right].
\end{aligned}$$

□

Let $C_B[0,\infty)$ denote the space of real valued contiuous and bounded function f on the interval $[0,\infty)$, endowed with the norm,

$$\|f\| = \sup_{x\in[0,\infty)} |f(x)|$$

For any $\delta > 0$,
the Peetre's K-functional is defined by:

$$K_2(f;\delta) = \inf_{l\in C_B[0,\infty)} \left\{\|f-l\| + \delta\|l''\|\right\},$$

where

$$C_B^2[0,\infty) = \left\{g \in C_B[0,\infty) : l', l'' \in C_B[0,\infty)\right\}.$$

By DeVore and Lorentz [21] there exist an absolute constant $C > 0$ such that

$$K_2(f;\delta) \leqslant C\omega_2\left(f;\sqrt{\delta}\right),$$

where modulus of continuity of second order is denoted by

$$\omega_2\left(f;\sqrt{\delta}\right) = \sup_{x\in[0,\infty)} \sup_{0<h\leqslant\sqrt{\delta}} |f(x+2h) - 2f(x+h) + f(x)|$$

Theorem 4.4. *Let $f \in C_B[0,\infty)$, then the inequality*

$$\left|\tilde{\mathcal{K}}_m\left(f;x\right)-f\left(x\right)\right| \leq M\omega_2\left(f,\frac{1}{2}\sqrt{\psi_2\left(x\right)+\frac{\left(\frac{(2m\bar{\lambda}_m(x)+1)}{2(m+1)}-x\right)^2}{2}}\right) + \omega\left(f,\left(\frac{(2m\bar{\lambda}_m\left(x\right)+1)}{2(m+1)}-x\right)\right),$$

holds true for all $x \in [0,\infty)$ and a positive constant M.

Proof We establish the auxiliary operators $\mathbb{K}_m : C_B[0,\infty) \to C_B[0,\infty)$

$$\mathbb{K}_m\left(f;x\right)=\tilde{\mathcal{K}}_m\left(v;x\right)+f\left(x\right)-f\left(\frac{(2m\bar{\lambda}_m\left(x\right)+1)}{2(m+1)}\right).$$

For the operators (2.3), we have

$$\left\|\tilde{\mathcal{K}}_m\left(f;x\right)\right\| \leq \left\|f\right\|,$$

implies

$$\left\|\mathbb{K}_m\left(f;x\right)\right\| \leq 3\left\|f\right\|. \tag{4.5}$$

Additionally, for $\hbar \in C_B^2[0,\infty)$, Taylor expansion is presented as

$$\tilde{\nu}\left(u\right)-\tilde{\nu}\left(x\right)=\left(u-x\right)\tilde{\nu}'\left(x\right)+\int_x^u\left(u-\mu\right)\tilde{\nu}''\left(\mu\right)d\mu, \qquad x\in\mathbb{R}[0,\infty)$$

By applying operators $\mathbb{K}_m\left(f;x\right)$ on both sides of the above equation and utilizing the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}\left|\mathbb{K}_m\left(\tilde{\nu};x\right)-\tilde{\nu}\left(x\right)\right| &= \left|\mathbb{K}_m\left(\int_x^u\left(u-\mu\right)\tilde{\nu}''\left(\mu\right)d\mu;x\right)\right| \\ &\leq \left|\tilde{\mathcal{K}}_m\left(\int_x^u\left(u-\mu\right)\tilde{\nu}''\left(\mu\right)d\mu;x\right)\right| \\ &\quad+\left|\int_x^{\frac{(2m\bar{\lambda}_m(x)+1)}{2(m+1)}}\left(\frac{(2m\bar{\lambda}_m\left(x\right)+1)}{2(m+1)}-\mu\right)\tilde{\nu}''\left(\mu\right)d\mu\right| \\ &\leq \left\|\tilde{\nu}''\right\|\left(\psi_2\left(x\right)+\frac{\left(\frac{(2m\bar{\lambda}_m(x)+1)}{2(m+1)}-x\right)^2}{2}\right).\end{aligned} \tag{4.6}$$

Using the estimates from equations (4.5) and (4.6), we can finally conclude that

$$\begin{aligned}
&\left|\tilde{\mathcal{K}}_m\left(f;x\right)-f\left(x\right)\right| \\
&\leq \left|\mathbb{K}_m\left(f-\tilde{v};x\right)-\left(f-\tilde{v}\right)\left(x\right)\right|+\left|f\left(\frac{\left(2m\tilde{\lambda}_m\left(x\right)+1\right)}{2\left(m+1\right)}\right)-f\left(x\right)\right|+\left|\mathbb{K}_m\left(\tilde{v};x\right)-\tilde{v}\left(x\right)\right| \\
&\leq 4\left\|f-\tilde{v}\right\|+\left(\psi_2\left(x\right)+\frac{\left(\frac{\left(2m\tilde{\lambda}_m(x)+1\right)}{2(m+1)}-x\right)^2}{2}\right)\left\|\tilde{v}''\right\|+\left|v\left(\frac{\left(2m\tilde{\lambda}_m\left(x\right)+1\right)}{2\left(m+1\right)}\right)-f\left(x\right)\right| \\
&\leq 4K_2\left(f,\frac{1}{4}\left(\psi_2\left(x\right)+\frac{\left(\frac{\left(2m\tilde{\lambda}_m(x)+1\right)}{2(m+1)}-x\right)^2}{2}\right)\right)+\left|f\left(\frac{\left(2m\tilde{\lambda}_m\left(x\right)+1\right)}{2\left(m+1\right)}\right)-f\left(x\right)\right| \\
&\leq M\omega_2\left(f,\frac{1}{2}\sqrt{\left(\psi_2\left(x\right)+\frac{\left(\frac{\left(2m\tilde{\lambda}_m(x)+1\right)}{2(m+1)}-x\right)^2}{2}\right)}\right)+\omega\left(f,\left(\frac{\left(2m\tilde{\lambda}_m\left(x\right)+1\right)}{2\left(m+1\right)}-x\right)\right).
\end{aligned}$$

□

Example 4.1. The function $f(x) = \cos(x)$ is approximated by the operators $\tilde{\mathcal{K}}_m$ in Figure 1 for $m = 10, 20, 40$. The operators (2.3) clearly converges to the function $f(x)$ as m is increased. We calculate the error $|\tilde{\mathcal{K}}_m\left(f;x\right) - f(x)|$ in Table 1 and also give a graph depicting the error in Figure 2.

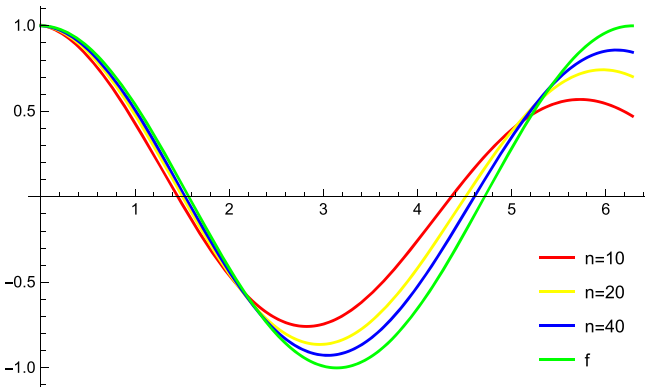


Figure 1. Considering $n = 10, 20, 40$, the convergence of operators $\tilde{\mathcal{K}}_m$ toward the function $f(x) = \cos(x)$.

Table 1. Error estimation for $f(x) = cos(x)$ when $n = 10, 20, 40, 60$.

x	$n = 10$	$n = 20$	$n = 40$	$n = 60$
$\frac{\pi}{2}$	0.107714	0.0648587	0.0356679	0.0245513
π	0.286407	0.152325	0.0777487	0.0520773
$\frac{3\pi}{2}$	0.236268	0.166841	0.0992171	0.0700621
2π	0.528039	0.296931	0.154454	0.103863

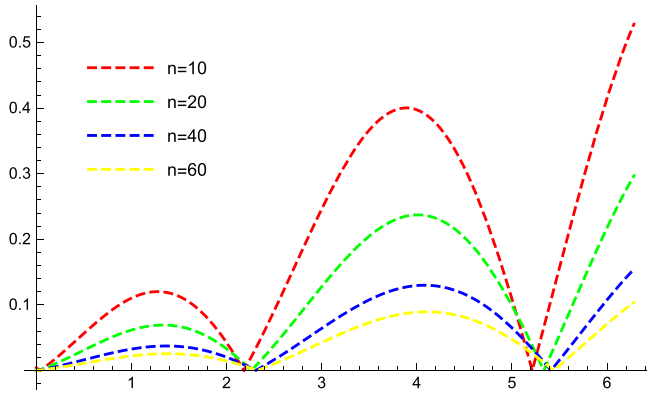


Figure 2. Considering $n = 10, 20, 40, 60$, the error of operators $\tilde{\mathcal{K}}_m$ toward the function $f(x) = \cos(x)$.

5. Significance statement

This work is carried out to find an extension of the Lupaş-Kantrovich operators preserves the exponential function e^{-x}. Using both the usual modulus of continuity and the exponentiated modulus of continuity, we estimate the convergence rate and there's also a global estimate and Voronovskaya results. Graphs and results are provided to demonstrate the effectiveness of modified operators.

Disclosure statement

This work has no conflicts of interest.

Funding

CSIR is funding this research with Reference No:08/133(0021)/2018-EMR-1 for the first author.

References

[1] King, J. P. (2003). Positive linear operators which preserve x^2. *Acta Math. Hungarica* 99(3):203–208.

[2] Acar, T., Aral, A., Gonska, H. (2017). On Szász–Mirakyan operators preserving e^{2ax}, $a > 0$. *Mediterr. J. Math.* 14(1):1–14.

[3] Aral, A., Inoan, D., Rasa, I. (2019). Approximation properties of Szász-Mirakyan operators preserving exponential functions. *Positivity* 23(1):233–246.

[4] Acar, T., Aral, A., Cárdenas–Morales, D., Garrancho, P. (2017). Szász–Mirakyan type operators which fix exponentials. *Results Math.* 72(3):1393–1404.

[5] Lipi, Km., Deo, N. (2020). General family of exponential operators. *Filomat* 34(12): 4043–4061.

[6] Dhamija, M., Deo, N. (2017). Better approximation results by Bernstein-Kantorovich operators. *Lobachevskii J. Math.* 38:94–100.

[7] Gupta, V., Deo, N. (2005). On the rate of convergence for bivariate Beta operators. *Gen. Math.* 13(3):109–116.

[8] Gupta, V., Acu, A. M. (2018). On Baskakov–Szász–Mirakyan–type operators preserving exponential type functions. *Positivity* 22(3):919–929.

[9] Kanat, K., Sofyalioglu, M. (2020). On Stancu type Szász-Mirakyan-Durrmeyer operators preserving e^{2ax}, $a > 0$. *Gazi Univ. J. Sci*. 34(1):196–209.

[10] Gupta, V., López-Moreno, A. J. (2018). Phillips operators preserving arbitrary exponential functions, e^{at}, e^{bt}. *Filomat*, 32(14):5071–5082.

[11] Tachev, G. (2019). On two modified Phillips operators. *Studia Universitatis Babeş-Bolyai Math* 64(3):305–312.

[12] Sofyalioğlu, M., Kanat, K. (2019). Approximation properties of generalized Baskakov–Schurer–Szász–Stancu operators preserving e^{-2ax}, $a > 0$. *J. Inequalities Appl.* 112:1–16.

[13] Yilmaz, Ö. G., Bodur, M., Aral, A. (2018). On approximation properties of Baskakov–Schurer–Szász operators preserving exponential functions. *Filomat* 32(15):5433–5440.

[14] Deo, N., Bhardwaj, N. (2011). Some approximation results for Durrmeyer operators. *Appl. Math. Comput.* 217(12):5531–5536.

[15] Deo, N., Ozarslan, M. A., Bhardwaj, N. (2014). Statistical convergence for general beta operators. *Korean J. Math.* 22(4):671–681.

[16] Yilmaz, Ö. G., Gupta, V., Aral, A. (2018). A note on Baskakov–Kantorovich type operators preserving e^{-x}. *Math. Methods Appl. Sci.* 43(13):7511–7517.

[17] Lupas, A. (1995). The approximation by means of some linear positive operators. In: Muller, M. W., Felten, M., Mache, D. H., eds. *Approximation Theory (Proceedings of the International Dortmund Meeting IDoMAT 95*, Held inWitten, March 13–17, 1995). Mathematical Research, Vol. 86. Berlin: Akademie Verlag, pp. 201–229.

[18] Gupta, V., Malik, N., Rassias, T. M. (2018). Moment generating functions and moments of linear positive operators. In: Daras, N., Rassias, T., eds. *Modern Discrete Mathematics and Analysis*. Springer Optimization and Its Applications, Vol. 131. Cham: Springer.

[19] Boyanov, B. D., Veselinov, V. M. (1970). A note on the approximation of functions in an infinite interval by linear positive operators. *Bull. Math. Soc. Sci. Math. Roum.* 14(62): 9–13.

[20] Holhoş, A. (2010). The rate of approximation of functions in an infinite interval by positive linear operators. *Studia Universitatis Babeş–Bolyai Math* 55(2):133–142.

[21] DeVore, R. A., Lorentz G. G. (1993). *Constructive Approximation.* Berlin: Springer.

Sādhanā (2023)48:222
https://doi.org/10.1007/s12046-023-02271-4
© Indian Academy of Sciences

An improved teaching learning based optimization method to enrich the flight control of a helicopter system

ABHISHEK CHAUDHARY* and BHARAT BHUSHAN

Electrical Engineering Department, Delhi Technological University, New Delhi, India
e-mail: abhishek@dtu.ac.in; bharat@dce.ac.in

MS received 6 January 2023; revised 11 May 2023; accepted 17 July 2023

Abstract. A helicopter is a multivariable, nonlinear, higher order and strongly coupled system. The helicopter dynamics is subjected to unknown external disturbances and sensitive to parametric uncertainties. Such complications require a sophisticated control approach which may handle all these difficulties. The paper presents a trajectory control approach using an improved teaching learning based optimization (iTLBO) method, which optimizes the parameters of proportional integral derivative controller. This control approach is validated by minimizing the yaw and pitch errors of two degrees of freedom helicopter system. It is observed that the experimental results in a laboratory setup does not stand true when subjected to weather changes during the helicopter flight. To overcome this situation, a real time like conditions are created in the laboratory using the two external fans creating strong wind disturbance and the proposed controller is implemented to eliminate the resulting external disturbances created due to high speed wind of external fans. Initially, modelling of helicopter system is achieved using the Lagrangian approach. Then, the TLBO is tuned to improve the controller performance using MATLAB platform. The effectiveness and superiority of iTLBO is proved over other optimization techniques available in literature like GA, PSO and TLBO by using linear and nonlinear benchmark functions. The optimized parameters hence obtained are implemented on helicopter system under the presence of external air disturbance and disturbance less results are obtained. Results obtained are validated using time response analysis on the MATLAB/Simulation platform.

Keywords. 2 DOF helicopter; trajectory control; disturbance rejection; teaching learning based optimization (TLBO); external disturbance; proportional integral derivative (PID) controller.

1. Introduction

Recently in military, industries and in agriculture, the use of robots, drones and unmanned helicopters is in great mandate as they have exceptional features which includes hovering, the capability of taking off vertical and capability to fly at comparatively lower altitudes by means of variable velocity [1–3]. The two degrees of freedom (2-DOF) helicopter by Quanser [4] is a multiple-input-multiple-output (MIMO), highly nonlinear, firmly coupled higher-order benchmark system. The LQR controller and LQR-I controller are planned by Quanser which are providing comfort in stability and in pursuing desired trajectory input. However, applied controller is not always able to precisely knob uncertainty like externally introduced disturbances. For years, considerable papers are being published using helicopter system as a test bench to investigate the value of altered controlling stratagems.

In last two decades, different control approaches have been offered by researchers for tracking down the desirable pitch position and yaw angle position of helicopter [5–10]. Here, uncertainties like, un-modeled elements and external disturbances are handled by designing a robust controller [5]. H∞ loop shaping framework is prolonged by applying two procedures of compensator synthesis methodology [6]. A discrete time neural control strategy is implemented to handle the angle of pitch axis and yaw axis of helicopter system [7]. A methodology of back stepping controller synthesis is presented in [8]. A different robust control scheme is presented after suitably coalescing adaptive control method [9]. Robust control strategy using multivariable adaptive sliding mode observer is implemented on 2DOF helicopter system [10]. Several control strategies are developed for three degrees of freedom unmanned helicopter systems also, as more than a few controller algorithms grounded on linear quadratic regulator (LQR), back stepping control, sliding mode control (SMC), model predictive control (MPC) controller etc. [11–14]. Recent work presents the designing of robust controlling of unmanned chopper is available in [15, 16], where Jacobian linearized model of the system is used to design the controller. Wind and payload disturbances are discussed in [15] and actuator

*For correspondence
Published online: 21 October 2023

faults are reflected in [16]. To ensure the asymptotic tracking performance, a compound control scheme having feedback and feedforward mechanism is used [17]. However, these nonlinear controllers are usually complex in calculation and designing. To compensate this, higher order controller designing with disturbance rejection capability of robust attitude tracking is used [18]. It clears how researchers are led to enhance the handling of many nonlinear models in the field of control system engineering by applying various types of linear and nonlinear controllers.

Considerable amount of linear controllers are offered by several literatures to handle many difficulties in underactuated systems [19–21]. These controllers are providing a simpler way of scheming closed loop handling of nonlinear systems. The speed of system response is heavily affected by the linearization of the nonlinear systems. It demands the expansion of several nonlinear controllers to handle the problems faced by underactuated systems. As a result a lot of nonlinear controllers like Lagrangian method, lambda method and back stepping controllers are evolved among researchers and the same is available in literature [22–24]. During early stage, these controllers confronted some disadvantages due to the limitations of existing lambda method in dealing with controllers lagging due to supplementary response. Apart from these mentioned controllers, to attain the non-linear control, the adaptive iterative learning methods are also developed [25]. The drawback of the iterative learning method is reliance on multiple number of calculations which increases exponentially when we increase state variables [26]. Motion of an underactuated Helicopter is mechanically constrained in pitch axes, yaw axis and voltage ratings which shows that states and input constraints of the helicopter system cannot be ignored at the very initial stage of controllers designing to avoid degradation of nonlinear helicopter system. Another method available to handle this nonlinearity is model predictive control, which is the constrained optimal controlling method [27, 28]. The LQR controller which is used in feedback to incorporate the constraint accompanying with the systems in the designer stage is implemented using linear model in [29]. The robust adaptive fuzzy controller design was implemented using a gradient descent algorithm [30]. Control techniques like adaptive sliding mode controller [31], model predictive controller [32] and adaptive disturbances rejection control [33] are also implemented and used by researchers in literature at length. Numerous control techniques available in the literature for achieving stable trajectories in helicopter system, the PID control is extensively adapted in maximum numbers of real engineering solicitations. The rewards with the PID controlling strategy are modest configuration, great reliability, and decent stability. It is also pertaining to mention the key drawback while implementing the traditional form of PID controller is their problems with the tuning process of unconstrained parameters.

1.1 *Related work*

In literatures, in order to accomplish the tuning of parameters of PID, there are several number of different approaches available such as self-tuning algorithms, genetic and evolutionary algorithms which further helped in improving the working stability of nonlinear under actuated systems like helicopter system, aerodynamic systems, ball balancer systems & nonlinear pendulums [36–39]. In [40], an overview is reported on applications of an evolutionary algorithm in the field of control engineering. It clearly shows a scope where such types of optimization problems may be solved using Evolutionary computation. Particle Swarm Optimization (PSO) by Kennedy and Eberhart, is used to stabilize the control design of a nonlinear system [41]. This metaheuristic particle search algorithm rivals the collective comportment of bird flocking and fish schooling in search of food. This optimization is widely being used by researchers ever since. Using PSO, a research was overtaken on attitude control strategy of unmanned helicopter system in [42]. Where PSO is used to tune the parameters of proportional integral derivative controller. In a similar approach, the parameters of LQR controller is optimized using PSO method to handle the control of helicopter system [43]. The Ant Colony Optimization method is another approach which is applied to find optimum PID parameters in [44]. The original concept of Ant Colony Optimization (ACO) was projected by Dorigo, which typically based on the comportment of ants which helps them finding the quickest path to a foodstuff source using pheromones. Ant colony optimization is also used to calculate the air route for helicopter submarine system [45]. These optimization techniques are also being used in hybridization methods in recent times. Like ACO is used [46] to boost the performance of type—2 fuzzy controller to handle the control of helicopter system. Also, the performances of Genetic Algorithm Optimization method and PSO method for control designing has been assessed [47]. Genetic Algorithm (GA) optimization technique was given by John Holland. This method is used to handle the routine problems in helicopter system [48]. GA is also used to handle the modelling technique of helicopter system in order to provide better stability in performance [49]. Another optimization technique called Grey Wolf Optimizer (GWO) is used to handle the controlling of helicopter system. GWO was introduced by Seyedali Mirjalili [50]. This optimization technique mimics the leadership hierarchy and hunting mechanism of grey wolves in nature. For path planning of combat aerial vehicles, GWO is used [51]. GWO integrated with model predictive control approach is used to handle the controller [52]. Grey Wolf Optimization is used to tune the parameters of PID controller [53]. Though, all the above mentioned algorithms have their different approach to handle the problem still these are just few of many other optimization techniques used in literature to tune the

parameters of PID controller in order to improve the output behavior of nonlinear systems.

Heuristic optimization search methods shows a malleable and well balanced machinery to boost the stability of nonlinear system under real time working conditions. Such a technique is Teaching Learning Based Optimization (TLBO), which was originally presented by R.V. Rao *et al* in 2011. Since then, it is being used in literature in diverse field of engineering which includes electrical, mechanical designs, thermal, manufacturing-civil-structural, computer science, electronic-VLSI, physics and chemical engineering [54–62]. It imitates influences of teacher on the outputs of learner in the class with teacher and learner being vital components of the algorithm [63]. The Teaching Learning Based Optimization (TLBO) method is considered better than other optimization methods discussed in literature due to two factors, which are:

(i) TLBO needs only two input parameters for the initialization of optimization process. These are maximum number of iterations & population size.
(ii) TLBO does not require any algorithmic specific parameter for the proper operation of optimization process, which helps providing the global solution more efficiently.

On comparing with other optimization techniques discussed in literature—PSO & GA requires three input parameters for the initialization of optimization process. While PSO, GA, ACO & GWO all of them requires algorithmic specific parameters for the execution of optimization process. Due to which even a minimum change in the initial data given by nonlinear system will change the overall behavior response dramatically. These loop holes in other techniques implemented on helicopter system, demands more work to be done for the better results during real time operation of helicopter model. Most studies available in literature are performed inside the laboratory on nonlinear systems. This will not hold true when real time setup of all the nonlinear systems will be operated with the outside laboratory conditions where, in case of helicopter system, real time environmental conditions like weather and air speed plays a vital role. The assumptions considered on the physical system and the authentication on the simulation platform sources a lack of practicality. This leads to the objectives of the presented work, which are:

(i) To develop a control scheme which will improve the behavior of helicopter system during practical situation like high speed wind and turbulence.
(ii) To introduce a controller to the modelling of helicopter system and then tune its parameters by using the improved control scheme.
(iii) To prove the superiority of improved method by using mathematical and graphical results obtained by real time simulation of helicopter model.

1.2 *Motivation and contributions*

In light of the above issues, this article presents an improved teaching learning based optimization (iTLBO) to overcome the drawbacks of other metaheuristic optimization in order to achieve the trajectory tracking of a helicopter under the unfavorable environment conditions. The unfavorable environment is generated by using high speed fans near the rotors of helicopter system in order to force the laboratory setup to experience the real time experience. The high speed of external air playing a critical role to cause instability to helicopter system. Inadequate tuning of the algorithm explicit parameter either upsurges the computational efforts or yields local optimum solutions. The proposed iTLBO approach overcomes the effects of parametric uncertainties in the helicopter system with the help of less computational struggle and great consistency when compared to other optimization techniques. The population based meta-heuristic algorithm possessed two different groups—Swarm Intelligence based and Evolutionary Based. PSO and TLBO falls under such categories only. PSO imitates the behavior of bird flock in search of food, is used to tune the parameters of PID and provide the stable and balanced working of pitch axis and the yaw axis of helicopter system in laboratory conditions [64, 65] but the implementation totally depends upon the algorithm specific parameters. The drawbacks of such conventional control techniques are overcome by making the formation of random number in proposed iTLBO approach even more dependent on student and teacher process instead of being critically random (which is discussed in detail in further sections of this paper). It is obvious that in the atmosphere of a class, a teacher trains students or learners such that the students will possess better knowledge in terms of marks, this same environment is replicated in iTLBO. Moreover, the proposed approach tackles the unfavorable environment conditions and produces better trajectory tracking of helicopter system. The major contributions of presented work can be summarized as:-

1. A dynamic fitness update is introduced to TLBO algorithm while tuning the controller parameter of a helicopter model.
2. The tuning of controller parameters is carried out by adding the PID controller in modelling of considered helicopter system.
3. Graphical Results obtained by proposed method is further compared with the results of LQR and core TLBO optimizaiton on MATLAB platform.
4. All the contributions are carried out by putting two high speed fans near the propeller of helicopter system to provide it some unfavorable real time like outside laboratory environment.

To achieve these set of objectives, the paper follows this sequence: the modelling conditions of two degree of freedom (2DOF) helicopter is discussed in section 2. In section

3, TLBO and iTLBO optimization techniques are explained which are further used on the MATLAB/SIMULINK model of the helicopter system to track down the trajectories of yaw and pitch angle. The experimental results of the controller design is communicated in section 4. Conclusion of the proposed work is done in section 5.

2. System description

In this section, two-degrees-of-freedom (2DOF) operation of a helicopter system is discussed. The two degree of freedom (2DOF) helicopter by Quanser [4], shown in figure 1, comprises of a yoke, a helicopter body, two DC motors—one for pitch and one for yaw, two propellers—front propeller and back propeller, two encoders—pitch encoder and yaw encoder. A fixed base platform is used to fasten the body of a helicopter system. The front propeller is being operated by the help of pitch DC motor, which control the altitude of a helicopter nose about pitch axis and hence pitch angle (θ) is controlled. The back propeller is being operated by the help of yaw DC motor, which operates the horizontal motions of the helicopter about the yaw axis, hence the yaw angle (Ψ) is controlled. The actual pitch angle and the actual yaw angle are being measured by using the pitch encoder and yaw encoder respectively. These measured angles in the form of signals are being transferred through a slip-ring. The proper modelling of considered 2DOF helicopter system (figure 1) is done based on the manipulations given in table 1.

One propeller motors is located at a distance given by r_p from the center and other propeller motor is located at a distance given by r_y from the center to determine the center of mass of the system. To continue with the modelling, the defining of transformation matrix of linearized model is important because the center of mass of model is dislocated from the central axis [66–68]. The required ABCD parameters of the nonlinear model is converted into Laplace Domain and hence used as the bench-problem on the MATLAB platform to perform the coding part of proposed iTLBO method.

3. Trajectory controller design

The controlling method for the cross coupled yaw and pitch axis of helicopter system is illustrated in figure 2. The parameters of controller G(s) are being handled using proposed algorithm. Two degree of freedom (2DOF) helicopter system is handled using optimization techniques like GA [69] and PSO [64]. For any problem having some large numbers of constraint, it provides a near optimum solution but it is matter-of-fact that a minimum change in algorithm parameter may change the overall effectiveness of algorithm. Using such algorithm during the unfavorable weather conditions, may affect the working of helicopter system at length. This attracts present work, to use an algorithm whose results on helicopter model is free from the parameters of algorithm. Using such an algorithm will definitely provide a more stable mechanism for cross coupled design of 2DOF helicopter as given in figure 2. In the figure, the controller 'G(s)' is optimized using proposed algorithm and implemented on error 'e(s)' signal, which is obtained by subtracting the pitch and yaw angle signal from the pitch and the yaw reference signal respectively. Further, these optimized parameters are handling the working of hardware

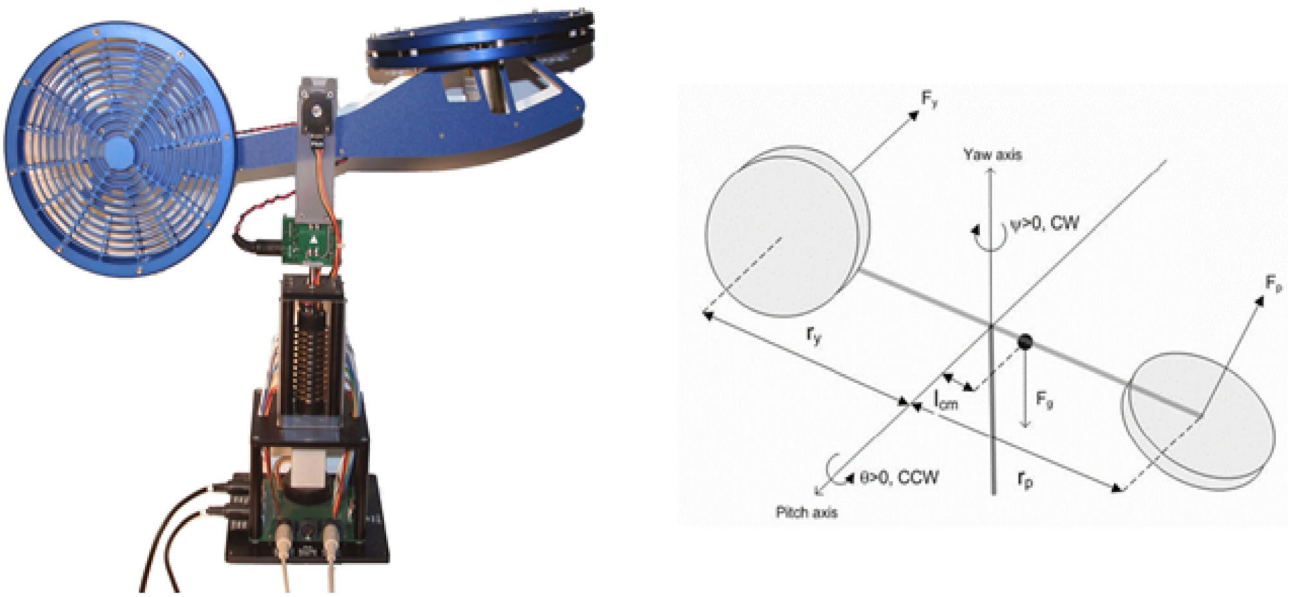


Figure 1. Laboratory setup and free body diagram of a two-DOF helicopter system [4].

Table 1. Conditions observed during operation of helicopter system.

S. No.	Conditions to be observed during operation on 2 DoF helicopter system
1	Pitch angle is denoted by θ. If θ = 0, it means the helicopter is horizontal
2	When the head of helicopter system moves upwards and downwards, the pitch angle increases and decreases respectively
3	If pitch angle moves positively i.e. θ(t) > 0, it means the helicopter body moving in vertical upward direction with respect to pitch axis.
4	If pitch angle moves negatively i.e. θ(t) < 0, it means the helicopter body moving in vertical downward direction with respect to pitch axis
5	Yaw angle is denoted by Ψ. If Ψ = 0, it means the body is neither moving clockwise nor moving anticlockwise
6	If the yaw angle increases i.e. Ψ(t) > 0, it means the body moves in horizontal direction and rotates counter clockwise with respect to yaw axis
7	If the yaw angle decreases i.e. Ψ(t) < 0, it means the body moves in horizontal direction and rotates clockwise with respect to yaw axis
8	The force produced due to pitch thrust is denoted by Fp. If Fp > 0, it means pitch angle increases i.e. θ > 0
9	The force produced due to yaw thrust is denoted by Fy. If Fy > 0, it means yaw angle increases i.e. Ψ > 0

system of helicopter model and the whole process continues until a more stable waveform for the pitch angle (θ) and the yaw angle (Ψ) is achieved. Control techniques designed for the trajectory tracking of this helicopter model during unfavorable weather conditions are discussed in following subsections.

3.1 *Teaching learning based optimization*

Two controlling parameters i.e. number of generations and population size are required for the implementation of TLBO algorithm. TLBO working is explained into two parts – "Teacher Phase" and "Learner Phase". Working of these two is explained herewith and further illustrated in figure 3 with the help of a flow chart.

3.1a *Teacher phase*: This phase of TLBO algorithm illustrates learning of students in a class through the teacher. In this phase, teacher aims to increase the mean results of the students for a particular subject taught by him/her depending on his/her capacity.

In such a way that at iteration i, 'n' be the number of learners and 'm' be number of subjects, the population size is 'k' and a particular subject 'j' be defined as:

$$k = 1, 2, 3, 4, 5, \ldots\ldots n \tag{1}$$

$$\text{and } j = 1, 2, 3, 4, 5, \ldots\ldots m \tag{2}$$

Making an allowance for all the subjects together the best result ($X_{j,best,i}$) obtained among the entire students be assumed as the result of best learner i.e. $X_{j,best,i}$ be the result of best student in subject j. The teacher is supposed to train the learners in order to have better results out of the students. Since, the teacher is assumed to be a highly learned person, so we will consider the identified best learner using the algorithm, as the teacher. The mean result of each subject (X_{mean}) and the best result of the teacher ($X_{j,best,i}$) are used in the following equation to calculate the difference mean ($Dmean_{j,k,i}$).

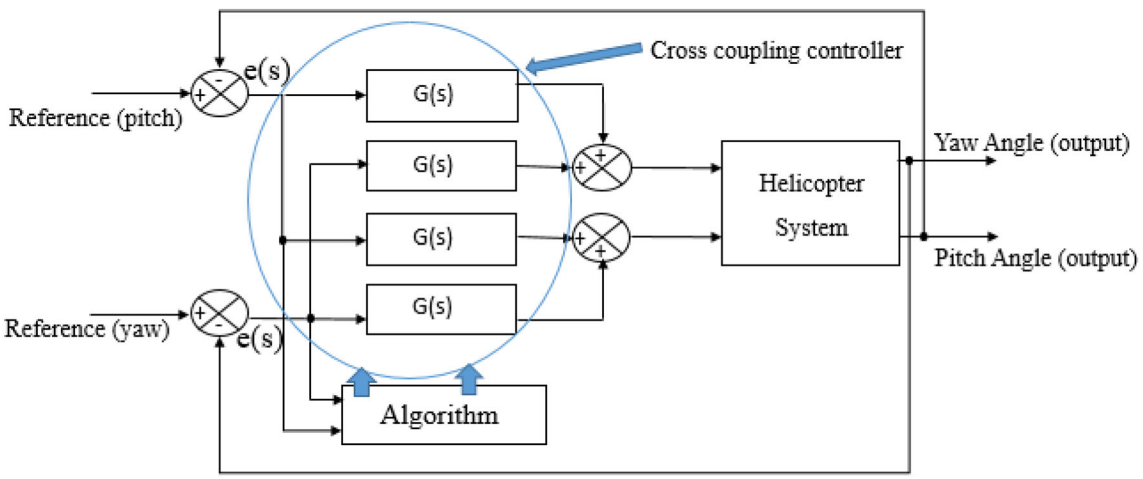


Figure 2. Block diagram of cross-coupled controller design for the trajectory control of helicopter system.

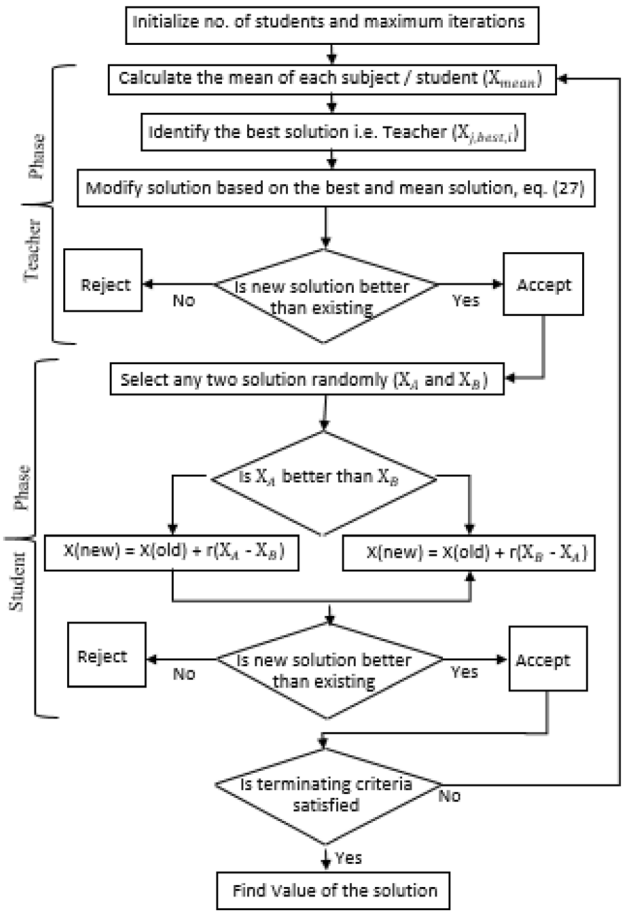


Figure 3. Flow Chart of teaching learning based algorithm.

$$\mathrm{Dmean}_{j,k,i} = \mathrm{r} * (\mathrm{X}_{best} - \mathrm{T}_f * \mathrm{X}_{mean}) \tag{3}$$

where T_f be the teaching factor which is the deciding factor for the value of mean supposed to be updated and r be any random number such that r will always be in the range [0,1]. T_f value is calculated randomly using the following equation:-

$$\mathrm{T}_f = \text{ round } (1 + \mathrm{r}) \tag{4}$$

The new solution ($\mathrm{X}_{j,k,i}(new)$) in teaching phase is updated based on $\mathrm{Dmean}_{j,k,i}$ using the following equation:

$$\mathrm{X}_{j,k,i}(new) = \mathrm{X}_{j,k,i}(old) + \mathrm{Dmean}_{j,k,i} \tag{5}$$

$\mathrm{X}_{j,k,i}(new)$ will be accepted provided it gives improved function value after applying the greedy solution. In the greedy solution demonstration new solution is accepted or rejected based on the fact that if it gives improved function value or not. At the end of this teacher phase, all these improved function values becomes input to the learner phase, which shows the dependency of learner phase on the teacher phase.

3.1b *Student/learner phase*: In this part of the algorithm, the learners or students improves their knowledge through interaction among each other all by themselves. A student will randomly choose the other student (called partner student) and boost his/her learning. Students learns if the other learner has more knowledge. For some population size of 'n' number of learners/students, A and B are the two randomly selected learners such that $\mathrm{X}_{A,i}(new) and \mathrm{X}_{B,i}(new)$ are the updated solution for $\mathrm{X}_{A,i}(old) and \mathrm{X}_{B,i}(old)$ which are defined in equations (6) and (7).

$$\mathrm{X}_{A,i}(new) = \mathrm{X}_{A,i}(old) + \text{ r } [\mathrm{X}_{A,i}(old) - \mathrm{X}_{B,i}(old)] \tag{6}$$

Equation (6) is valid for solving maximization problems if updated function value for the current student is more than that of the partner students and valid for solving minimization problems if updated function value for the current student is less than that of the partner student.

$$\mathrm{X}_{A,i}(new) = \mathrm{X}_{A,i}(old) + \text{ r } [\mathrm{X}_{B,i}(old) - \mathrm{X}_{A,i}(old)] \tag{7}$$

Equation (7) is valid for solving maximization problems if updated function value for the current student is less than that of the partner students and valid for solving minimization problems if updated function value for the current student is additional than that of the partner student.

3.2 *Improved TLBO (iTLBO)*

3.2a *Updating process for teaching factor*: Value of teaching factor given by Eq. (4) is not considered as the parameter of the algorithm as it is decided randomly by the algorithm. After performing many operations on benchmark function, it was observed that algorithm performs faster if the range of T_f is between 1 and 2. Still selection of teaching factor is totally dependent on the randomized selection of the algorithm only. Moreover, algorithm gives acceptable value if this value is taken either 1 or 2. All these cause lack of data dependency in the algorithm. For the sake of simplification of implementation, the value of teaching factor is preferred to take either 1 or 2 by rounding up criteria given in (4). But, if instead of taking the shortcut, let us put some light on this randomized behavior and to eliminate this randomized selection make the teaching factor dependent on data used by algorithm.

To fulfill this desire, an improved factor (i_f) is introduced in order to eliminate the random value (r) command from the TLBO algorithm and the proposed value of this i_f is calculated during the optimization process as:

$$i_f = \frac{1}{1 + (\mathrm{X}_{j,best,i} - \mathrm{X}_{mean})} \tag{8}$$

In TLBO, teaching factor is preferred as a round off number, this also indicates some room for improvement. By eliminating this issue too from the algorithm the equation of teaching factor updated as:-

$$T_f = \frac{2 + (X_{j,best,i} - X_{mean})}{1 + (X_{j,best,i} - X_{mean})} \tag{9}$$

3.2b *Updated data-driven solution*: The new solution during teaching phase is updated using Eq. (5) and during learner phase is updated using Eqs. (6–7). After updating the teaching factor in Eq. (9), the value given by Eq. (5–7) will also get updated and hence the new solution in teacher phase will be referred as:

$$X_{j,k,i}(new) = X_{j,k,i}(old) + \left\{\frac{1}{1 + (X_{j,\ best,i} - X_{mean})}\right\} * \left\{X_{best} - \frac{2 + (X_{j,\ best,i} - X_{mean})}{1 + (X_{j,\ best,i} - X_{mean})} X_{mean}\right\} \tag{10}$$

Equation (10) is the proposed result of new solution generated by teacher phase in iTLBO algorithm. This also shows positive effect on the student phase of iTLBO, as the new solution in the generation of TLBO algorithm in learning phase also dependent on random behavior. To improve this, Eq. (8) is integrated with Eqs. (6–7) and gives us the improved new solutions for learner phase as:

$$X_{A,i}(new) = X_{A,i}(old) + \left\{\frac{1}{1 + (X_{j,best,i} - X_{mean})}\right\} * \chi_1 \tag{11}$$

where $\chi_1 = \{X_{A,i}(old) - X_{B,i}(old)\}$ and Eq. (11) is valid for solving maximization problems if χ_1 is positive and valid for solving minimization problems if χ_1 is negative.

$$X_{A,i}(new) = X_{A,i}(old) + \left\{\frac{1}{1 + (X_{j,best,i} - X_{mean})}\right\} * \chi_2 \tag{12}$$

where $\chi_2 = [X_{B,i}(old)—X_{A,i}(old)]$ and Eq. (12) is valid for solving maximization problems if χ_2 is positive and valid for solving minimization problems if χ_2 is negative.

Eliminating the randomized behavior in the algorithm makes the iTLBO more reliable and accurate. The data driven updated value given stable input and quick response with both the test benchmark functions and two degree of freedom helicopter system. These are discussed in the next section of this work.

3.2c *Testing iTLBO on benchmark functions*: To check the effectiveness and superiority of proposed algorithm two experiments are conducted. Experiments are based on benchmark test functions. Two different constrained benchmark functions, one is with linear characteristic and other is with nonlinear characteristic of objective functions are considered. These are discussed in this subsection further.

- Benchmark function 1

A maximization nonlinear problem having one nonlinear equality constraint and ten design variables is shown in [Eqs. (13), (14)]. At the optimum point there is one active constraint given the global maximum at $y^* = (1/\sqrt{h}, 1/\sqrt{h}, 1/\sqrt{h}, \ldots\ldots..)$ having one objective function value $f(y^*) = 1$. The inequality constraint shown as $|h| \leq \varepsilon$, which is converted from equality constraint, where $\varepsilon = 0.001$.

$$\text{Max } f(y) = (\sqrt{h})^h \prod_{i=1}^{h} y_i \tag{13}$$

$$\text{S.T. } g(y) = \sum_{i=1}^{4} y_i^2 - 1 = 0 \tag{14}$$

where $h = 10$ and $0 \leq y_i \leq 10$ ($i = 1, 2, 3, \ldots\ldots..h$)

This problem was solved using particle evolutionary swarm optimization (PESO) [70], multi-membered evolutionary strategy (M-ES), cultural differential evolution (CDE), Co-evolutionary differential evolution (CoDE), artificial bee colony (ABC) [74], teaching learning based optimization (TLBO) [63]. Setting the maximum number of generations up to 2000 and size of population as 50 in improved TLBO algorithm, with best result for problem is calculated and shown in table 2. iTLBO requires less function evaluations to prove it's superiority over the other considered optimization methods.

- Benchmark function 2

A minimization linear problem having three linear inequality constraints, three nonlinear inequality constraints and eight design variables as shown in Eqs. (15–21). At the optimum point there are three active constraint given the optimum solution at y^* = (286.165, 579.3066, 5109.9707, 1359.9709, 5109.9707, 295.601, 182.0177, 217.982, 395.6012) having one objective function value as $f(y^*)$ = 7049.248021. The inequality constraint shown as $|h| \leq \varepsilon$,

Table 2. Different optimization results on benchmark function 1.

Optimization	Best solution	Function evaluations
PESO [70]	0.9939	350, 000
M-ES [76]	1	240, 000
CDE [75]	0.9954	100,100
ABC [74]	1	240, 000
TLBO [63]	1	100, 000
Proposed	1	90, 000

which is converted from equality constraint, where ε = 0.000001.

$$\text{Min } f(y) = y_1 + y_2 + y_3 \tag{15}$$

$$\text{S.T.} f_1(y) = -1 + 0.0025(y_4 + y_6) \leq 0 \tag{16}$$

$$f_2(y) = -1 + 0.0025(y_5 + y_7 - y_4) \leq 0 \tag{17}$$

$$f_3(y) = -1 + 0.01(y_8 - y_5) \leq 0 \tag{18}$$

$$f_4(y) = -y_1 y_6 + 833.33252 y_4 + 100 y_1 - 83333.333 \leq 0 \tag{19}$$

$$f_5(y) = -y_2 y_7 + 1250 y_5 + y_2 y_4 - 1250 y_4 \leq 0 \tag{20}$$

$$f_6(y) = -y_3 y_8 + 1250000 + y_3 y_5 - 2500 y_5 \leq 0 \tag{21}$$

where $-100 \leq y_1 \leq 10000$
$-1000 \leq y_i \leq 10000$ (i = 2, 3)
$-100 \leq y_i \leq 10000$ (i = 4, 5, 6, 7, 8)

This problem was solved using particle swarm optimization (PSO) [70], artificial bee colony (ABC) [74], teaching learning based optimization (TLBO) [63]. Setting the maximum number of generations up to 2000 and size of population as 50 in iTLBO algorithm, with best result is evaluated and shown in table 3. Improved TLBO requires less function evaluations to solve the benchmark function problem.

3.2d *Describing function for helicopter model*. Describing function for two degree of freedom helicopter is formulated on the MATLAB script. For this, the properties of higher order transfer function is used. A second order helicopter model is designed after implementing RGA method presented by Bristol [71]. Transfer function derived in this process is used as plant and the controller used in the configuration as G(s) is PID controller which is optimized using the algorithm iTLBO and arranged exactly as shown in figure 1. The step information characteristics of the system hence develop i.e., rise time (γ_1), settling time (γ_2), peak overshoot (γ_3) and peak undershoot (γ_4) are calculated using Simulink script. To have the stable output for this setup, experimental values of above mentioned four characteristics of nonlinear model must be as minimum as possible during simulation. Hence to develop the cost function used in minimization process, a combination of code and Simulink is processed to calculate the cost function J(x) used in this process as:

$$J(x) = [\min(\gamma_1) + \min(\gamma_2) + \min(\gamma_3) + \min(\gamma_4)] \tag{22}$$

This cost function is used on the experimental setup during simulation and this helped in extracting out the optimal values of PID parameters.

Table 3. Different optimization results on benchmark function 2.

Optimization	Best solution	Function evaluations
PESO [60]	7049.38	350, 000
M-ES [76]	7051.903	240, 000
CDE [75]	7049.248	100,100
ABC [74]	7053.904	240, 000
TLBO [53]	7049.24	100, 000
Proposed	7049.24	90, 000

3.2e *Implementing iTLBO for PID parameter optimization using describing function*. For implementing proportional-integral-derivative controller as the cross-coupled controller (as shown in figure 1) on helicopter model, MATLAB SIMULINK platform is used. The error signal e(s) (in figure 1), is given as:

$$e(s) = X_d(s) - X(s) \tag{23}$$

where $X_d(s)$ is the desired set point for the two control axis i.e. pitch axis and yaw axis and $X(s)$ is the actual process value. PID controller is implemented on this error signal. The set points, error signals e(s) and the process values are calculated separately for both the axis i.e. yaw axis and pitch axis and different PID controllers are used for both the axis. To handle the coupling, four PID controllers are used as shown in figure 2. Same iTLBO algorithm is used to optimize the parameters of all four PID controllers. To implement the PID controller in frequency domain on helicopter system, where $k_{d,heli}$, $k_{i,heli}$ and $k_{p,heli}$ are derivative, integral and proportional gain used to control the helicopter system, for which the initial operating points are given as:

$$\beta_d(s) = \left(k_{d,heli}.s + \frac{k_{i,heli}}{s} + k_{p,heli}\right)[X_d(s) - X(s)] \tag{24}$$

Initial parameters of PID and the boundary conditions are calculated following the explanation by George Ellis [72]. Then controller gains are calculated after each iteration of iTLBO as shown in section 3.2.1 and 3.2.2 and followed by describing function given in Eq. (23). To attain optimized parameters of PID controller on MATLAB script, some steps are followed in each iteration. These steps are:

i. Define Three dimensions of the problem as $k_{d,heli}$, $k_{i,heli}$ and $k_{p,heli}$ along with the upper and lower bounds.
ii. Define Number of learners and size of maximum iteration. Then in teacher phase calculate mean of each subject and identify teacher.
iii. Update mean of each subject and teacher by solving these equations in this sequence—Eqs. (8), (3), (9) and (5).
iv. Use the updated value to check the new solution using describing function in Eq. (22).

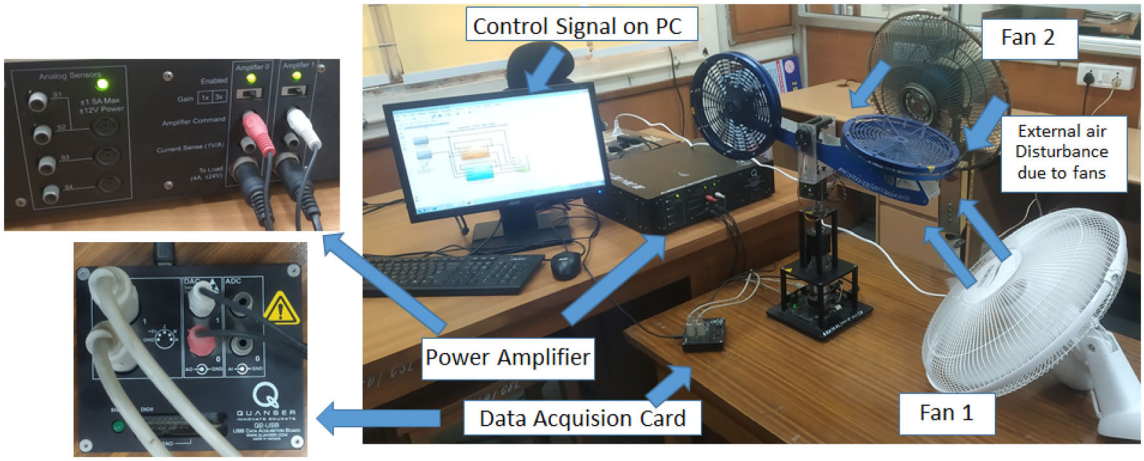


Figure 4. Experimental setup of 2 DOF helicopter system under external disturbance created in Laboratory.

v. Two random subjects/students will be selected and compared in student phase as given in Eq. (10).
vi. If one student is better than the other new solution will be updated according to Eqs. (11) and (12).
vii. Validate the new solution using describing function in Eq. (5) and if it is found better than the existing then accept it.
viii. New fitness value will be used for next iteration and hence the current iteration terminates.
ix. As the mean of student and position of teacher gets updated after each iteration, the value of three dimensions calculated using (5).
x. Steps (iii)–(viii) are repeated until the maximum iterations reached and optimized value of three dimensions i.e. optimized value of PID controller is obtained.
xi. Put these optimized value of $k_{d,heli}$, $k_{i,heli}$ and $k_{p,heli}$ in PID controller [G(s)] as shown figure 2 to perform the control of yaw and pitch axis of helicopter system during experimental analysis.

4. Experimental analysis

The modelling of two degree of freedom (2DOF) helicopter system as shown in section 2, is further realized on simulation platform using Quanser hardware system [4], as depicted in figure 4. This system aids in the development of real time control of pitch angle and yaw angle of a 2DOF helicopter. It consist of a power amplifier (VoltPAQ-X2) [73], Q-2 USB data acquisition (DAQ) card, actuator sensor and two dc motors. The Pittman model 9234 pitch motor is used having a resistance of 0.83 ohms and a torque constant of 0.0182 Nm/A. The Faulhaber series 2842−006 C yaw motor is used having a terminal resistance of 1.6 ohms and a torque constant of 0.0109 Nm/A. Two propellers [5], Graupner 15 cm having a thrust force constant 0.43 N/V and a Graupner 20 cm having a thrust force constant 0.104 N/V are used for yaw and pitch respectively. Two encoders having a fine resolution of 4096 counts per revolution and a fine resolution of 8192 counts per revolution are used as pitch and yaw encoder respectively. These encoders plays a vital role in providing the position feedback when they are connected with DAQ board of the helicopter setup. The controller developed in section 3 is implemented on the model of helicopter system as explained in section 2 and the whole real time operation of 2 DOF helicopter system is analyzed for the trajectory tracking of pitch angle and yaw angle.

The performance of proposed controller in terms of improved pitch angle (θ) and yaw angle (Ψ), under the presence of external disturbance (which is created using two Fans, as shown in figure 4) is assessed by the graphs in which variation of these angles is recorded over the time. While placing the fans at different locations near the body of helicopter model, it is observed that orientation of pitch axis changes depending upon the locations of the fans but the pitch and yaw angles gets similar error irrespective of the location of the fans. This is due to the fact that the two propellers of the helicopter model are connected with a fixed base. Hence, only pitch and yaw angles are sensitive to the external disturbances. Further, pitch and yaw trajectories traced using the proposed approach under the influence of external disturbance is compared with other controller's response under the same external disturbance in figures 5 and 6.

The response is recorded with square reference signal using different controllers. Figures 5a and 6a shows the trajectory of pitch axis and yaw axis respectively, obtained by using the default LQR controller. Figures 5b and 6b shows the trajectory of pitch axis and yaw axis respectively, obtained by using the TLBO optimization method on PID controller. Figures 5c and 6c shows the trajectory of pitch

axis and yaw axis respectively, obtained by using the iTLBO controller. The depiction clearly shows that the proposed iTLBO using PID, eliminates the disturbance to almost zero, in real time while improving the peak overshoot, peak time and settling time of the helicopter system on MATLAB/ Simulation platform.

The superiority of proposed algorithm is further discussed in table 4, in which the response parameters i.e. settling time, peak time and peak overshoot obtained from the MATLAB Simulation analysis of the results obtained in figures 5 and 6 are discussed. It is observed that peak overshoot of LQR is high which causes huge oscillations and made it difficult to handle the external disturbances. Some results available in literature are also compared in table 4, which are further used to prove the superiority of proposed algorithm over the results available in literature. Afterwards, by implementing TLBO, peak overshoot is reduced but the external disturbance is still an issue, which creates the need of improved controller. The results of iTLBO-PID controller, shows that not only the peak overshoot is reduced but the oscillations due to external disturbance is also reduced to almost zero and excellent balance of both yaw propeller and pitch propeller is achieved using the proposed method.

Moreover, the value of absolute error for pitch angle is calculated from running simulation model of two degree of freedom helicopter system using Eq. (23) and may be re-written:-

$$\text{Error}[e(s)] = \text{Reference angle} - \text{output angle} \tag{25}$$

Convergence of absolute error graph is obtained using Eq. (25) by applying three different controlling techniques i.e. LQR, TLBO & proposed iTLBO. Using simulation shown in figure 2, signal e(s) is sent to workspace of MATLAB code and then error value is traced using three different controlling techniques. Figure 7 shows the path followed by error shown by pitch response of helicopter system when controlled using three controlling approaches i.e. LQR, TLBO & proposed iTLBO. As seen clearly in figure 7, the error is converged to nearly zero after 7.8 s while using LQR controller, got conversed after 4.1

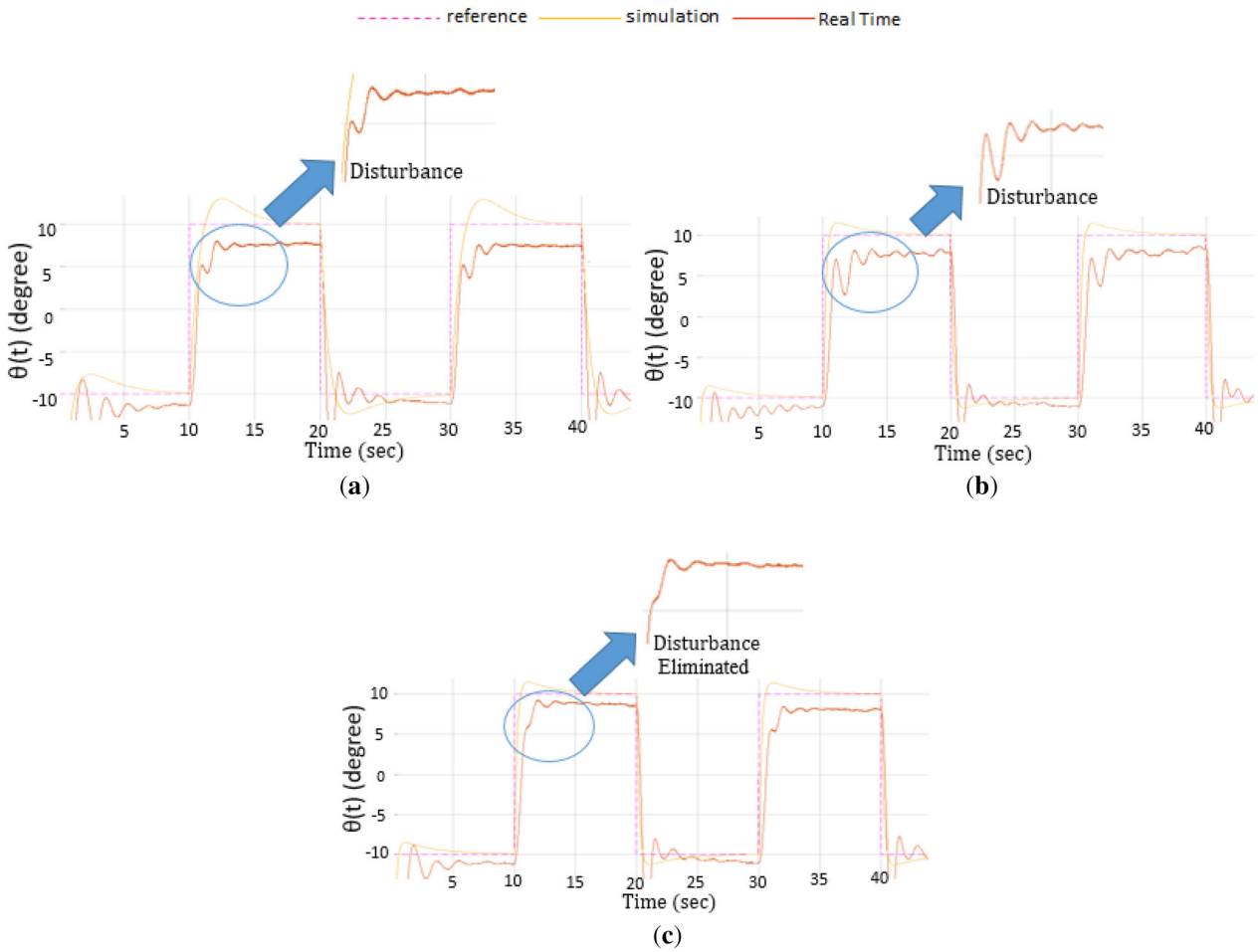


Figure 5. Pitch angle trajectory with the square reference signal during external disturbance (**a**) default Quanser [5] LQR controller (**b**) TLBO [63] algorithm (**c**) proposed iTLBO algorithm.

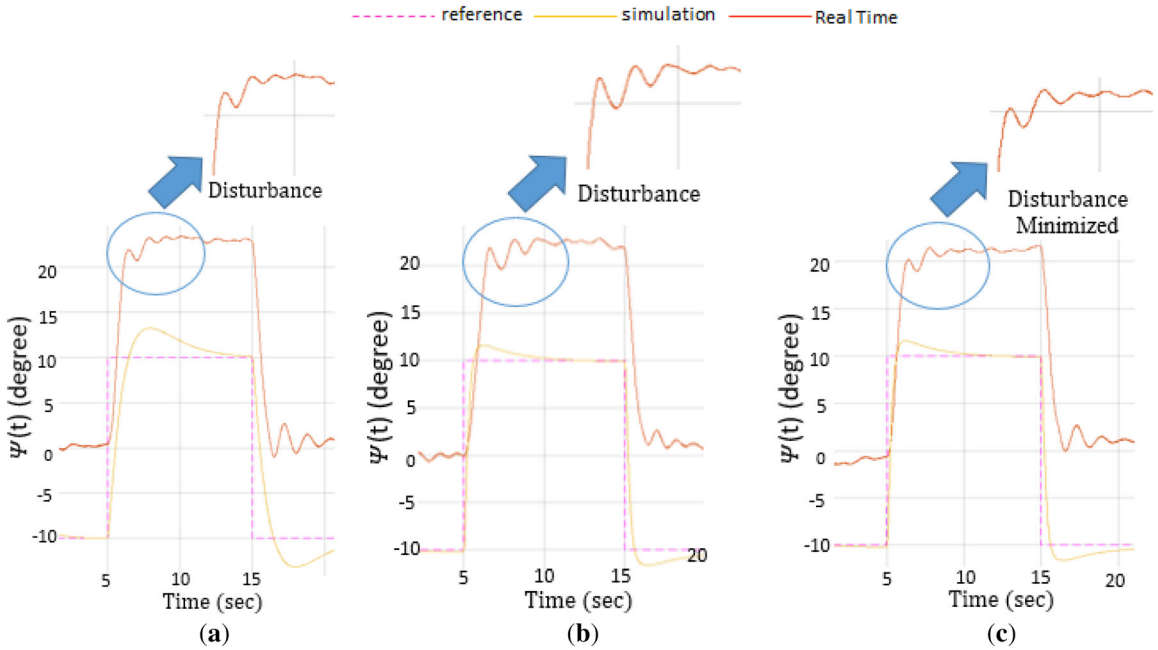


Figure 6. Yaw angle trajectory with the square reference signal during external disturbance using (**a**) default Quanser [5] LQR controller (**b**) TLBO [63] algorithm (**c**) proposed iTLBO algorithm.

Table 4. Time response parameters for various controllers during simulation.

	Pitch response			Yaw response		
Method	t_s (sec)	t_p (sec)	M_p (%)	t_s (sec)	t_p (sec)	M_p (%)
LQR	7.81	2.8	8.99	5.12	2.1	18.3
TLBO	4.11	1.2	4.30	4.01	0.95	8.19
proposed	2.83	1.0	0.92	2.67	0.79	6.23

seconds when TLBO optimization method is used and got conversed to nearly zero after 2.9 seconds when proposed iTLBO optimization method is used on the pitch response of helicopter controller.

iTLBO proves its dominance when compared to other optimization techniques as the results obtained using proposed algorithm is compared with other optimization techniques developed in literature are discussed in table 5. This table compares different optimization techniques available in literature with the proposed results. Comparison is based on four different factors which are—Response of control action, sensitivity to disturbance, external disturbance & time for stabilization. Let's discuss these four factors one by one. Response of control action refers to the dead time and peak time taken by the pitch controller from initial position.

If peak time is more than 2 s, the controller response will be considered slow. If peak time is in the range of 1–2 s, the controller action will be considered Normal. If the control action takes less than or equals to 1 second to give peak time, then fast or immediate action of controller is observed. Sensitivity to disturbance refers to the peak overshoot given by the pitch trajectory of helicopter in simulation results during considered controller. If peak overshoot is more than 7%, the response given by controller will be considered poor or highly sensitive to disturbances. If the peak overshoot is in the range of (4–7)%, the controller action is moderately sensitive to the disturbance in the system. If the controller response showing peak overshoot in the range (2–4)%, the controller action is less sensitive to the disturbance. But, if the controller response on simulation platform is showing less than 2% of peak overshoot then the considered controller is least or very low sensitive to the disturbances.

External disturbance is referred to the two external fans used to provide turbulence like conditions by giving high speed air environment to the propellers of helicopter system. Time for stabilization refers to the settling time taken by the pitch response on simulation platform of helicopter system. If the settling time is more than 6 s, it means

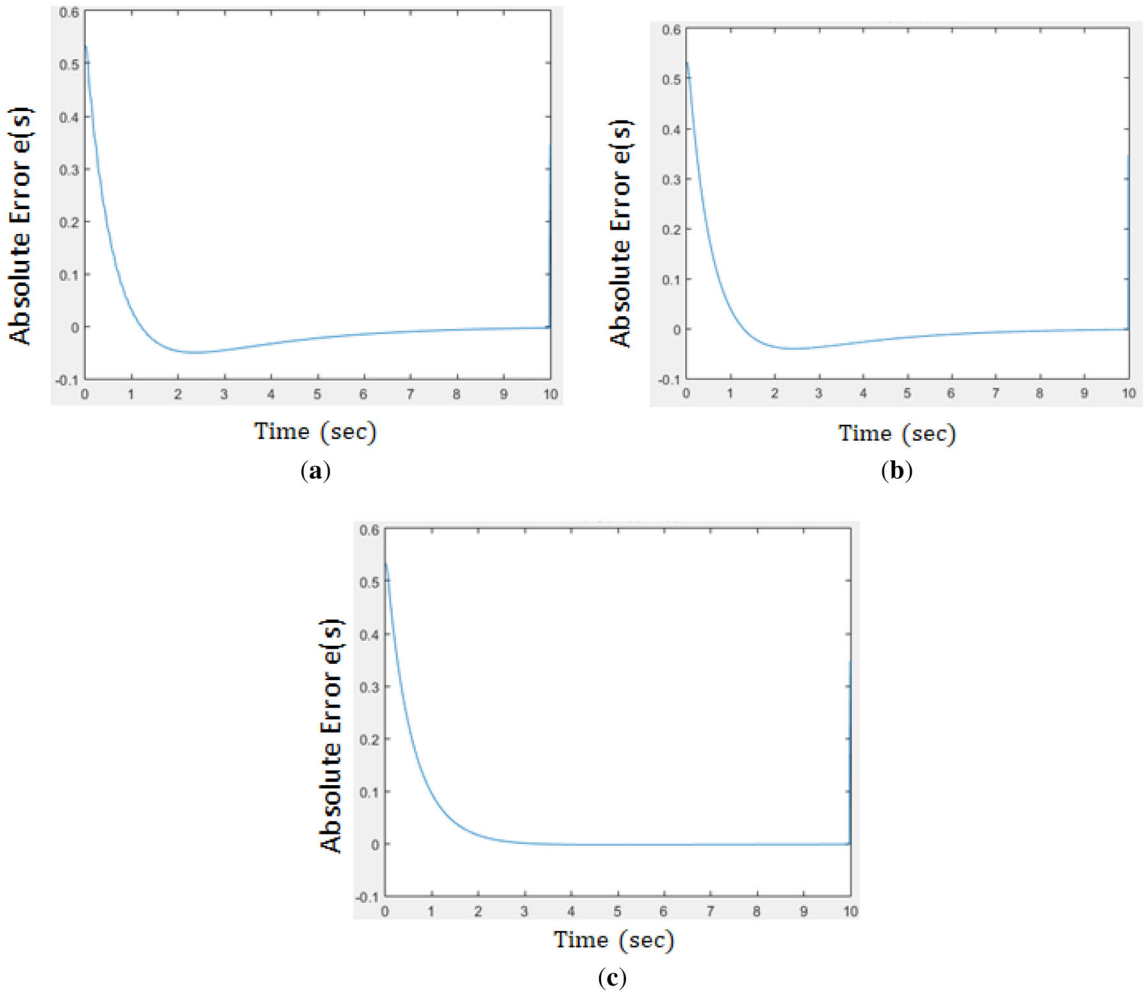


Figure 7. Absolute Error vs Time graph obtained by (**a**) default LQR controller [5] (**b**) TLBO algorithm [63] (**c**) proposed iTLBO algorithm.

Table 5. Comparison of performance Index of proposed iTLBO using different methods available in literature.

Method	Response of control action	Sensitivity to disturbance	External disturbance	Time for stabilization
LQR [5]	Slow	High	Cannot handle	High
PSO [42, 43, 64]	Normal	Moderate	–	Less
ACO [45, 46]	Slow	High	–	Moderate
GA [48, 49, 69]	Normal	Moderate	–	Less
Fault classification based PID [75]	Immediate	Low	–	Moderate
GWO [51]	Normal	Low	–	–
Proposed	Immediate	Very low	Handled smoothly	Very low

considered controller is taking high time to reach stability. If the settling time is in the range of (4–6) seconds, it means the considered controller is taking moderate time to reach stability. If the settling time is in the range of (3–4) seconds, then the system is taking less time to attain stability. But, if the settling time is less than 3 s, it means the time taken by system for stabilization is very low.

5. Conclusion

In this work, pitch axis and yaw axis trajectories of a two degree of freedom helicopter system are controlled under the presence of external disturbance by implementing an improved teaching learning based optimization (iTLBO) algorithm. In the process, tuning of gain parameters of

proportional integral derivative (PID) controller is performed. Comparing to other optimization techniques in literature, the performance of iTLBO is more satisfactory on the linear and nonlinear benchmark models. The results obtained after experimental analysis of helicopter system proved that the developed approach turned out to be superior in eliminating the external disturbance. All the results are validated on simulation platform and gives noteworthy performance within the framework of traditional controller structure. Results obtained are authorized using time response analysis on the MATLAB/Simulation platform. To achieve these results, the describing function is derived using time response parameters only. The proposed controller is proved adaptable and adequate which gives satisfactory performance on helicopter system when tested under the presence of external air disturbance in the laboratory. iTLBO does not require any algorithmic specific parameters and eliminates randomness from TLBO algorithm, as the result control action is more reliable and achieves remarkable performance while handling yaw axis and pitch axis of helicopter system.

It gives a hope for more future work may be done using such improved algorithm on the benchmark functions and nonlinear models like helicopter system. Moreover, control techniques like fuzzy control, neural control, model predictive control and wavelet methods may be used in hybridization along with optimization techniques like iTLBO. Which may be used to control the trajectories of helicopter system. Just like the air pressure by external disturbance is handled in the proposed work, more research may be focused on the vibrations experienced by helicopter system in different weather conditions. This will further help the system to show robust performance and to be more susceptible to withstand disturbances.

Declarations

Conflict of interest Author(s) declare(s) no conflict of interest to disclose to any known person or agency.

References

[1] Najm A A and Ibraheem I K 2019 Nonlinear PID controller design for a 6-DOF UAV quadrotor system. *Eng. Sci. Technol. Int. J.* 22(4): 1087–1097

[2] Padfield G D 2008 Book on helicopter flight dynamics. Wiley, Hoboken

[3] Vilchis J A, Brogliato B, Dzul A and Lozano R 2003 Nonlinear modelling and control of helicopters. *Automatica* 39(9): 1583–1596

[4] Quanser Inc 2010 Quanser 2-DOF Helicopter Manual. Tech. rep. Quanser

[5] Apkarian J, Levis M and Fulford C 2012 User manual of 2-DOF helicopter experiment setup and configuration. Ontario, Canada: Quanser

[6] Halsey K M and Glover K 2005 Analysis and synthesis of nested feedback systems. *IEEE Trans. Autom. Cont.* 50(07): 984–996

[7] Hernandez-Gonzalez M, Alanis A Y and Vargas 2012 Decentralized discrete-time neural control for a quanser 2-DOF helicopter. *Appl. Soft Comput.* 12: 2462–2469

[8] Samadi B and Rodrigues L 2014 A sum of squares approach to backstepping controller synthesis for piecewise affine and polynomial systems. *Int. J. Robust Nonlinear Contr.* 24: 2365–2387

[9] Nuthi P and Subbarao K 2015 Experimental verification of linear and adaptive control techniques for a 2-DOF helicopter. *J. Dyn. Syst. Meas. Contr.* 137: 64501–64506

[10] Khayati K 2015 Multivariable adaptive sliding mode observer based control for mechanical systems. *Can. J. Electr. Comput. Eng.* 38(3): 253–265

[11] Liu Z and Shi H 2010 Control strategy design based on fuzzy logic and LQR for 3-DOF helicopter model. In: *International Conference on Intelligent Control and Information Processing*. IEEE, pp. 262–266

[12] Ríos H, Rosales A, Ferreira A and Dávilay A 2010 Robust regulation for a 3-DOF helicopter via sliding-modes control and observation techniques. In: *Proceedings of the 2010 American Control Conference*. IEEE. pp. 4427–4432

[13] Witt J, Boonto S and Werner H 2007 Approximate model predictive control of a 3-DOF helicopter. In: *46th IEEE Conference on Decision and Control.* IEEE. pp. 4501–4506

[14] Yu Y, Lu G and Zhong YS 2012 Robust decentralized control for a 3-DOF helicopter. In: *Proceedings of the 31st Chinese Control Conference*. IEEE. pp. 2708–2714

[15] Zhu X and Li D 2021 Robust attitude control of a 3-DOF helicopter considering actuator saturation. *Mech. Syst. Signal Process.* 149: 107209

[16] Zhu X and Li D 2021 Robust fault estimation for a 3-DOF helicopter considering actuator saturation. *Mech. Syst. Signal Process.* 155: 107624

[17] Wang X Y, Yu X and Li S H 2019 Composite block backstepping trajectory tracking control for disturbed unmanned helicopters. *Aerosp. Sci. Technol.* 105(6): 386–398

[18] Amin R and Li A 2017 Modelling and robust attitude trajectory tracking control of 3-DOF four rotor hover vehicle. *Aircr. Eng. Aerosp. Technol.* 89(1): 87–98

[19] De Jager 1998 Control of under actuated systems: a monocycle example. *IFAC Proc.* 31: 61–66

[20] Olivares M and Albertos P 2013 On the linear control of underactuated systems: the fywheel inverted pendulum. In: *10th IEEE International Conference on Control and Automation (ICCA)*. IEEE. pp 27–32

[21] Aguilar-Avelar C and Valenzuela Moreno J 2015 A composite controller for trajectory tracking applied to the Furuta pendulum. *ISA Trans.* 57: 286–294

[22] Chang D E 2008 Some results on stabilizability of controlled lagrangian systems by energy shaping. *IFAC Proc.* 41: 3161–3166

[23] Choukchou-Braham A, Cherki B, Djemaï M and Busawon K 2014 Underactuated mechanical systems from the Lagrangian formalism. Analysis and control of underactuated mechanical systems. Springer, Cham, pp 15–33

[24] Rudra S, Barai R K and Maitra M 2017 Block backstepping control of the underactuated mechanical systems. Block

backstepping design of nonlinear state feedback control law for underactuated mechanical systems. Springer, Singapore, pp 31–52

[25] Li X 2020 Robot target localization and interactive multi-mode motion trajectory tracking based on adaptive iterative learning. *J. Ambient Intell. Humaniz. Comput.* 11(2): 6271–6282

[26] Sutharsan S, Kirubarajan T, Lang T and Mcdonald M 2012 An optimization-based parallel particle flter for multitarget tracking. *IEEE Trans. Aerosp. Electron. Syst.* 48: 1601–1618

[27] Tarczewski T and Grzesiak L M 2015 Constrained state feedback speed control of PMSM based on model predictive approach. *IEEE Trans. Industr. Electron.* 63(6): 3867–3875

[28] Lou G, Gu W, Sheng W, Song X and Gao F 2018 Distributed model predictive secondary voltage control of islanded microgrids with feedback linearization. *IEEE Access* 6: 50169–50178

[29] Kiefer T, Graichen K and Kugi A 2009 Trajectory tracking of a 3DOF laboratory helicopter under input and state constraints. *IEEE Trans. Control Syst. Technol.* 18(4): 944–952

[30] Jahed M and Farrokhi M 2013 Robust adaptive fuzzy control of twin rotor MIMO system. *Soft Comput.* 17(10): 1847–1860

[31] Mondal S and Mahanta C 2012 Adaptive second-order sliding mode controller for a twin rotor multi-input-multi-output system. *IET Control Theory Appl.* 6(14): 2157–2167

[32] Gao J, Wu P, Li T and Proctor A 2017 Optimization-based model reference adaptive control for dynamic positioning of a fully actuated underwater vehicle. *Nonlinear Dyn.* 87(4): 2611–2623

[33] Ghabraei S, Moradi H and Vossoughi G 2015 Multivariable robust adaptive sliding mode control of an industrial boiler–turbine in the presence of modeling imprecisions and external disturbances: a comparison with type-I servo controller. *ISA Trans.* 58: 398–408

[34] Wang Y, Jin Q and Zhang R 2017 Improved fuzzy PID controller design using predictive functional control structure. *ISA Trans* 71: 354–363

[35] Wang J, Zhu Y and Qi R 2020 Adaptive PID control of multi DOF industrial robot based on neural network. *J. Ambient Intell. Humaniz. Comput.* 11: 6249–6260

[36] Abdo M M, Vali A R, Toloei A R and Arvan M R 2014 Stabilization loop of a two axes gimbal system using self-tuning PID type fuzzy controller. *ISA Trans.* 53: 591–602

[37] Zhang J, Zhuang J, Du and Wang S 2009 Self-organizing genetic algorithm based tuning of PID controllers. *Inf. Sci. (Ny)* 179: 1007–1018

[38] Feng H, Yin C-B and Weng W 2018 Robotic excavator trajectory control using an improved GA based PID controller. *Mech. Syst. Signal Process.* 105: 153–168

[39] Hou Y Y 2017 Design and implementation of EP-based PID controller for chaos synchronization of Rikitake circuit systems. *ISA Trans.* 70: 260–268

[40] Fleming P J and Purshouse R C 2002 Evolutionary algorithms in control systems engineering: a survey. *Control Eng.* 10(11): 1223–1241

[41] Solihin M I and Akmeliawati R 2010 Particle swam optimization for stabilizing controller of a self-erecting linear inverted pendulum. *Int. J. Electr. Electron. Syst. Res.* 3: 410–415

[42] Zhiqiang L, Xi Y, Qiana T and Linwen Y 2022 Research on attitude control strategy of unmanned helicopter based on hybrid particle swarm optimization algorithm. *J. Phys. Conf. Ser.* 2290: 012072

[43] Yu GR and Hsieh PH 2019 Optimal design of helicopter control systems using particle swarm optimization. In: *IEEE International Conference on Industrial Cyber Physical Systems* (ICPS), 346–351

[44] Chiha I, Liouane N and Borne P 2012 Tuning PID controller using multiobjective ant colony optimization. *Appl. Comput. Intell. Soft Comput.* 2012: 1–7

[45] Li Lin 2020 Application of ant colony algorithm to air route planning in helicopter submarine searching. *J. Phys. Conf. Ser.* 1650: 032004

[46] Amar, Achour Z and Hamerlain M 2014 Ant colony optimization of type-2 fuzzy helicopter controller. In: *IEEE International Conference on Robotics and Biomimetics*. IEEE ROBIO 1548–1553

[47] Panda S and Padhy N P 2008 Comparison of particle swarm optimization and genetic algorithm for FACTS-based controller design. *Appl. Soft Comput.* 8(4): 1418–1427

[48] Suprayogi S and Cakravastia A 2014 Genetic algorithm for solving a helicopter routing problem. In: *The 6th International Conference on Operations and Supply Chain Management* OSCM Bali, Indonesia

[49] Yang F, Chen Z and Wei C 2013 Nonlinear system modeling and identification of small helicopter based on genetic algorithm. *Int. J. Intell. Comput. Cybern.* 6(1): 45–61

[50] Mirjalili S, Mirjalili S M and Lewis A 2013 Grey wolf optimizer. *Adv. Eng. Softw.* 69: 46–61

[51] Zhang S, Li Z, Zhiming and Pan W 2016 Grey Wolf optimizer for unmanned combat aerial vehicle path planning. *Adv. Eng. Softw.* 99: 121–136

[52] Yingxun W, Tian Z, Zhihao C, Jiang Z and Kun W 2020 Multi-UAV coordination control by chaotic grey wolf optimization based distributed MPC with event-triggered strategy. *Chin. J. Aeronaut.* 33(11): 2877–2897

[53] Padhy S and Panda S 2021 Application of a simplified Grey Wolf optimization technique for adaptive fuzzy PID controller design for frequency regulation of a distributed power generation system. *Prot. Control Mod. Power Syst.* 6(2): 1–6

[54] Akin A and Aydogdu I 2015 Optimum design of steel space frames by hybrid teaching-learning based optimization and harmony search algorithms. World Academy of Science Engineering and Technology. *Civ. Environ. Eng.* 2(7): 739–745

[55] Alneamy J S M and Alnaish R A H 2014 Heart disease diagnosis utilizing hybrid fuzzy wavelet neural network and teaching learning based optimization algorithm. *Adv. Artif. Neural Syst.* 14(3): 6–6

[56] Aydogdu I and Akin A 2014 Teaching and learning-based optimization algorithm for optimum design of steel buildings. *Comput. Civ. Build. Eng.* 2(7): 2167–2175

[57] Babu B and Palaniswami S 2015 Teaching learning based algorithm for OPF with DC link placement problem. *Int. J. Electr. Power Energy Syst.* 73(1): 773–781

[58] Baghlani A 2014 Water level stabilization in open channels using Chebyshev polynomials and teaching–learning-based optimization. *J. Hydroinform.* 16(5): 1097–1109

[59] Chen D, Zou F, Li Z, Wang J and Li S 2015 An improved teaching–learning-based optimization algorithm for solving global optimization problem. *Inf. Sci.* 297: 171–190

[60] Cheng Y H 2013 A novel optimization method for picking PCR oligonucleotide primers. *Int. J. Comput. Sci. Electron. Eng.* 1(4): 518–523

[61] Daljit K and Ranjit K 2013 A design of IIR based digital hearing aids using teaching-learning-based optimization. *Int. J. Comput. Eng. Appl.* 3(2–3): 180–190

[62] Das S P and Padhy S 2015 A novel hybrid model using teaching–learning-based optimization and a support vector machine for commodity futures index forecasting. *Int. J. Mach. Learn. Cybern.* 9(3): 97–111

[63] Rao R, Savsani V and Vakharia D P 2011 Teaching–learning-based optimization: a novel method for constrained mechanical design optimization problems. *Comput.-Aid. Des.* 43(3): 303–315

[64] Kumar E V, Subramanian G and Jovitha J 2016 Adaptive PSO for optimal LQR tracking control of 2 DoF laboratory helicopter. *Appl. Soft Comput.* 41: 77–90

[65] Ramalakshmi AP, Manoharan PS and Deepamangai P 2013 PID tuning and control for 2-DOF helicopter using particle swarm optimization 8297:662–672

[66] Beloli AS, Florêncio JL and Cavalca MSM 2013 A 2 DoF low cost control workstation for control techniques application. In: *22nd International Congress of Mechanical Engineering* (COBEM 2013). Brazil. 13: 3101–3112

[67] Zheng B and Zhong Y 2011 Robust attitude regulationof a 3-DOF helicopter benchmark: theory and experiments. *IEEE Trans. Ind. Electron.* 58(2): 660–670

[68] Hernandez-Gonzalez M, Alanis A and Hernandez-Vargas E 2012 Decentralized discrete-time neural controlfor a Quanser 2-DOF helicopter. *Appl. Soft Comput.* 12(8): 2462–2469

[69] Yanpeng H, Yanping Y, Shu L and Yaoming Z 2020 Fuzzy controller design of micro-unmanned helicopter relying on improved genetic optimization algorithm. *Aerosp. Sci. Technol.* 98: 105685

[70] Zavala A, Aguirre A and Diharce E 2005 Constrained optimization via evolutionary swarm optimization algorithm (PESO). *Conf. Genetic Evolut. Comput.* 2005: 209–216

[71] Bristol E 1966 On a new measure of interaction for multivariable process control. *IEEE Trans. Autom. Control.* 11(1): 133–134

[72] Ellis G 2012 Control system design guide (fourth edition) 6.4.1.3, Butterworth-Heinemann, Book, pp. 97–119, ISBN 9780123859204

[73] VOLTPAQ-X2' Reliable Linear Voltage Amplifier for Superior Real Time Hardware-in-the-Loop (HIL) Performance. Markham, ON, Canada

[74] Karaboga D and Basturk B 2007 Artificial Bee Colony (ABC) Optimization Algorithm for Solving Constrained Optimization Problems. 4529: 789–798. https://doi.org/10.1007/978-3-540-72950-1_77

[75] Becerra R L and Coello C 2006 Cultured differential evolution for constrained optimization. *Comput. Methods Appl. Mech. Eng.* 195: 4303–4322

[76] Mezura-Montes E and Coello C 2005 A simple multimembered evolution strategy to solve constrained optimization problems. *IEEE Trans. Evolut. Comput.* 9(1): 1–17

ECS Journal of Solid State
Science and Technology

The Electrochemical Society
Advancing solid state & electrochemical science & technology

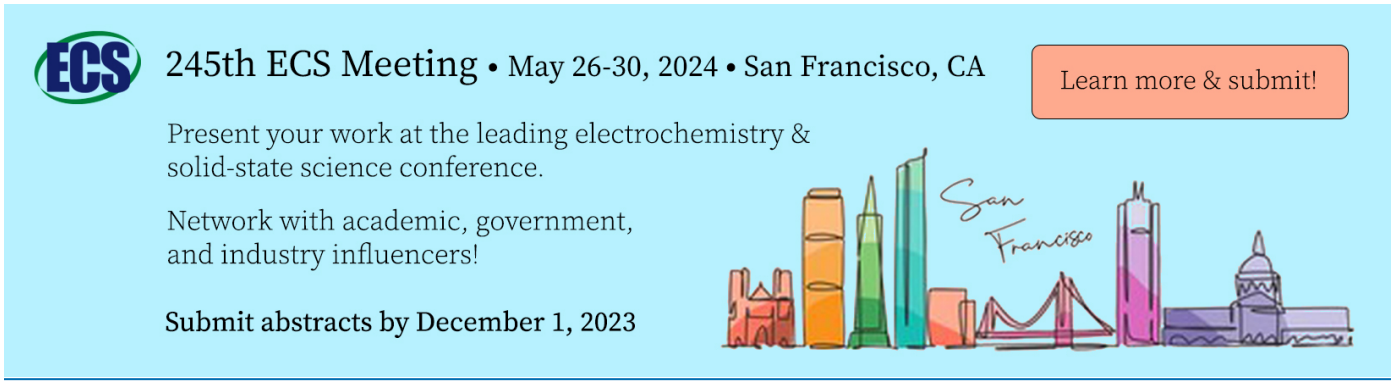
An In-Depth Analysis of Variation in Characteristic Performance of OLED with Respect to Position of Charge Generation Layer

To cite this article: Sugandha Yadav *et al* 2023 *ECS J. Solid State Sci. Technol.* **12** 106001

View the article online for updates and enhancements.

You may also like

- Three-dimensional Simulations and Spacecraft Observations of Sub-ion Scale Turbulence in the Solar Wind: Influence of Landau Damping
S. Kobayashi, F. Sahraoui, T. Passot et al.
- Guiding-centre theory for kinetic-magnetohydrodynamic modes in strongly flowing plasmas
S Lanthaler, J P Graves, D Pfefferlé et al.
- POLARIZATION AND COMPRESSIBILITY OF OBLIQUE KINETIC ALFVÉN WAVES
P. Hunana, M. L. Goldstein, T. Passot et al.



This content was downloaded from IP address 14.139.251.98 on 03/11/2023 at 06:12

ECS Journal of Solid State Science and Technology, 2023 **12** 106001
2162-8777/2023/12(10)/106001/10/$40.00 © 2023 The Electrochemical Society ("ECS"). Published on behalf of ECS by IOP Publishing Limited

An In-Depth Analysis of Variation in Characteristic Performance of OLED with Respect to Position of Charge Generation Layer

Sugandha Yadav,[1] Poornima Mittal,[1,z] and Shubham Negi[2]

[1]*Department of Electronics & Communication Engineering, Delhi Technological University, New Delhi, 110042, India*
[2]*Department of Electronics and Communication Engineering, Chandigarh University, Mohali Punjab, 140413 India*

In this paper, a high performance blue organic light emitting diode having a charge generation layer (CGL) is proposed and compared with the other five CGL and non-CGL based devices. The utilized CGL layer in the different structure consists of two materials; HAT-CN (hexaazatriphenylene-hexacarbonitrile) and TAPC (1,1-bis[(di-4-tolyamino)phenyl)]cyclohexane,) for electrons and holes generation, correspondingly. In the proposed novel structure, the CGL layer is incorporated outside of the emissive layer (EML) which significantly enhances the device performance in terms of current and luminescence. The device exhibits luminescence and current values as 3636.3 cd m^{-2} and 0.44 A, respectively. Furthermore, this paper represents in-depth internal analysis of the six devices (D_1-D_6). This analysis is provided by drawing horizontal and vertical cutlines inside the devices. The proposed device is analysed and compared with other mentioned devices in terms of several parameters such as Langevin recombination rate, electron concentration, hole concentration, band energy, total current density, electron affinity, hole QFL (quasi-Fermi level), conduction current density, potential distribution and electron/hole mobility. In comparison with D_1, D_2, D_3, D_4 and D_5, the current of the proposed device (D_6) is about 16.9, 2.2, 1.7, 3 and 1.6 times improved, correspondingly. Moreover, structural analysis is also included to understand the performance of the devices more precisely.
© 2023 The Electrochemical Society ("ECS"). Published on behalf of ECS by IOP Publishing Limited. [DOI: 10.1149/2162-8777/acfd5f]

Manuscript submitted August 29, 2023; revised manuscript received September 15, 2023. Published October 5, 2023.

During the past few years, full colour displays with high resolution have been developed by many researchers.[1–4] These displays containing organic materials have various attractive features likes flexibility, light weight, fast response, easy fabrication, wide viewing angle, better lifetime etc.[5–8] On the other hand, there is still scope of improvement in such devices in terms of current density, luminesce and lifetime. Various researchers have investigated these displays and improved them by adding few additional layers like electron/hole transport layer, electron/hole injection layer, hole blocking layer etc. These different layers have refined the performance of these display device very significantly.[9,10] But, in recent time some novel layers have been utilized to enhance the performance of the display devices such as charge generation layers,[11] spacer layers[12] and mixed interlayer.[13]

This paper demonstrates a positional analysis of charge generation layer (CGL) to enhance the performance of an OLED in terms of various parameters such as current density and luminescence. Charge generation layer plays vital role to improve overall device performance.[11] While the recombination process, the OLED can utilize only 20% of the total light. The remaining light is lost due to optical losses and Plasmon. With the utilization of CGL, this lost light can be prevented which eventually enhance the performance of the device in terms of EL emission and efficiency.[14] Herein, the CGL is the combination of two different type of semiconductor layers (p type & n type) which act as electrodes.[15] Charge is produced depending upon the external applied voltage and inserted towards emissive layer where the photons are generated after the recombination.

Many researchers have investigated CGL to improve the performance of OLEDs.[16–19] Kim et al.[20] explained the emission properties of blue fluorescent multi-tandem OLED using CGL with the materials Al/MoO_X. In this research, the current efficiency and maximum quantum efficiency are achieved as 46.6 cd A^{-1} and 25.6%, respectively. Moreover, three OLEDs which are blue, green and red are fabricated by Yuan et al.[21] utilizing a CGL of the materials HAT-CN/TAPC. Here, the device EL efficiencies are improved having balanced injection of electrons and holes. Afterwards, a CGL layer of C70/Pentacene is reported by Jelty et al.[14] They proposed a high-performance organic LED which contains organic photovoltaic behaviour. The proposed device exhibited current density and luminescence as 15.63 cd A^{-1} and 15630 cd cm^{-2}, correspondingly.

Herein, a highly efficient organic LED is proposed using a charge generation layer which is a combination of two materials which are 1,1-bis[(di-4-tolyamino)phenyl)]cyclohexane (TAPC) and 1,4,5,8,9,11-hexaazatriphenylene-hexacarbonitrile (HAT-CN). The CGL is placed outside the emissive layer in the proposed device. Incorporation of CGL in OLED provided better charge injection which results in higher current density and better luminous property. Additionally, this paper also compares five different CGL-based devices in terms of luminescence and current density. Moreover, this paper also covers in-depth internal analysis of six different devices (D_1- D_6) using horizontal and vertical cutline analysis and investigated the devices on the basis of various important parameters discussed in Internal Analysis of all Devices section.

This complete paper is assembled in total four sections. Introduction Section provides introduction of OLED and discusses the importance of charge generation layer. This section also highlights the work of few researchers on the basis CGL layer. Thereafter, Various OLED structures based on charge generation layer section includes six different blue OLED devices and compares their performance in terms of current and luminescence. Internal Analysis of all Devices Section incorporated an in-depth internal analysis of all the six devices. The internal analysis is provided with the help of horizontal and vertical cutline analysis. Eventually, In Conclusion section, necessary conclusions are drawn.

Various OLED Structures Based on Charge Generation Layer

As mentioned in Introduction section, CGL is very important layer for increasing the charge carrier concentration which eventually improves various parameters such as recombination rate, current density, and luminous properties of the device. Figure 1 represents one reference device i.e., blue OLED D_1[22] with no CGL layer and four different blue OLEDs (D_2-D_5) having charge generation layer. Herein, these devices will be discussed in terms of their structures, materials, thickness, and performance parameters. In all the devices except D_1, the CGL is placed at various positions where in CGL, TAPC is working as hole generator and HAT-CN is working as electron generators. In D_2 and D_3 devices, the CGL layer is utilized below cathode and above anode, correspondingly. Moreover, in case of D_4, CGL is incorporated at both sides of the device like below cathode and above anode. Addition to this, in D_5, CGL is situated inside emissive layer.

Proposed device.—Now, the device D_6 which is the proposed device, contains CGL outside of the emissive layer. Additionally, in

[z]E-mail: poornimamittal@dtu.ac.in

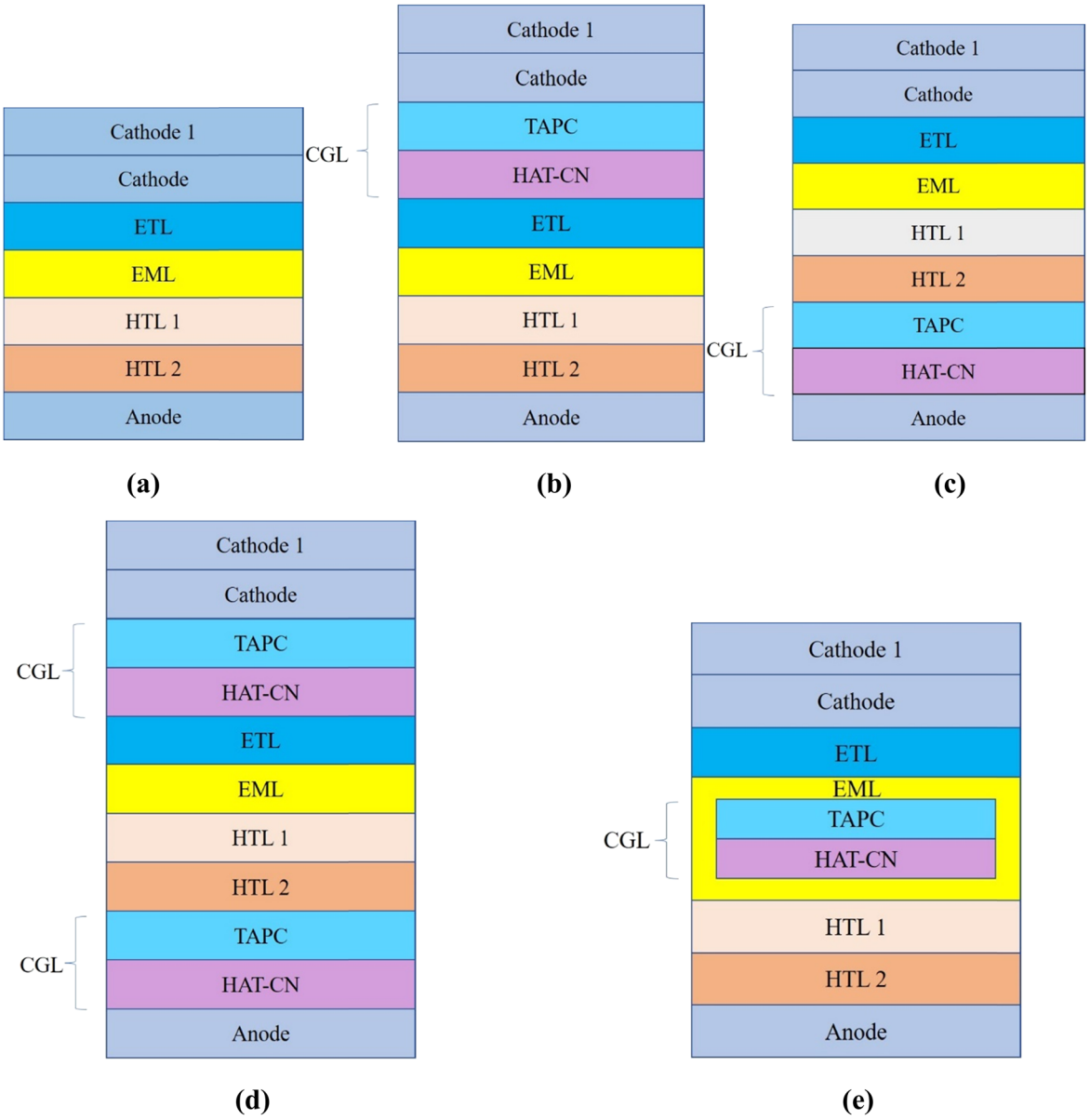


Figure 1. Different structures of Blue OLEDs (a) Reference device with no CGL (D_1),[22] (b) CGL below cathode (D_2), (c) CGL above anode (D_3), (d) CGL at both side of the device (D_4) and (e) CGL inside emissive layer (D_5).

this case, the position of TAPC and HAT-CN is also reversed as compared to the other devices. Figure 1 is illustrating the positional variation of D_1-D_5 devices in terms of CGL layer. The structure of the proposed device (D_6) is shown in Fig. 2.

The materials and thickness of all the devices mentioned here are depicted in Table I. All the devices are analysed with the same materials and same possible thickness. Only in the case of D_5 and D_6, there are few minor variations in thickness of the emissive layer and charge generation layer due to their structural diversities.

All the above-mentioned devices (D_1-D_6) are simulated using Silvaco Atlas 2-D device simulator. Herein, for the validation of the fabricated device,[22] a model i.e. Langevin's recombination rate is utilized in the simulator. The expression for the model is provided below:

$$R_L(n,\ p) = r_l(x,\ y,\ t)\,(np - n_i^2) \qquad [1]$$

Here, r_l, n, p and n_i denote the recombination rate coefficient, electron concentration, hole concentration, and intrinsic concentration, respectively. Moreover, Poole-Frenkel mobility model is applied to simulate the carrier's transportation of a device structure. This model provides the results under specified boundary conditions. The model is as follows:

$$\mu(E) = \mu_0 e^{\left[-\frac{\Delta}{kT} + \left(\frac{\beta}{kT} - \gamma\right)\sqrt{|E|}\right.} \qquad [2]$$

where μ, μ_0 and E are field dependent mobility, zero field mobility and electric field, respectively. Besides this, Δ, k, T, β and γ represent activation energy at zero electric field, Boltzmann constant, effective temperature for charge carriers, hole Pool–Frenkel factor and fitting parameter, correspondingly.

Mechanism of CGL in the OLED devices (D_2-D_5) and proposed device (D_6).—In D_2, D_3 and D_4, the utilization of CGL is not very close to the emissive layer. Hence, the device performance is not enhanced very significantly in these devices. Among these devices, D_3 is exhibiting the best performance. In D_3, the CGL is placed above the anode. Here, electrons in HAT-CN are attracted from anode side and holes in TAPC are repelled from anode side which

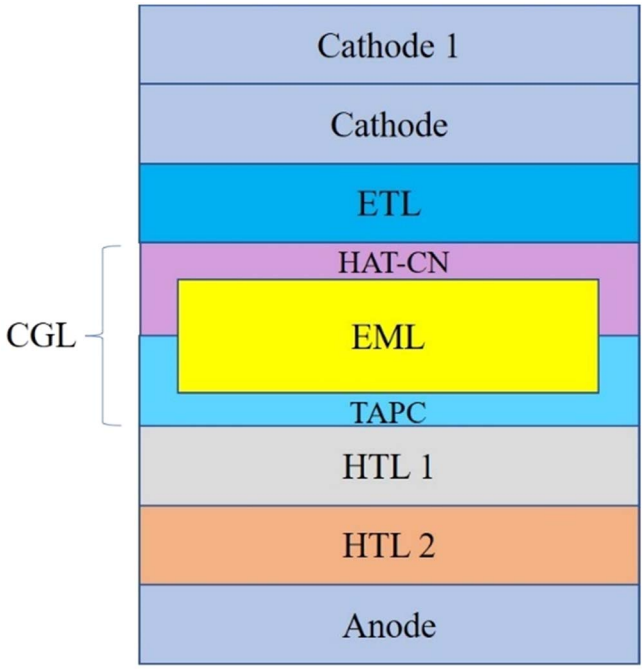


Figure 2. Proposed device (D_6).

will move towards the emissive layer. Also, it is known that the mobility of holes is better in comparison with electrons in organic electronics. Hence, hole concentration will be improved and eventually recombination rate will be improved.

In case of D_5, CGL is situated inside the emissive layer which significantly improves the recombination of charge carriers as compared to D_1, D_2, D_3, and D_4. The reason behind this is very clear. As it is mentioned already that the bottom layer (HAT-CN) of CGL is behaving like electron generator and top layer (TAPC) of CGL is acting as hole generator. Now, the electrons present in HAT-CN will get attracted from anode side and moves through emissive layer. In the same manner, the holes present in TAPC will be attracted from cathode side and will move through emissive layer. Now lot of charge carriers generated from CGL participate in recombination process as these charge carriers will be mixed up with the charge carriers generated from electrodes (cathode and anode) inside the emissive layer. Very few charge carriers cross EML. This is how the charge carrier concentration will be increased inside EML which will eventually increase the recombination rate. Here, a very good amount of current will be produced (~0.28 A) which is almost 10 times increased on comparing with D_1 (reference device).

Among all the mentioned six devices, the proposed device (D6) is exhibiting the remarkable performance. In this device, the CGL is incorporated outside of the emissive layer. Also, the position of top and bottom layer of CGL are reversed as depicted in Fig. 2. In device D5, HAT-CN is injecting electrons towards emission layer on the anode side and these electrons will get recombined with holes coming from anode. Similarly, TAPC is injecting holes towards the emission layer on the cathode side. These will get recombined with electrons coming from cathode. But as the direction of hole and electron injection from CGL is different, then there could be a lower recombination if charge carriers injected from electrodes are lower.

This problem is solved by D6, wherein emission layer is in between CGLs. Additionally, the position of two CGL is also changes now. Therefore, holes are injected into the EML by CGL as well as anode in the same direction. Similar is the case with electrons provided by the cathode and CGL HAT-CN. The carriers injected by both the electrodes and CGL will meet at the EML and therefore, it is expected that maximum recombination will be achieved. This can be perceived by the results shown in Fig. 3 wherein, maximum luminance is observed for D6. Herein, all five devices (D_2-D_6) are analysed and compared in terms of current density and luminescence with reference device D_1.

Figure 3 is illustrating the current and luminescence curves of all the six devices (D_1-D_6). It is clearly seen that D_6 is exhibiting the best performance in terms of current and luminescence as 0.44 A and 3636.3 cd m^{-2}, correspondingly as highlighted in Table II. The Luminescence of the novel proposed structure (D_6) is obtained about 16.8, 2.3, 1.7, 3 and 1.6 times better in comparison with D_1, D_2, D_3, D_4 and D_5, respectively.

Internal Analysis of all Devices

In this section, all the mentioned devices will be studied and investigated with the help of internal analysis for the better understanding. This in-depth analysis will be performed by drawing a cutline on the device in horizontal and vertical manner. All the devices will be analysed in terms of various important parameters such as hole/electron concentration, recombination rate, electric field distribution etc.

Vertical cutline analysis.—In this sub-section, a vertical cutline is performed on the simulated structures of all the six devices at 200 μm as highlighted in Fig. 4. Here, several parameters such as hole concentration, electron concentration, total current density and Langevin recombination rate, conduction current density, electron affinity etc. are investigated to understand the performance of different devices. In Fig. 5a, it is seen that the hole concentration is higher at bottom side of the devices (0.18 μm-0.36 μm) as holes travels from anode side. The minimum and maximum concentration is exhibited by D_1 and D_6 device, correspondingly. In all the cases, it is maximum near the emissive layer where holes will be accumulated eventually for the recombination process. If the proposed device (D_6) is compared here with the reference device (D_1), the hole concentration of D_6 is about 3 times higher due to the

Table I. Materials and thickness for each layer for devices D_1-D_6.

		Thickness (nm)					
Layer	Materials	D_1	D_2	D_3	D_4	D_5	D_6
Anode	ITO	20	20	20	20	20	20
HTL1	PEDOT:PSS	90	90	90	90	90	90
HTL2	NPB	60	60	60	60	60	60
EML	DPVBi & BCzVBi	20	20	20	20	40	20
ETL	BPhen	30	30	30	30	30	30
Cathode	LiF	1	1	1	1	1	1
Cathode1	Al	150	150	150	150	150	150
CGL	TAPC	NA	10	10	20	10	20
CGL	HAT-CN	NA	10	10	20	10	20

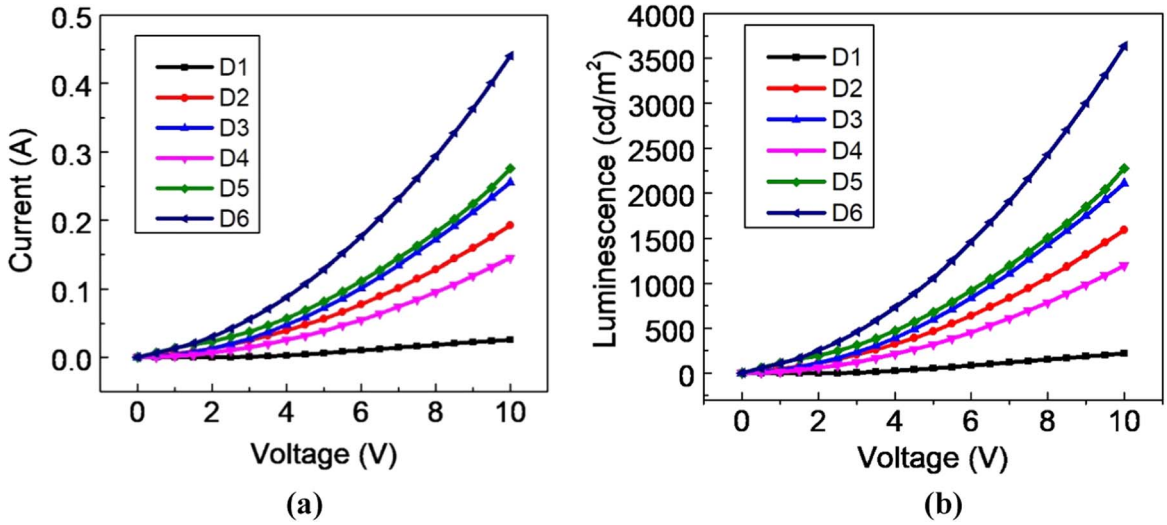


Figure 3. (a) Current and (b) luminescence curve for all six devices (D_1-D_6).

Table II. Parametric comparison of D_1-D_6.

Device	Current (A)	Luminescence (cd/m^2)
D_1 (Reference)	0.026	215.9
D_2 (Modified)	0.193	1592.3
D_3 (Modified)	0.256	2112.2
D_4 (Modified)	0.145	1198.4
D_5 (Modified)	0.276	2277.2
D_6 (Proposed)	0.44	3636.3

incorporation of CGL outside the emissive layer. The small peaks around 0.1 to 0.2 μm are highlighting the presence of TAPC (acts a hole generator) layer at cathode side in D_2 and D_4.

On the other hand, Fig. 5b is illustrating the electron concentration of all the devices. In this case, it can be noticed that, the electron concentration is maximum around 0.2 μm in all the devices because this region belongs to the emissive layer. Herein, the minimum and maximum concentration is obtained by D_1 and D_6 device, respectively. Now, if both the figures are observed (5 (a) & (b) deeply, the reason of the significant enhancement in charge concentration of the proposed device (D_6) around 0.2 μm is very clear. In the proposed device, CGL is situated outside the emissive layer which directly improves the carrier concentration of the device.

Moreover, Fig. 5c is depicting the total current density of the devices (D_1- D_6). The proposed device exhibits the highest current density which is almost 73, 5.7, 4.4, 7.8 and 1.4 times higher than D_1, D_2, D_3, D_4 and D_5, correspondingly. This is because of higher electron/hole concentration in D_6 as discussed above. All the discussed parameters play very vital role for landmark enhancement in the Langevin recombination rate of the proposed device as illustrated in Fig. 5d. In this figure, it can be seen that recombination peaks are achieved around 0.2 μm for the obvious reason as this area belongs to emissive layer. The materials TAPC and HAT-CN inject the charge carrier (holes & electrons) directly from the outside of EML which automatically improves the Langevin recombination rate of the proposed device.

Furthermore, Fig. 6a is showing the potential distribution of the proposed device (D_6). In this figure, the voltage is maximum at 0.4 μm (cathode side) and it keeps decreasing towards anode and at last becomes zero at 0 μm. Next, Figs. 6b–6d are comparing the performance of D_5 and D_6 in terms of conduction current density, electron affinity and acceptor trap ionized density. There is very much improvement in all the three parameters mentioned in case of D_6 in comparison with D_5.

Horizontal cutline analysis.—Here, comparison of the parameters is highlighted by performing a horizontal cutline inside HAT-CN and TAPC, individually. Both the analysis based on HAT-CN and TAPC are discussed in below sub-sections.

Cutline analysis inside HAT-CN layer.—In this case, a horizontal cutline is drawn at 0.195 μm inside HAT-CN in both D_5 and D_6 devices as depicted in Fig. 7. The parameters such as Langevin recombination rate, conduction current density, conduction band, valence band and hole QFL (quasi-Fermi level) are investigated to understand the performance comparison of D_5 and D_6 devices. The peak Langevin recombination rate of D_6 is almost 10 times improved as compared to D_5 as depicted in Fig. 8a. In this figure, the drop and peak in the values from 40 to 360 μm is due to the positional change in TAPC and HAT-CN in D_5 and D_6. Similarly, conduction current density of D_6 is also about 8 times enhanced on comparing with D_5 as illustrated in Fig. 8b.

Furthermore, band energies are also incorporated in Figs. 8c and 8d. Figure 8c is representing the valence and conduction band of D_5 and Fig. 8d is highlighting the valence and conduction band of D_6. Moreover, Fig. 8e is illustrating hole QFL of the both devices. This parameter is showing the change in Fermi level as charge carriers are excessively increased in both devices. Comparivelty, hole QFL is obtained higher in case of D_6 as in D_6 more carriers are accumulated for the recombination because of the positional variation of CGL.

Cutline analysis inside TAPC layer.—In this sub-section, a horizontal cutline is drawn at 0.215 μm inside the TAPC layer as represented in Fig. 9. Herein, all six devices (D_1-D_6) are analysed and compared in terms of hole mobility, electron mobility and Langevin recombination rate. It can be clearly observed from the Fig. 10a that D_6 exhibits extremely high hole mobility i.e. 100 cm^2/Vs which is 12, 6.4, 1.9, 8.1 and 1.6 times increased as compared to D_1, D_2, D_3, D_4 and D_5 devices, respectively. In the same manner, electron mobility of D_6 is obtained as 51.2 cm^2/Vs which is about 11, 5.2, 2, 6.2 and 1.9 times increased on comparing with D_1, D_2, D_3, D_4 and D_5, correspondingly as outlined in Fig. 10b. Furthermore, Fig. 10c is showing the peak recombination rate of all the devices in graphical manner. The values of recombination rate are slightly lower as compared to the values of the parameter in case of HAT-CN (discussed in 3.2.1). This happens only due to mobility difference between holes and electrons. Addition to this, the parameter hole QFL is also compared as depicted in Fig. 10d. This parameter is also showing slightly lower values as compared to HAT-CN values for the same.

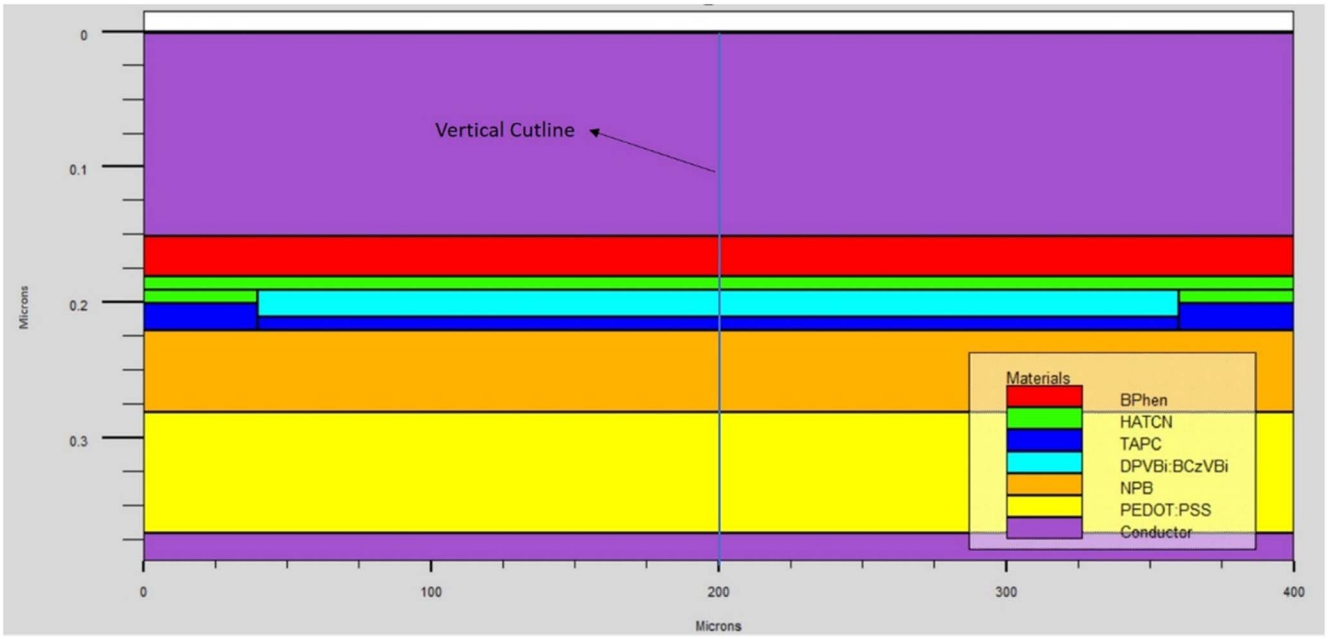


Figure 4. Structure of the proposed device (D_6) along with vertical cutline at 200 μm.

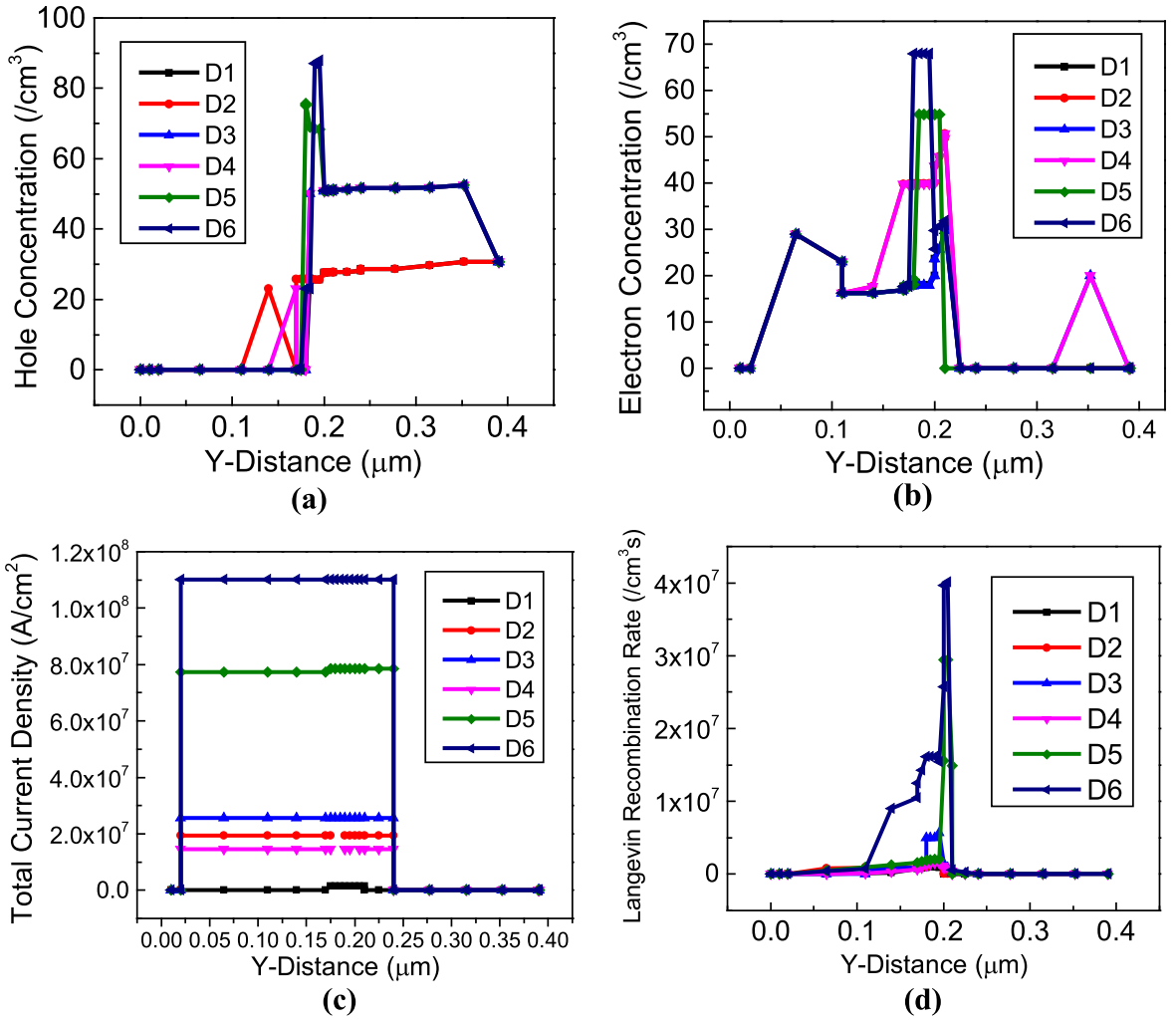


Figure 5. Comparison of various parameters with vertical cutline in terms of (a) Hole concentration, (b) Electron concentration, (c) Total current density, (d) Langevin recombination rate of all six devices (D_1-D_6).

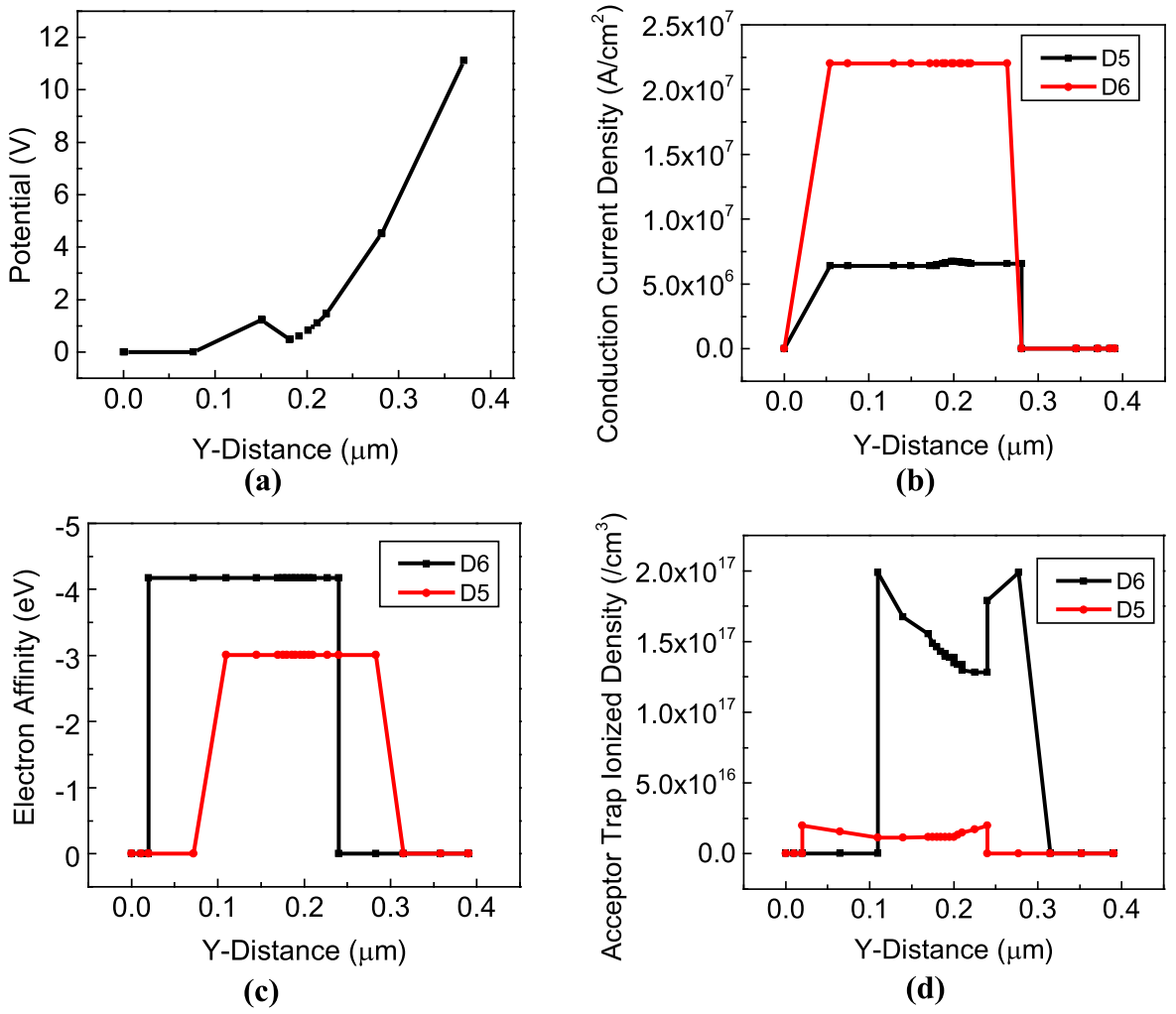


Figure 6. (a) Potential distribution of the proposed device (D_6), (b) Conduction current density, (c) Electron affinity and (d) Acceptor trap ionized density of D_5 and D_6.

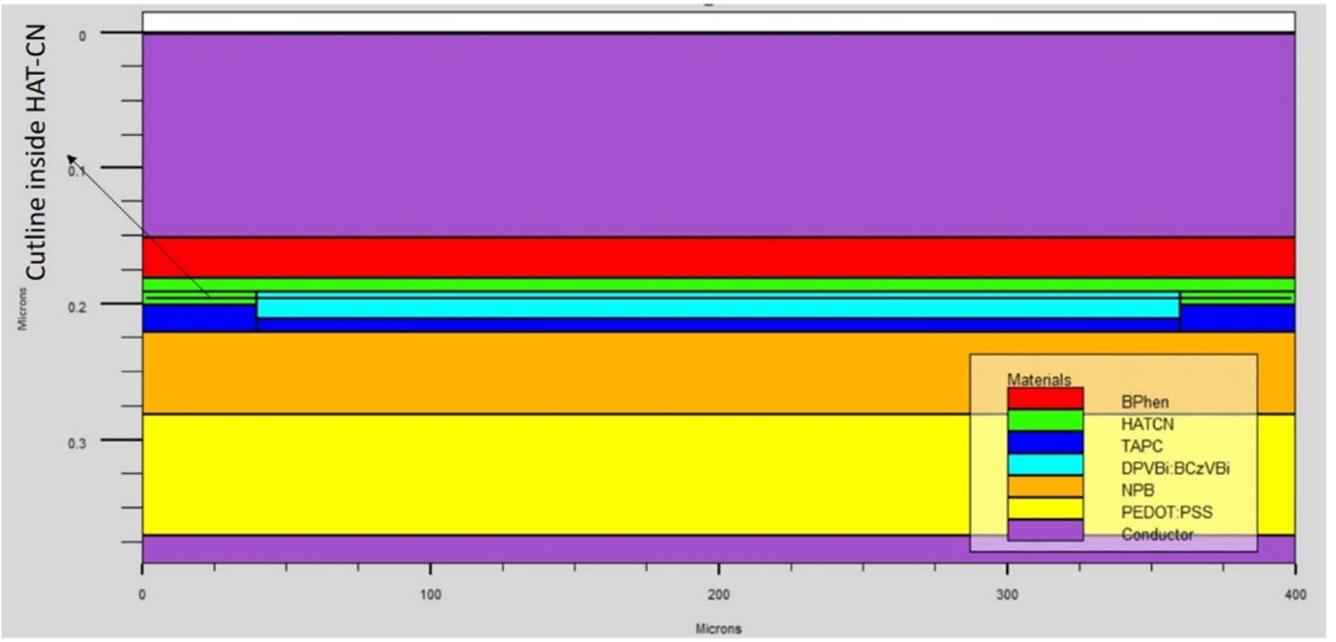


Figure 7. Structure of the proposed device (D_6) along with horizontal cutline inside HAT-CN at 0.195 μm.

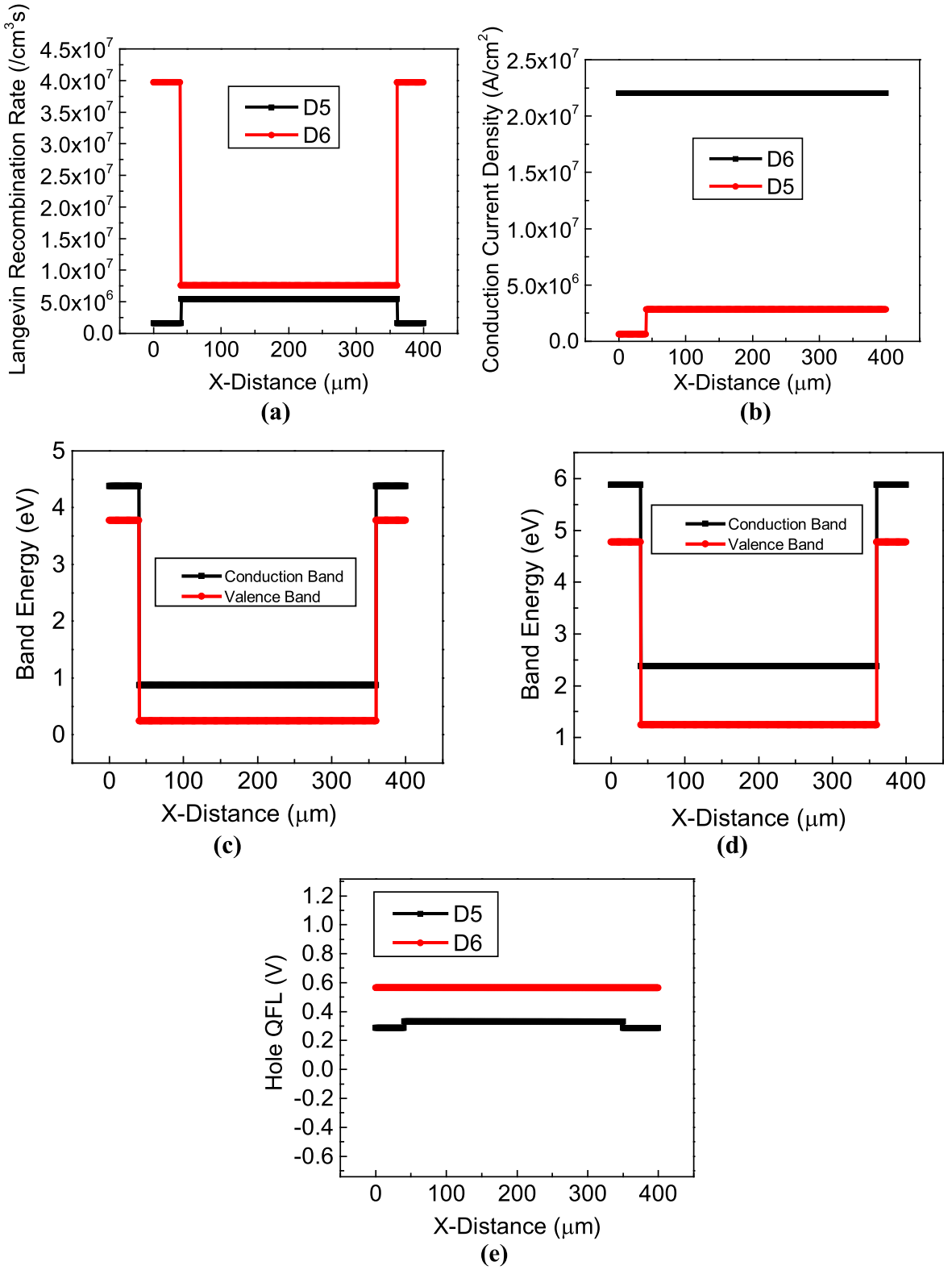


Figure 8. Comparison of various parameters of D_5 and D_6 with horizontal cutiline (a) Langevin recombination rate, (b) conduction current density, (c) Band energy of D_5, (d) Band energy of D_6 and (e) Hole QFL.

Structural internal analysis.—Apart from the cutline analysis, this paper also includes the structural layer wise analysis as illustrated in Fig. 11. These figures are the zoomed versions of the D_5 and D_6 structures. The parameters net doping and conduction current density are analysed more deeply by investigating the perticular affected regions mainly around emissive layer as shown in below figures.

Figures 11a and 11b are showing the net doping of D_5 and D_6, correspondingly. It can be clearly noticed that maximum net doping which is highlighted in red colour is achieved in the wider region as

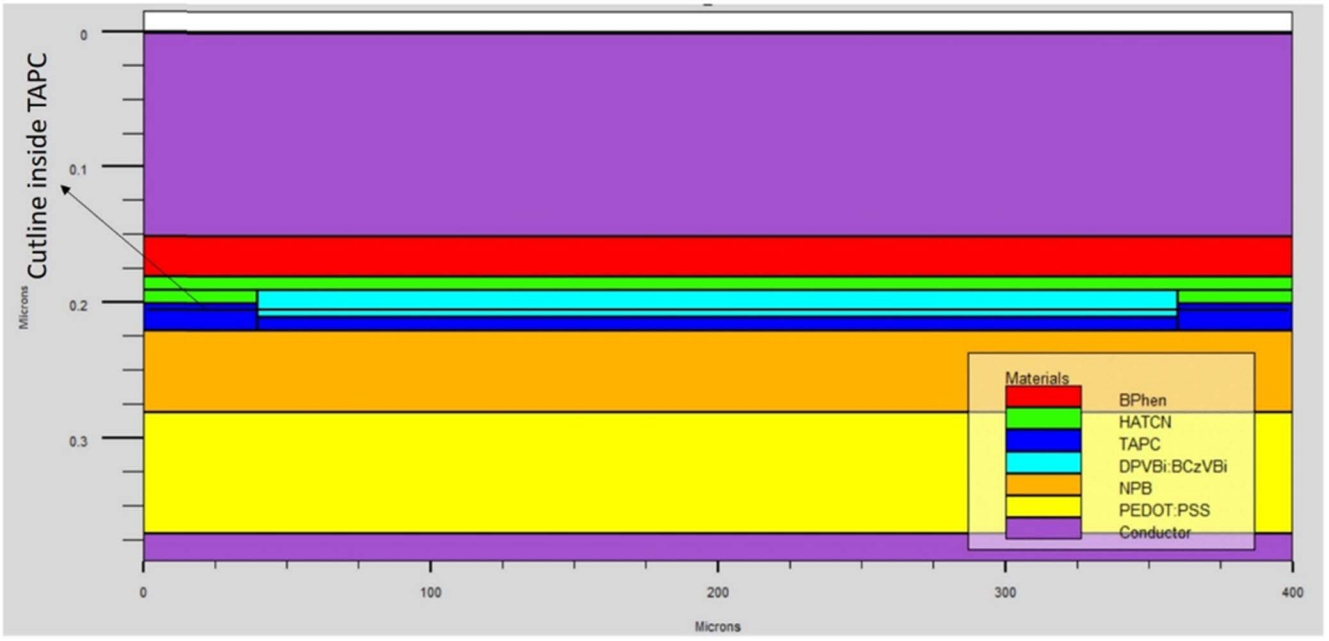


Figure 9. Structure of the proposed device (D_6) along with horizontal cutline inside TAPC at 0.215 μm.

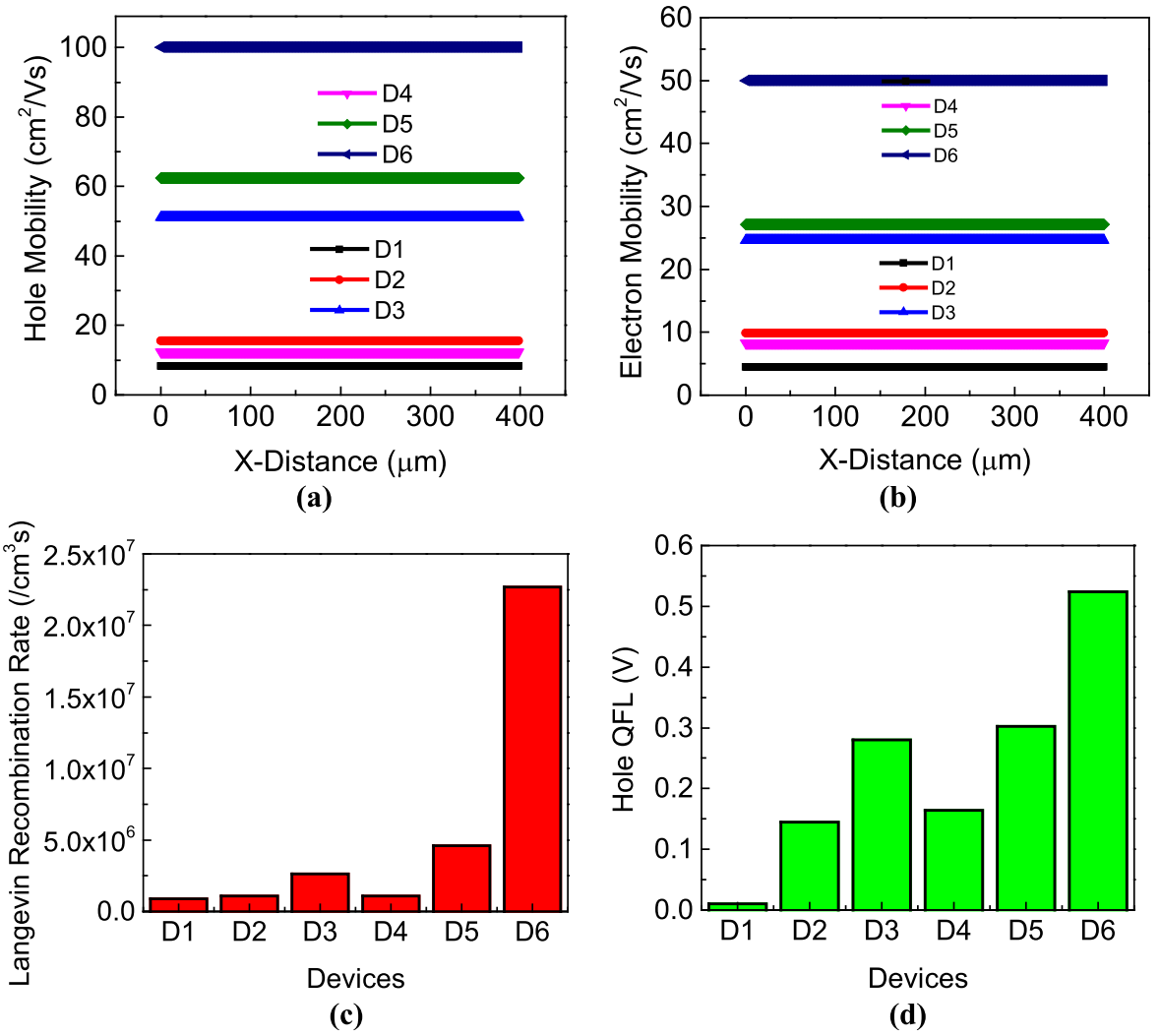
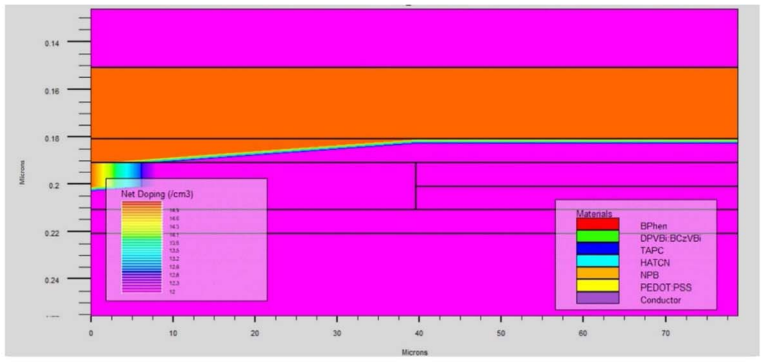
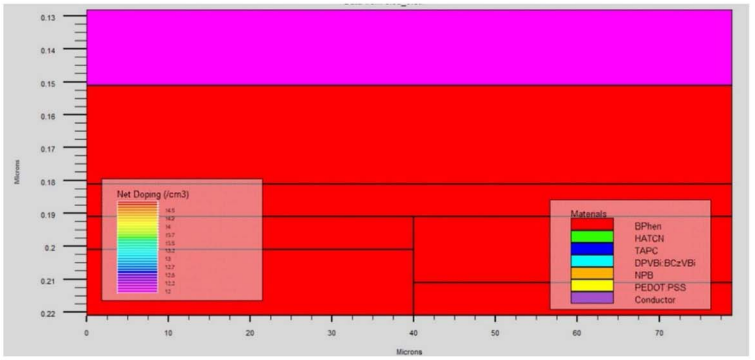


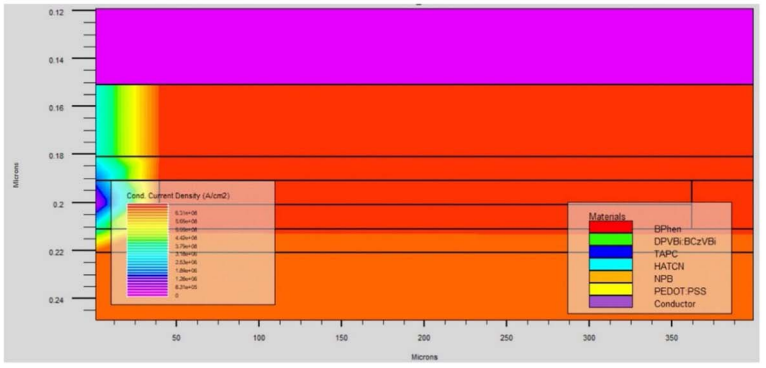
Figure 10. Comparison of (a) Hole mobility and (b) Electron mobility (c) Langevin recombination rate of all six devices (D_1-D_6) with horizontal cutline.



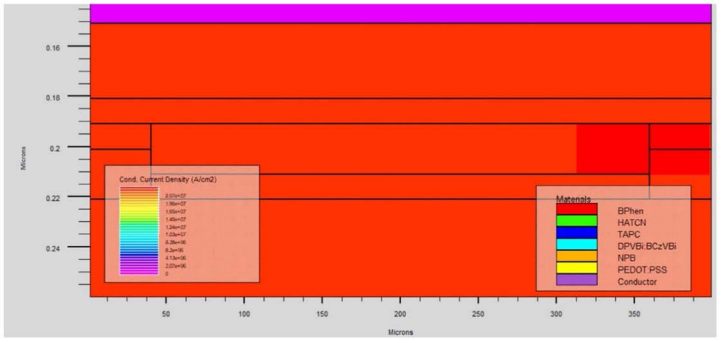
(a)



(b)



(c)



(d)

Figure 11. Zoomed view of D_5 and D_6 in terms of Net doping (a) D_5 and (b) D_6 and conduction current density (c) D_5 and (d) D_6.

compared to D_5. In D_5, it is only higher near the edge of the emissive layer. This illustration clearly depicts that in D_6 there are large amount of charge carriers accumulated which is possible because of 100% utilization of the charge carriers from CGL layer. Hence, device performance of D_6 is very much improved comparatively. Additionally, Figs. 11c and 11d are representing the conduction current density of D_5 and D_6, respectively. Herein, this parameter is the maximum in whole region of emissive layer in case of D_6 as

shown in Fig. 11d. But, in Fig. 11c which is showing the performance of D_5, conduction current density is lower at outside of CGL layer particularly at left side and below side of CGL layer.

Conclusions

This paper represents a proposed device with CGL layer of two materials HAT-CN (as electron injector) and TAPC (as hole injector). The novel proposed structure (D_6) is deeply analysed and compared with other five non-CGL and CGL based devices. Herein, device D_6 exhibits maximum luminescence as 3636.3 cd m^{-2} which is approximately 16.8, 2.3, 1.7, 3 and 1.6 times improved as compared with D_1, D_2, D_3, D_4 and D_5, correspondingly. Moreover, in this paper an in-depth internal analysis is also incorporated which highlights the performance of all the six devices on the basis of various parameters such as hole/electron concentration, recombination rate of charge carriers, electron affinity, conduction current rate, band energies, hole QFL, acceptor trap density ionized density, net doping and hole/electron mobility. This internal analysis concludes that the proposed device exhibits the best performance among all the device in every aspect.

Acknowledgments

Author's Contributions. Sugandha Yadav: Conceptualization, Methodology, Data curation, Writing—original draft, Writing—review & editing. Poornima Mittal: Conceptualization, Methodology, Supervision, Data curation, Writing—review & editing, Validation. Shubham Negi: Supervision, Writing—review & editing, Visualization, Investigation, Validation. Availability of data and materials. Data and materials utilized and supporting the results of this study are available within the article.

Funding

Not Applicable

ORCID

Poornima Mittal https://orcid.org/0000-0002-9479-8628

References

1. P. Chamola and P. Mittal, "Flexible organic solar cell to power modern cardiac pacemakers: versatile for all age groups, skin types and genders." *Phys. Scr.*, **98**, 035018 (2023).
2. P. Negi, Mittal, and B. Kumar, "Characteristic performance of OLED based on hole injection, transport and blocking layers." *Recent Pat.*, **14**, 373 (2020).
3. E. Angioni et al., "A single emitting layer white OLED based on exciplex interface emission." *J. Mater. Chem. C*, **4**, 3851 (2016).
4. S. Yadav, P. Mittal, and S. Negi, "Recent advancements over a decade for organic light-emitting diodes: from structural diversity, role of layers, colour emission, material classification, performance improvement, fabrication to applications." *Bull. Mater. Sci.*, **45**, 109 (2022).
5. H. Lee, "Phosphorescent white organic light-emitting diodes with stable white color depending on luminance." *Curr. Appl Phys.*, **21**, 116 (2021).
6. B. Huang, X. Ban, K. Sun, Z. Ma, Y. Mei, W. Jiang, B. Lin, and Y. Sun, "Thermally activated delayed fluorescence materials based on benzophenone derivative as emitter for efficient solution-processed non-doped green OLED." *Dyes Pigm.*, **133**, 380 (2016).
7. X. Zheng, Y. Liu, Y. Zhu, F. Ma, C. Feng, Y. Yu, H. Hu, and F. Li, "F., Efficient inkjet-printed blue OLED with boosted charge transport using host doping for application in pixelated display." *Opt. Mater.*, **101**, 109755 (2020).
8. S. Negi, P. Mittal, and B. Kumar, "Impact of different layers on performance of OLED." *Microsyst*, **24**, 498 (2018).
9. S. Altazin, S. Züfle, E. Knapp, C. Kirsch, T. D. Schmidt, L. Jäger, Y. Noguchi, W. Brütting, and B. Ruhstaller, "Simulation of OLEDs with a polar electron transport layer." *Org. Electron.*, **39**, 244 (2016).
10. S. Negi, P. Mittal, and B. Kumar, "Numerical modeling and parameters extraction of novel triple hole block layer-based organic light-emitting diode for display." *J. Soc. Inf. Disp.*, **28**, 956 (2020).
11. C. H. Gao, X. B. Shi, D. Y. Zhou, L. Zhang, Z. K. Wang, and L. S. Liao, "Highly efficient white organic light-emitting diodes with controllable excitons behavior by a mixed interlayer between fluorescence blue and phosphorescence yellow-emitting layers." *Int. J. Photoenergy.*, **2013** (2013).
12. Y. Chen, J. Chen, D. Ma, D. Yan, L. Wang, and F. Zhu, "High power efficiency tandem organic light-emitting diodes based on bulk heterojunction organic bipolar charge generation layer." *Appl. Phys. Lett.*, **98**, 114 (2011).
13. Y. M. Kim et al., "Enhanced brightness and efficiency of organic light-emitting diodes with an LiF in the Alq/sub." *IEEE Electron Device Lett.*, **27**, 558 (2006).
14. A. Jetly and R. Mehra, "Design of tandem organic light emitting diode using efficient charge generation layer." *Opt. Mater.*, **88**, 304 (2019).
15. D. Zhao, H. Liu, Y. Miao, H. Wang, B. Zhao, Y. Hao, F. Zhu, and B. Xu, "A red tandem organic light-emitting diode based on organic photovoltaic-type charge generation layer." *Org. Electron.*, **32**, 1 (2016).
16. F. A. Angel, J. U. Wallace, and C. W. Tang, "Effect of lithium and silver diffusion in single-stack and tandem OLED devices." *Org. Electron.*, **42**, 102 (2017).
17. H. Sun, Q. Guo, D. Yang, Y. Chen, J. Chen, and D. Ma, "High efficiency tandem organic light emitting diode using an organic heterojunction as the charge generation layer: an investigation into the charge generation model and device performance." *Acs Photonics*, **2**, 271 (2015).
18. S. Hamwi, J. Meyer, M. Kröger, T. Winkler, M. Witte, T. Riedl, A. Kahn, and W. Kowalsky, "The role of transition metal oxides in charge-generation layers for stacked organic light-emitting diodes." *Adv. Func. Mater.*, **20**, 1762 (2010).
19. Y. Liu, X. Wu, Z. Xiao, J. Gao, J. Zhang, H. Rui, X. Lin, N. Zhang, Y. Hua, and S. Yin, "Highly efficient tandem OLED based on C60/rubrene: MoO3 as charge generation layer and LiF/Al as electron injection layer." *Appl. Surf. Sci.*, **413**, 302 (2017).
20. J. H. Kim and S. H. Ju, "Emission characteristics of multi-tandem OLED using MoO x with CGL." *J. Korean Inst. Met. Mater.*, **48**, 105 (2015).
21. J. Yuan, Y. Dai, Q. Sun, X. Qiao, D. Yang, J. Chen, and D. Ma, "High efficiency organic light-emitting diodes based on HAT-CN/TAPC heterojunction charge generation layer as charge injectors." *Semicond. Sci. Technol.*, **34**, 105010 (2019).
22. I. Titov, M. Köpke, N. C. Schneidewind, J. Buhl, Y. Murat, and M. Gerken, "OLED-OPD matrix for sensing on a single flexible substrate." *IEEE Sens. J.*, **20**, 7540 (2020).

Analysing Effectiveness of Different Physiological Biomarkers in Detecting Stress

Anshuman Raj Chauhan
Department of Computer Engineering
Delhi Technological University
Delhi, India
its7arc@gmail.com

Akhil
Department of Computer Engineering
Delhi Technological University
Delhi, India
sah.akhil.01@gmail.com

Sanjay Kumar
Department of Computer Engineering
Delhi Technological University
Delhi, India
sanjay.kumar@dtu.ac.in

2023 IEEE World Conference on Applied Intelligence and Computing (AIC) | 979-8-3503-1006-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/AIC57670.2023.10263957

***Abstract*—Stress is a common part of our everyday life and helps us to handle certain situations in a better way. However, exposure to stress over a long period of time can cause mental and physical health complications. Many health complications associated with stress can be prevented with its early detection. Stress in humans is accompanied by certain physiological changes and monitoring those changes can help in its detection. In this, we study the impact of different physiological changes on stress detection. Different machine learning and deep learning techniques viz. Random Forest, AdaBoost, Decision Trees, Gaussian Naive Bayes and Artificial Neural Networks were used for stress detection using WESAD dataset, which is a dataset for stress and affect detection. Artificial Neural Network outperformed the machine learning techniques and achieved an accuracy of 93.6%. To analyse the effectiveness of each physiological biomarker in detecting stress, the artificial neural network was trained separately on every possible combination of biomarker set. The series of experiments showed that only three physiological biomarkers viz. three axes acceleration, electrodermal activity and body temperature, are sufficient to detect stress effectively.**

***Index Terms*—Stress Detection, Physiological Biomarker, Artificial Neural Network, Deep Learning,**

I. Introduction

Stress can be defined as the psychological and physiological response of the body to physical, emotional or mental threats in the surrounding [1]. Stress results in activation of a cascade of biological responses (physiological changes) [2]. This stress response is a good evolutionary trait that helps us to adapt to dynamically changing environments and allows our bodies to be better prepared for a "do-or-die" situation, for instance encounter with a wild animal. In today's world, stress is prevalent in responsibilities and meeting deadlines. Although, short term stress can act as motivator to perform such tasks, frequent exposure to stressful situations can have poor impact on an individual's mental and physical health. Stress can also have negative impact on immune system of humans and animals and is associated with increased risk of cancer [3]. Such adverse effects of stress on human's health calls for stress management system that can be used to detect the same so that necessary relieving measures can be followed on time.

Conventional method of detecting stress conditions using questionnaire is quite unreliable and uncertain as it depends on individual's responses. To minimize this unreliability, many researchers have tried to detect stress conditions based on the person's physiological parameters as stress is characterised by shifts in various bio-signals caused by release of stress hormones. Such research works leverage different sensor-fitted wearable devices to monitor various physiological changes and use that data to train machine learning or deep learning models to detect stress. Using a number of devices worn on different body parts yields quality data that is good for a laboratory setting but too cumbersome for daily use. Modern smartwatches solves this problem with its compact design and ability to measure a number of physiological biomarkers that are sufficient to detect stress conditions. For this reason, this study focuses only on those physiological biomarkers that are measurable by currently available smartwatches. Such a simple stress monitoring device can be used to identify potentially stress inciting daily activities by providing continuous data on stress levels. This can play an essential role in managing stress levels in individuals.

The objective of this study is to study the effectiveness of various physiological changes in detecting stress using machine learning and deep learning techniques. Only the smartwatch measurable physiological biomarkers are anlaysed in this study. The work uses publicly available WESAD dataset for stress detection. First the performance of different machine learning and deep learning techniques on WESAD dataset for stress detection was compared. The best performing model was later on used for analysing the effectiveness of each physiological biomaker by training that model on every possible set of biomarkers.

The organization of rest of the paper is as follows: Section II comprises of the related works by other researchers in this domain. Section III elaborates about the dataset used in this work. Section IV contains details about the proposed methodology followed by the analysis of the experiments performed in Section V. Finally, the paper ends with Section VI containing the concluding remarks.

II. Related Work

Recent times have seen a push towards better health maintenance citing the detrimental effects stress has on an individual's mental and overall health. There is extensive research in this field and this section discusses about some of the works done in the domain of stress detection using sensor fitted smart

979-8-3503-1006-1/23/$31.00 ©2023 IEEE

DOI: 10.1109/AIC.2023.12
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 06:05:01 UTC from IEEE Xplore. Restrictions apply.

wearable devices. Healy et. al. [4] in 2005 published their work and became first researchers to show that physiological sensors can be used for stress detection. The purpose of their work was to use physiological biomarkers to detect driver stress which can be used to monitor how different road and traffic conditions affect driver's focus. They measured electrocardiogram, electromyogram, skin conductance, and respiration data for multiple drivers on different routes. They were able to achieve 97% accuracy for distinguishing three levels of driver stress using the measured datas. Keshan et al. [5] used data from Healy et. al.'s work to develop a personalized individual stress detection mechanism. They used machine learning algorithms to achieve an accuracy of 88.24% using ECG signals alone. Sharma et. al. [6] used stationary wavelet transform to extract entropy based features of EEG signals and used those features for classification using different supervised machine learning algorithms.

Due to scarcity of data being available in this field, few researches in this field have been solely focused on introducing new datasets. In this regard, Koelstra et. al. [7] introduced a multimodal dataset for affective state analysis. The dataset was created by measuring EEG signals for 33 participants while watching a 40 minute excerpt of music video. Later on, Koldijk et al. [8] presented a multimodal dataset named SWELL-KW for stress and user modelling in 2014. They monitored 25 people using camera to record facial expressions, kinetic 3D sensors to record body postures and wearable body sensors to measure heart rate (variability) and skin conductance while the subjects performed typical knowledge work under stress triggering working conditions. Although these datasets helped researchers to study the correlation between physiological changes and affective states of people, they either lacked in number of physiological biomarkers used or number of affective states considered. This gap was bridged by WESAD dataset introduced by Schmidt et. al. [9] in 2018. They used RespiBAN (a device worn on chest), and Empatica E4 (a device worn on wrist), to measure several physiological biomarkers and considered three affective states viz. state of amusement, neutral state and state of stress in their. This dataset has been leveraged by several researchers in their works [10] [11] [12]. For instance, it was used by Bobade et. al. [13] in their work which assessed the performance of different machine learning and deep learning techniques for stress detection. Garg et al. [14] worked on the similar lines and used WESAD dataset to train several machine learning techniques.

In the past few years, researchers have to tried to leverage the power of complex deep learning techniques. Zhang et. al. [15] proposed a framework using convolutional neural networks (CNN) and bidirectional long short-term memory (BiLSTM) suitable for real-time application that used ECG signal for a window of just 10 sec, unlike previous methods that needed a bigger window. Similar to Healy et. al.'s [4] work, a recent work by Mou et. al. [16] leveraged an attention based CNN and LSTM model for driver stress detection. Using a complex deep learning model they were able to achieve an accuracy of 95.5%. Kumar et. al. proposed a neural network with a number of layers having hierarchical learning capabilities of a CNN. They used a fusion strategy to combine high-level features generated by a hierarchy of deep neural networks. Cho et. al. [17] suggested using transfer learning method involving training of model on a bigger dataset, for which they used a driving dataset, and fine tuning it on smaller dataset, for which they used a mental arithmetic data set. Feng et. al. [18] explored the applicability of few-shot learning algorithm in this field. They proposed usage of a Siamese Neural Network for stress detection using a multimodal data from 72 dating couples.

III. Experimental Dataset

This study uses WESAD dataset [9] which is a publicly available multimodal dataset for stress and affect detection. The dataset was created using physiological features and motion data collected from 15 subjects, which were graduate student with mean age of 27.5 ± 2.4 years, measured using chest and wrist worn devices. The RespiBAN measured three axes acceleration (ACC), respiration rate (RESP), electrocardiogram (ECG), electrodermal activity (EDA), electromyogram (EMG) and body temperature (BT), all sampled at 700 Hz. The Empatica E4 measured body temperature (BT), electrodermal activity (EDA), three axes acceleration (ACC), and blood volume pulse (BVP). The data was gathered for approximately 2 hours during which the subjects were exposed to three different affective states viz. stress, amusement, and relaxed (baseline). The length of duration for which the subjects were subjected to the relaxed (baseline), amused and stressed situation was approximately 20 minutes, 6.5 minutes and 10 minutes respectively.

IV. Methodology

In this section we present our proposed methodology for analysing the effectiveness of different physiological biomarkers in detecting stress. This study focuses on finding a high accuracy model for stress detection and use it to analyse the importance of different physiological biomarkers. We used WESAD dataset to train different machine learning and deep learning models. Necessary pre-processing steps were followed prior to the training process. We analysed the effectiveness of each physiological biomarker by training the highest performing classifier on every possible combination of biomarker set. Fig 1 summarizes the various steps of this study in form of a flowchart. The successive sub-sections discusses about each of these steps in detail.

A. Dataset Pre-processing

WESAD dataset contains bio-signlas sampled at a high frequency. To convert those signals to discrete samples a discretization step was performed. A five second window was used for this task and values within this interval was replaced with their mean value and mapped to median time stamp. This step was followed by a normalisation step to scale the values of each biomarker and bring them to same range.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 06:05:01 UTC from IEEE Xplore. Restrictions apply.

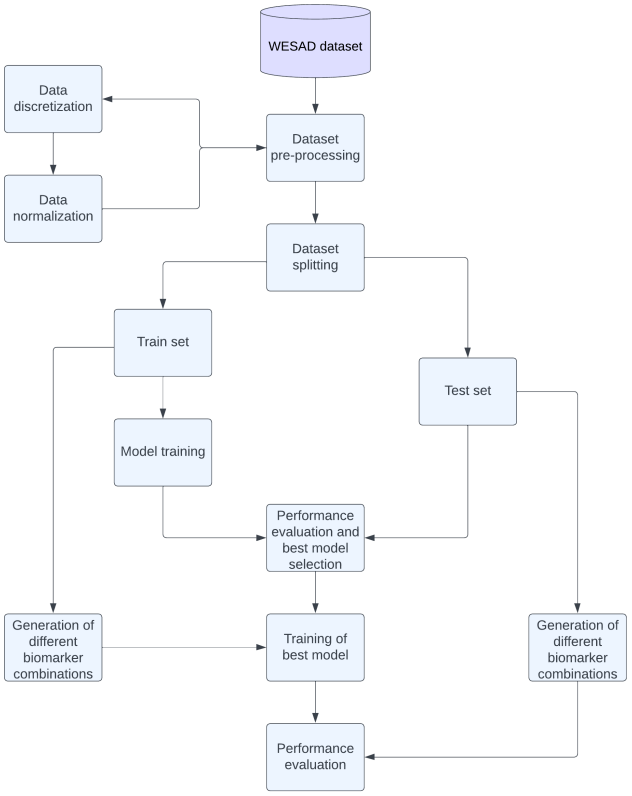


Fig. 1. Schematic flow diagram of Stress Detection Methodology.

B. Classification and Biomarker effectiveness analysis

Stress detection can be viewed as a classification task as it involves predicting the current affective state of an individual using their physiological data. We used 4 different machine learning tecniques viz. Decision Tree (DT), Random Forest (RF), AdaBoost (AB) and Gaussian Naive Bayes (GNB) and an Artificial Neural Network for the classification task. The models were trained with different parameters to obtain the best performing model out of them. Table I contains the description of hyperparameters of all those models.

We used the best performing classifier among all the classifiers for analysing the effectiveness of different physiological biomarkers. A total of 32 different combinations of biomarkers were possible and the classifier was trained on every single one of those sets. To get a more accurate result, the classifier was run on each biomarker set for five times and mean accuracy along with standard deviation for these runs were recorded. The next section contains a detailed discussion on the experiments and their results.

V. Experimental Analysis

A. Comparison of different machine learning and deep learning techniques

The performance of different machine learning and deep learning models is summarised in the table II. It is evident from table II that Gaussian Naive Bayes classifier is the worst

TABLE I
Description of Hyperparameters for different models

Model	Hyperparameter description
Decision trees	
DT-1	Maximum depth = 4
DT-1	Maximum depth = 5
DT-1	Maximum depth = 6
DT-1	Maximum depth = 7
Random Forests	
RF-1	Maximum depth = 4
RF-1	Maximum depth = 5
RF-1	Maximum depth = 6
RF-1	Maximum depth = 7
AdaBoost	
AB-1	Number of estimator = 20
AB-1	Number of estimator = 40
AB-1	Number of estimator = 60
AB-1	Number of estimator = 80
Artificial Neural Networks	
ANN-1	Layer-wise neurons = 6, 5, 4, 1
ANN-2	Layer-wise neurons = 10, 16, 10, 1
ANN-3	Layer-wise neurons = 64, 32, 16, 1
ANN-4	Layer-wise neurons = 32, 64, 32, 1

performer with an accuracy of 46% and F1 score of 48.4%, while artificial neural network (ANN) is the best with an accuracy of 93.6% and F1 score of 92.4%. Although random forest classifier outperformed decision tree classifier for three different set of values of hyperparameters, it was not able to beat decision tree classifier for a particular set of hyperparameter which made decision tree as second best performer. The ANN, as a deep learning technique, outperformed other machine learning techniques with a significant margin because of its ability to handle multi-dimensional data and complex relations in a much easier manner. We also trained the best performing ANN, i.e. ANN-3 on datasets corresponding to the individual subjects, results for which is summarized in table III.

B. Physiological biomarker importance analysis

Table IV shows the performance of ANN-3, the best performing model, on test data for different biomarker set. From the experimental results we concluded that acceleration (ACC) is the most important physiological biomarker as it achieved an accuracy of 64.2% on test data when used alone, which is the highest among all the physiological biomarkers when used alone to train the model. This observation is further supported by the fact that most of the biomarker set with highest accuracy contain ACC biomarker. On the other hand, ECG shows the worst performance when used alone to train the model and gives just 40.4% accuracy. It rather negatively impacts the performance of classifier even when combined with other biomarkers, which is evident from the fact that the best performance was obtained when ECG signal was not present in biomarker set (93.3% accuracy with acceleration, electrodermal activity and body temperature only). EDA is the second most effective biomarker in detecting stress as accuracy of classifier when using EDA alone is next to ACC biomarker when it is used alone. Also this biomarker is a

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 06:05:01 UTC from IEEE Xplore. Restrictions apply.

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT MACHINE LEARNING AND DEEP LEARNING TECHNIQUES

		Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
Decision Trees	DT-1	65.6	65.4	68.5	66.9
	DT-2	73.9	70.2	74.3	72.2
	DT-3	77.9	74.4	78.7	76.5
	DT-4	83.9	81.7	82.7	82.1
Random Forests	RF-1	68.6	65.7	68.5	67.1
	RF-2	75.9	73.8	79.0	76.3
	RF-3	79.3	76.0	80.8	78.3
	RF-4	80.7	77.3	82.1	79.7
AdaBoost	AB-1	67.2	65.8	72.2	68.9
	AB-2	79.4	75.5	79.6	77.5
	AB-3	79.7	76.2	81.0	78.5
	AB-4	78.3	75.4	80.1	77.7
GNB	GNB	46.9	47.5	49.4	48.4
Artificial Neural Networks	ANN-1	71.3	73.6	79.4	76.4
	ANN-2	90.8	87.6	90.2	88.9
	ANN-3	**93.6**	**92.1**	**92.7**	**92.4**
	ANN-4	92.3	90.1	93.5	91.8

TABLE III
PERFORMANCE OF ANN-3 WHEN TRAINED ON SUBJECT-WISE DATASET

Study subject	Accuracy (%)
2	87.2
3	85.0
4	94.3
5	100.0
6	100.0
7	100.0
8	100.0
9	96.7
10	100.0
11	93.4
13	98.9
14	98.9
15	100.0
16	100.0
17	98.9

TABLE IV
PERFORMANCE OF ANN-3 ON DIFFERENT SET OF PHYSIOLOGICAL BIOMARKERS

Biomarker set	Mean Accuracy % (std deviation)
ACC, EDA, BT	93.3 (0.7)
ACC, ECG, EDA, BT	92.7 (3.0)
ACC, EMG, EDA, BT	92.2 (1.3)
ACC, ECG, EMG, EDA, BT	91.5 (2.1)
ACC, EDA	90.0 (2.5)
ACC, EMG, EDA	89.8 (1.8)
ACC, ECG, EDA	88.7 (2.9)
ACC, ECG, EMG, EDA	87.7 (3.7)
ACC, ECG, EMG, BT	79.1 (2.0)
ACC, EMG, BT	77.7 (2.8)
ACC, EMG	70.7 (4.1)
ACC, ECG, BT	70.6 (1.3)
ACC, ECG, EMG	70.4 (3.9)
ACC, BT	70.3 (2.7)
ECG, EDA, BT	69.7 (5.8)
EDA, BT	67.8 (5.2)
ECG, EMG, EDA, BT	67.7 (4.6)
ACC, ECG	64.9 (2.1)
ACC	64.2 (3.7)
EMG, EDA, BT	62.1 (5.4)
EDA	58.8 (1.9)
ECG, EMG, EDA	58.7 (4.4)
EMG, EDA	56.3 (3.8)
ECG, EDA	55.1 (2.6)
ECG, EMG, BT	51.3 (4.5)
EMG	48.0 (4.0)
EMG, BT	47.2 (3.5)
BT	45.7 (5.0)
ECG, BT	45.4 (1.1)
ECG, EMG	43.0 (4.9)
ECG	40.4 (10.6)

part of the biomarker set that achieved the highest accuracy for stress detection. Another important observation made from the results is that although body temperature is second worst performer when used alone for classification task but improves the performance when used with other biomarkers unlike ECG signals and EMG signals that tend to deteriorate the model performance.

VI. CONCLUSION

In this work, we analyzed the performance of various machine learning and deep learning models for the task of stress detection. Then, using the best performing model we analysed the effectiveness and importance of different physiological biomarkers in detecting stress. Based on the findings of the experiments, we concluded that three axes acceleration has the most significant impact on the task of stress detection. Moreover, it was found that electrocardiogram and electromyogram negatively impact the performance of classifiers used for stress detection and just three axes acceleration, electrodermal activity and body temperature are sufficient for detecting stress with high accuracy. The findings of this paper can help researchers to develop efficient models for stress detection that uses minimalist physiological biomarkers without compromising with the detection accuracy.

REFERENCES

[1] J. Choi, B. Ahmed, and R. Gutierrez-Osuna, "Development and evaluation of an ambulatory stress monitor based on wearable sensors," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 2, pp. 279–286, 2012.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 06:05:01 UTC from IEEE Xplore. Restrictions apply.

[2] H. Yaribeygi, Y. Panahi, H. Sahraei, T. Johnston, and A. Sahebkar, "The impact of stress on body function: A review," *EXCLI Journal*, vol. 16, pp. 1057–1072, 07 2017.
[3] T. Yang, Y. Qiao, S. Xiang, W. Li, Y. Gan, and Y. Chen, "Work stress and the risk of cancer: a meta-analysis of observational studies," *International Journal of Cancer*, vol. 144, 11 2018.
[4] J. Healey and R. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 2, pp. 156–166, 2005.
[5] N. Keshan, P. V. Parimi, and I. Bichindaritz, "Machine learning for stress detection from ecg signals in automobile drivers," in *2015 IEEE International Conference on Big Data (Big Data)*, 2015, pp. 2661–2669.
[6] L. D. Sharma, V. K. Bohat, M. Habib, A. M. Al-Zoubi, H. Faris, and I. Aljarah, "Evolutionary inspired approach for mental stress detection using eeg signal," *Expert Systems with Applications*, vol. 197, p. 116634, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417422001233
[7] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis ;using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.
[8] S. Koldijk, M. Sappelli, S. Verberne, M. Neerincx, and W. Kraaij, "The swell knowledge work dataset for stress and user modeling research," 11 2014.
[9] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing wesad, a multimodal dataset for wearable stress and affect detection," 10 2018, pp. 400–408.
[10] K. Nkurikiyeyezu, A. Yokokubo, and G. Lopez, "Effect of person-specific biometrics in improving generic stress predictive models," *Sensors and Materials*, vol. 32, pp. 703–722, 02 2020.
[11] R. Li and Z. Liu, "Stress detection using deep neural networks," *BMC Medical Informatics and Decision Making*, vol. 20, 12 2020.
[12] T. Ninh, S. Smyth, M.-T. Tran, and C. Gurrin, *Analysing the Performance of Stress Detection Models on Consumer-Grade Wearable Devices*, 09 2021.
[13] P. Bobade and M. Vani, "Stress detection with machine learning and deep learning using multimodal physiological data," in *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2020, pp. 51–57.
[14] P. Garg, J. Santhosh, A. Dengel, and S. Ishimaru, "Stress detection by machine learning and wearable sensors," in *26th International Conference on Intelligent User Interfaces - Companion*, ser. IUI '21 Companion. New York, NY, USA: Association for Computing Machinery, 2021, p. 43–45. [Online]. Available: https://doi.org/10.1145/3397482.3450732
[15] P. Zhang, F. Li, R. Zhao, R. Zhou, L. Du, Z. Zhao, X. Chen, and Z. Fang, "Real-time psychological stress detection according to ecg using deep learning," *Applied Sciences*, vol. 11, p. 3838, 04 2021.
[16] L. Mou, C. Zhou, P. Zhao, B. Nakisa, M. N. Rastgoo, R. Jain, and W. Gao, "Driver stress detection via multimodal fusion using attention-based cnn-lstm," *Expert Systems with Applications*, vol. 173, p. 114693, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417421001342
[17] H.-M. Cho, H. Park, S.-Y. Dong, and I. Youn, "Ambulatory and laboratory stress detection based on raw electrocardiogram signals using a convolutional neural network," *Sensors*, vol. 19, p. 4408, 10 2019.
[18] T. Zhang, A. El Ali, A. Hanjalic, and P. Cesar, "Few-shot learning for fine-grained emotion recognition using physiological signals," *IEEE Transactions on Multimedia*, pp. 1–1, 2022.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 06:05:01 UTC from IEEE Xplore. Restrictions apply.

Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-17324-3

Artificial intelligence in cerebral stroke images classification and segmentation: A comprehensive study

Gourav Kumar Sharma[1] · **Santosh Kumar**[1] · **Virender Ranga**[2] · **Mahendra Kumar Murmu**[1]

Received: 23 May 2023 / Revised: 7 August 2023 / Accepted: 28 September 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract
A brain stroke, commonly called as a cerebral vascular accident (CVA) is one of the deadliest diseases across the globe and may lead to various physical impairments or even death. Therefore, timely detection, diagnosis, and treatment of said medical emergency are urgent requirements to minimize life loss, which is not affordable in any sense. Nowadays, with the advancements in Artificial Intelligence, Machine Learning, and Deep Learning, these techniques are popularly used in the domain of medical image analysis in general, and brain stroke imaging in particular, to automate image analysis more accurately and less human error-prone. The aim of this state-of-the-art study is to explore the latest computer-aided diagnosis (CAD) techniques to detect, classify, and segment cerebral strokes. Available methods or techniques, and tools used for analysis have been presented and compared with their attained results. In this study, available open-access datasets in the domain of brain stroke analysis have been explored and presented. Moreover, the research also includes the major challenges and provides researchers with applicable future directions.

Keywords Brain stroke classification · Cerebral stroke segmentation · Artificial intelligence · Stroke lesion · Machine learning · Ischemic stroke · Hemorrhagic stroke · Deep learning

1 Introduction

Brain structure is one of the complex organs of the human body. The brain, along with the spinal cord, is responsible for carrying and disseminating the controlling sensory signals throughout the connected body parts [1]. The main compositional brain parts are the cerebrum, cerebellum, and brain stem [2]. The frontal lobe is mainly responsible for body movement, intelligence, and emotions. The parietal lobe controls vision, hearing, sensing, and language interpretations. The occipital lobe supervises chromatic interpretations, and

✉ Gourav Kumar Sharma
gourav03vatsa@gmail.com

[1] National Institute of Technology, Kurukshetra, India

[2] Delhi Technological University, Delhi, India

the temporal lobe manages organizational and memory-related activities [3]. The cerebrum is the biggest member of the brain and is structured in both hemispherical shapes i.e., left and right. The connections and message-passing mechanisms are like the left hemisphere control right body parts whereas, the right hemisphere supervises the left ones. Left arm or leg or mouth weakness or paralyzed behavior is an indicator of a brain stroke in the left hemisphere and vice-versa [4]. The cerebrum's surface is known as the cortex and it contains around 16 billion neurons and is grey in color. Grey matter, the outer part, is darker whereas, white matter, the inner part, is lighter in color. The cerebellum is the second largest part of the brain, also called the mini-brain, situated below and back side of the cerebrum in the brain's anatomy [5]. The cerebellum is similar to the cortex in the sense that it consists of two hemispheres. The main functions of the cerebellum include muscular movement coordination and body balancing. The brainstem, the middle portion of the brain, provides a link between the cerebrum and the spinal cord. Brainstem parts are the midbrain, pons, and medulla. The midbrain helps in hearing, movement, and visualization. Pons acts as a bridge between the midbrain and medulla, and assists in facial features, chewing, and balancing. The medulla is the bottom-most part of the brainstem and joins with the spinal cord, and helps in vomiting, sneezing, etc. [6]. A typical brain structure is shown in Fig. 1.

A brain stroke also named a cerebral vascular accident (CVA) is a medical emergency, which requires immediate medical examinations or treatment procedures. The obstacle or interruption in blood-flowing vessels to a particular part of the brain is known as ischemic stroke, and its contribution is the most in stroke cases. Whereas, the rupture or disruption of a blood-carrying vessel leads to hemorrhagic stroke. According to the American Heart Association (AHA) statistics, the maximum cases approximately 87% are ischemic strokes whereas, about 17% belong to hemorrhagic strokes [8]. CVA attacks have been observed irrespective of sex, age, or regionality.

With the rapid increase in cerebral strokes worldwide, associated severity, and deadliest nature, efficient and timely treatment is very crucial. According to the estimates published by the world health organization, stroke is the second leading cause of mortality across the globe [9], as shown in Fig. 2. Therefore, timely detection, diagnosis, and treatment of stroke are highly recommended to reduce the chances of long-lasting side effects or even death. The stroke diagnosis, in terms of computerized tomography and magnetic resonance imaging, is tiresome and subject to the availability of medical professionals like neuroradiologists and physicians. And also, identifying the severity of stroke is essential in treatment

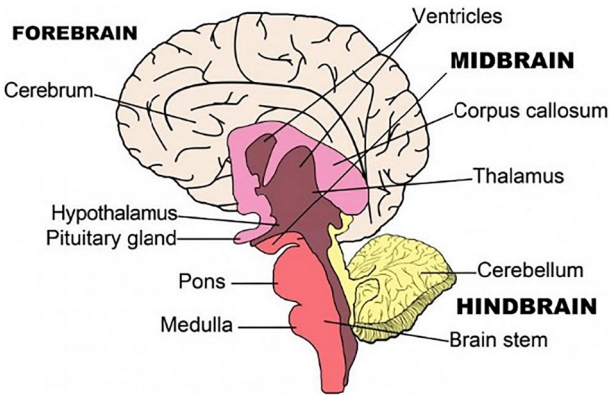


Fig. 1 A typical brain structure [7]

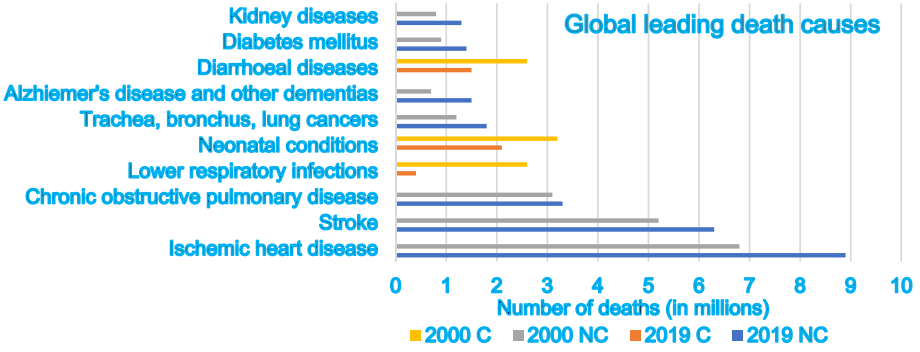


Fig. 2 Top ten mortality rates estimates published by WHO

planning. Due to manual examination error-prone behavior, CAD-based AI-driven automated systems are the need of the hour to efficiently detect, classify, and segment strokes from cerebral images [10]. CAD-based disease diagnosis systems are not meant for the replacement of medical doctors, they just provide aid in clinical practices for automating the treatment procedure.

The methods by which the images of the body parts are generated are known as imaging modalities. Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) are the most general types of imaging modalities that are performed on brain stroke diagnosis. Computed tomography, X-ray-based imaging, is the most common type of imaging that is performed in cerebral stroke [11]. Neurologists aim to diagnose at a low cost, with less imaging time, and make sure availability in most healthcare centers. Non-contrast CT is the very first neuro examination in cerebral ischemic stroke and shows a hypodense structure on CT [12]. Core and penumbra regions of ischemia are detected with the help of CT perfusion imaging with parametric maps such as cerebral blood flow (CBF), cerebral blood volume (CBV), time to peak (TTP), mean transit time (MTP), and time to maximum (T_{max}) which are color-coded for detailing the functioning of the brain [13]. CT angiography is used with the help of contrast injection into the blood-carrying veins through the body parts to identify aneurysms or blockages [14]. Despite the above-mentioned characteristics of CT imaging, it is not able to identify little infarcts and detailed conformation of cerebral hemispheres.

Magnetic resonance imaging (MRI), magnetic field and radio waves-based imaging, are used to generate early infarcts in soft body organs like the brain [11]. This imaging modality is very sensitive and gives the detailed structure of the considered organ. T1 and T2-weighted, fluid attenuation inversion recovery (FLAIR), and diffusion-weighted imaging (DWI) are the various forms of this modality. T1-weighted shows dark whereas T2-weighted shows bright as brain abnormality. In the FLAIR image, the cerebral spinal fluid details are not revealed in contrast to T2-weighted. DWI images are used for the early identification of stroke lesions [12].

An overview of our systematic review process is depicted diagrammatically in Fig. 3. There are many research studies conducted to diagnose brain strokes using AI techniques. However, the potential targets of this study are as follows:

- To explore the advancements in the field of cerebral stroke detection, classification, and segmentation through artificial intelligence.

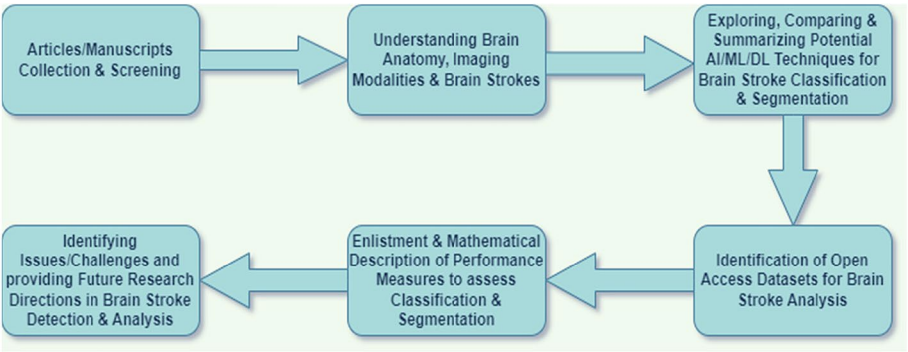


Fig. 3 Overview of systematic review process

- To review and compare the performance of the existing methods in brain stroke diagnosis.
- To present the future research directions in cerebral stroke detection and segmentation, and analyze the key challenges.

1.1 Organization of the rest of the article

The organization of the rest of the paper is as follows. In Sect. 2, we provide materials and methods, AI-assisted CAD-based, utilized in brain stroke classification and segmentation. Images visualization for brain stroke analysis is presented in materials. The literature selection strategy describes how the articles have been searched in different available scientific databases and also a detailed representation of the selection of the articles has been given. Then subsequently artificial intelligence, machine learning, and deep learning techniques have been described. Finally, in the methods subsection, the general procedure of CAD-based stroke diagnosis has been presented. And brain stroke classification, segmentation, and both classification-segmentation methods available in the literature have been reported comparatively and systematically. Section 3, presents a discussion related to benchmark standard datasets, performance measures for classification, and segmentation of brain strokes. Section 5 describes the applicable challenges and future directions for brain stroke analysis. Section 6 presents the limitations of the current review. Finally, the concluding remarks have been presented in Sect. 6.

2 Materials and methods

2.1 Materials

Materials are subjects that are utilized to conduct research. CT and MRI images are important materials in brain stroke analysis research. Visual representations of cerebral strokes on CT and MRI modalities, ICH (Intracerebral Hemorrhage) on CT, ICH on MRI, and various imaging modalities like T1-weighted, T2-weighted, DWI, and FLAIR are shown above in Fig. 4, Fig. 5, Fig. 6 and Fig. 7 respectively.

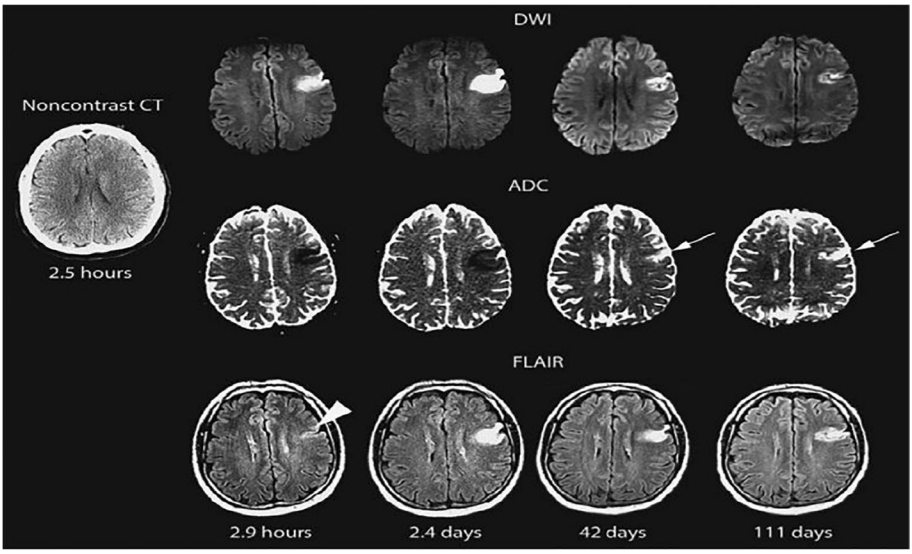


Fig. 4 Visual representation of cerebral strokes on CT, and MRI imaging (DWI—Diffusion-Weighted Imaging, Apparent Diffusion Coefficient, Fluid Attenuated Inversion Recovery) [15]

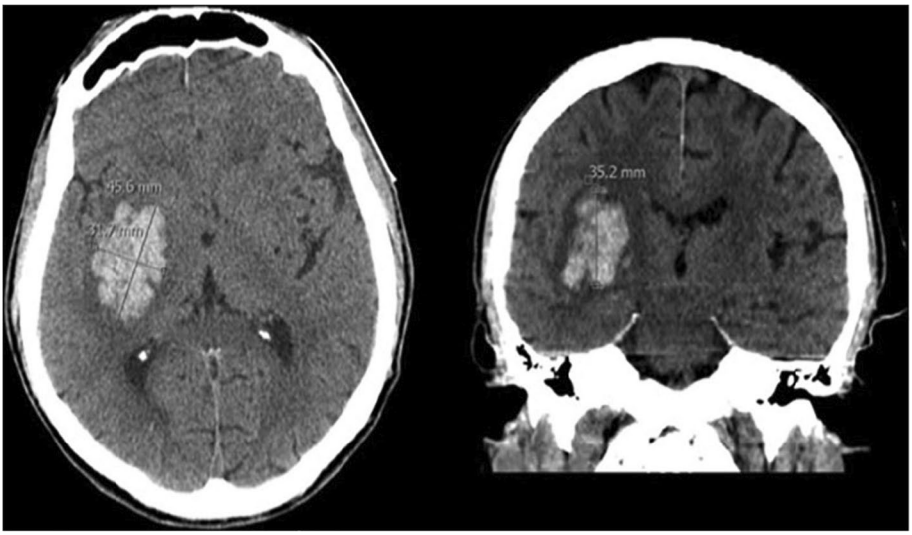


Fig. 5 Visualization of ICH on CT [16]

2.2 Literature selection strategy

There are various research articles have been published in scientific databases such as IEEEXplore™, ScienceDirect™, Pubmed™, and Google Scholar™ in the field of brain stroke detection, and segmentation. We scrutinize the searched articles in a phased

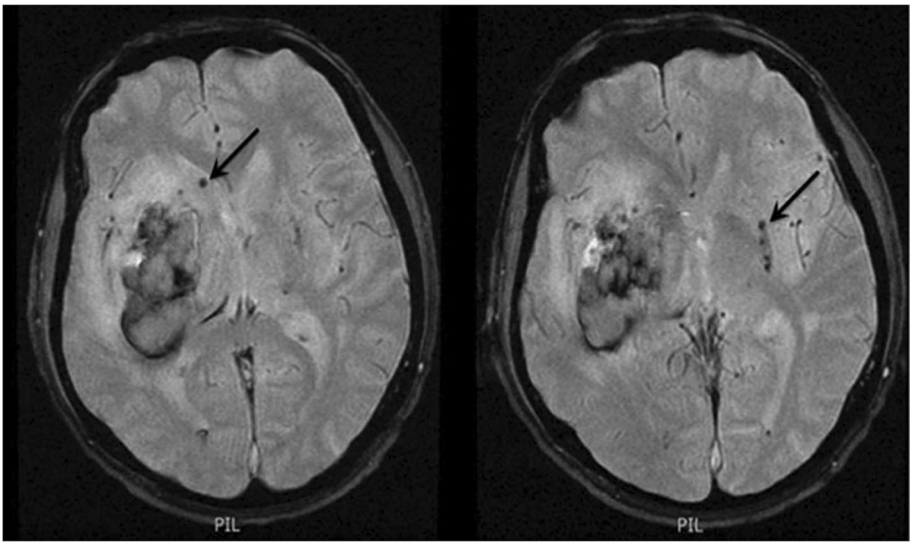


Fig. 6 Visualization of ICH on MRI [16]

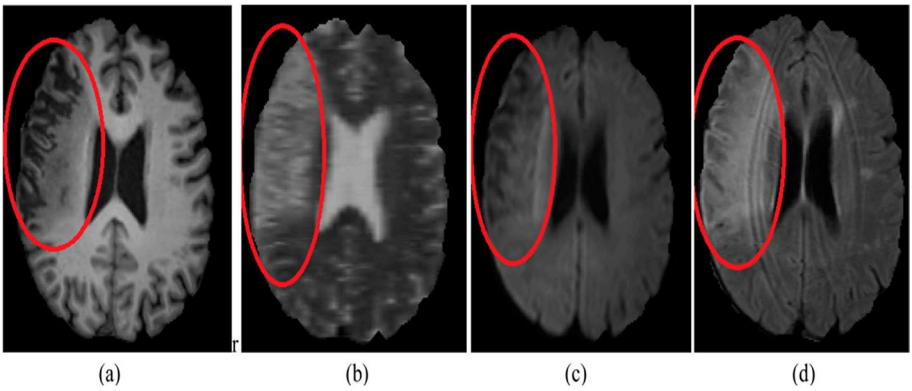


Fig. 7 (a) T1-weighted (b) T2-weighted (c) DWI (d) FLAIR [17]

manner, which is depicted in Fig. 8, shown below. To explore the latest technological advancements, we consider the articles published from 2017 to the present.

Inclusion and exclusion criteria are as follows: The total number of articles downloaded is 619. Three-level filtering is applied to the downloaded articles so that the support articles can be finalized for this study. In first-degree analysis, 347 articles are excluded based on the screening of title, abstract, and year of publication. In second-degree analysis, articles based on artificial intelligence techniques and datasets are filtered and a total of 145 more articles are excluded. Finally, in the last phase of screening only 38 out of the remaining articles have been considered based on full-length review and study.

The popular search key terms considered in scientific databases are 'Brain stroke detection', 'brain stroke classification', 'brain stroke diagnosis', 'Ischemic stroke detection using machine learning', 'Artificial intelligence in brain stroke', 'hemorrhagic stroke detection',

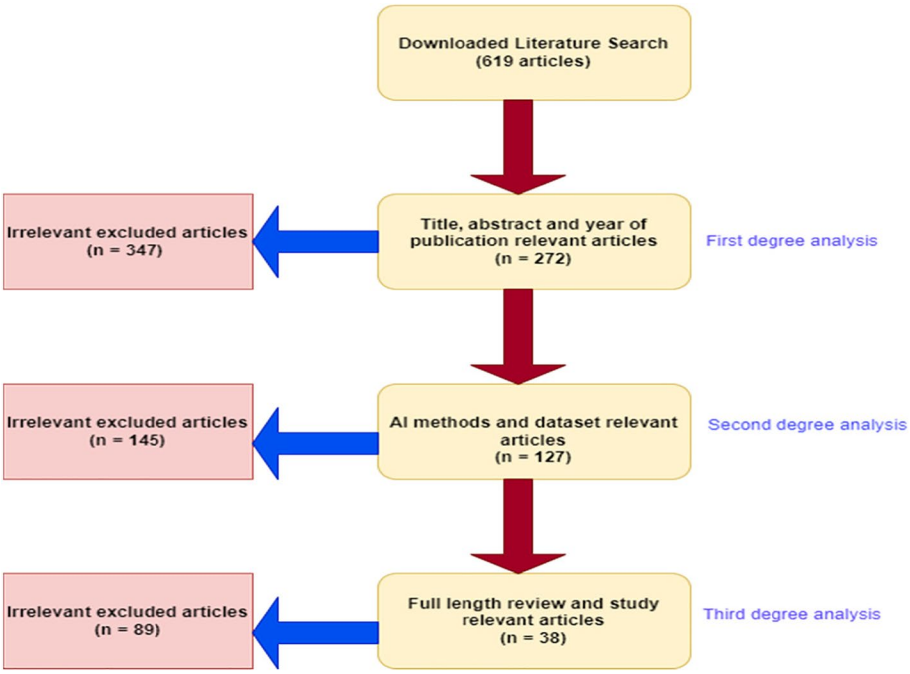


Fig. 8 Literature search and selection strategy

'stroke lesion segmentation'. 'Stroke prediction' and 'Computer-aided diagnoses'. And it took around two months to search these articles across the scientific research databases. All the selected published articles are in English language only. The details regarding research questions imposed before each considered study are given in Table 1, shown above.

2.3 Artificial intelligence (AI)

AI is a machine-level intelligence, using which the machine can behave and mimics human natural intelligence [18]. Machine learning is an AI technique extensively used in medical image diagnosis [19] to recognize and extract useful information, which can be helpful in critical medical applications [20]. The popular machine learning techniques used in brain

Table 1 Research questions imposed on individual study

Q1. Which ML or DL technique was used?
Q2. What kind of cerebral stroke issue was addressed?
Q3. What is the size of the utilized dataset?
Q4. Is image pre-processing done?
Q5. What are the stroke classes if it is a classification problem?
Q6. Which image features are utilized in the study?
Q7. What are the performance measures on which the model's evaluation has been done?
Q8. Are there any clinical interpretations?

stroke image analysis are supervised and unsupervised. In supervised learning, for training the human-labeled data, which is very tiresome in medical imaging, is used in which the actual output is known in advance. The predicted output is as close as the actual one. The supervised learning methods used in medical imaging are prediction and classification [21]. On the other hand, unsupervised learning does not utilize human-labeled data but instead explores its pattern in numerous datasets, which makes this learning very helpful in some novelty explorations like genotypes and phenotypes [22]. In this section, we briefly review some machine-learning strategies.

The Support vector machine (SVM) is a supervised category classification technique that categorizes the given object into one of two classes. It is mainly used in medical imaging to classify diseases. The basic concept behind SVM is to create a dividing hyperplane in high dimension, maximizing the margin hyperplane, soft margin to push the data object to come to the right side of the hyperplane, and kernel functions to project the features into high dimensional space [23]. Logistic regression (LR) is another supervised learning binary classification approach. It is used to associate some independent variables to one binary dependent variable. It uses a probabilistic approach to predict the outcome variable and is being potentially used in medical research [24]. A decision tree (DT) is a supervised category algorithm used for both regressions as well as classification [25]. It makes use of the decision process generated by a tree-like structure, which is maintained by some features in the dataset using a probabilistic and mathematical approach. DT is continuously used in medical research to support clinical decisions [26]. Naive Bayes (NB) classifier is a popular probabilistic algorithm, which relies on the principle of conditional probability. NB makes use of an assumption that the predictors are independent of each other. Liver and heart diseases are the best examples that exploit naive bayes classification strategy [27]. Random forest (RF) is one of the ensemble learning-based classifiers, which heavily relies on the majority decisions of multiple generated decision trees. RF classifier is potentially utilized in medical image analysis such as brain tumor detection [28].

An artificial neural network (ANN) is a parallel processing system inspired by architecture and connections inside a human brain. The ANN architecture is modeled in multiple layers, usually, input, hidden and output layers, connected with dedicated and weighted links. Each layer is made up of multiple processing elements, called neurons, which get activated when a certain condition is satisfied [29]. ANNs are employed in applications like regression and classification. Due to the adaptive self-learning nature and data type flexibility, ANN works well in areas where conventional statistical approaches fail. ANNs are currently utilized in a variety of healthcare applications [30]. Deep learning is one of the latest developments in the domain of artificial intelligence. In contrast to shallow learning, there is no limit to hidden layers in deep learning. Convolutional neural network (CNN) is a DL technique that is mainly used in processing and recognizing patterns in images. Convolutions are applied to produce feature maps of the inputted image. CNNs are extensively used in medical images to generate features without any pre-processing or human intervention. This characteristic makes CNN a powerful tool in medical image processing, segmentation other radiological examination procedures [31]. Autoencoder, an encoder-decoder-based, unsupervised approach primarily used in the extraction and representation of image features in high dimensional space. An example of an autoencoder is U-Net architecture and applications include brain structural abnormality identification [32]. Generative adversarial network (GAN) is more typical than autoencoders and extensively used in image processing tasks like image enhancement, augmentation, and reconstruction of medical images [33]. GAN is usually composed of two networks namely; generative and discriminative networks. The generative network is responsible to generate appropriate

candidates from input data whereas, the discriminative network is used to compare the generated candidate with the real one.

2.4 Methods

The general process of the CAD system in stroke diagnosis is shown in Fig. 9. CAD-based stroke diagnosis framework generally includes the following phases: image acquisition, pre-processing, segmentation, feature extraction, prediction or classification, and then clinical interpretations by medical experts.

A total of 38 studies have been reviewed in this work. Out of 38, 23 studies belong to brain stroke classification, 13 belong to stroke lesion segmentation and 2 are from both classification and segmentation. In the following section, we provide works and studies performed by eminent researchers in brain stroke classification, brain stroke lesion segmentation, and both classification and lesion segmentation.

2.4.1 Brain stroke classification

In this section, investigations conducted by eminent researchers in brain stroke classification, and a total of 23 articles have been included. A summary of artificial intelligence techniques in brain stroke classification [34] along with relevant details is presented in Table 2, shown below.

Pedro P. Reboucas Filho et al., in [35], proposed a novel method to identify and classify brain strokes based on CT images. Images are stored in Digital Imaging in Communication and Medicine (DICOM) format. Region of Interest Segmentation for image smoothing and noise reduction 3*3 median filter was used, and for segmentation thresholding Hounsfield unit was utilized. Feature Extraction Analysis of Brain Tissue Densities was used and compared with five other methods Grey Level Co-Occurrence Matrix, Central Moments, Local Binary Patterns, Statistical Moments, and Zernike's Moments. For classification SVM, Multilayer Perceptron (MLP), Optimum Path Forest (OPF), k-Nearest Neighbour (kNN), and Bayesian classifiers were used to classify the brain into one of the three categories healthy brain or ischemic stroke, or hemorrhagic stroke. An accuracy of 99% was achieved with Analysis of Brain Tissue Densities using an optimal configuration with the

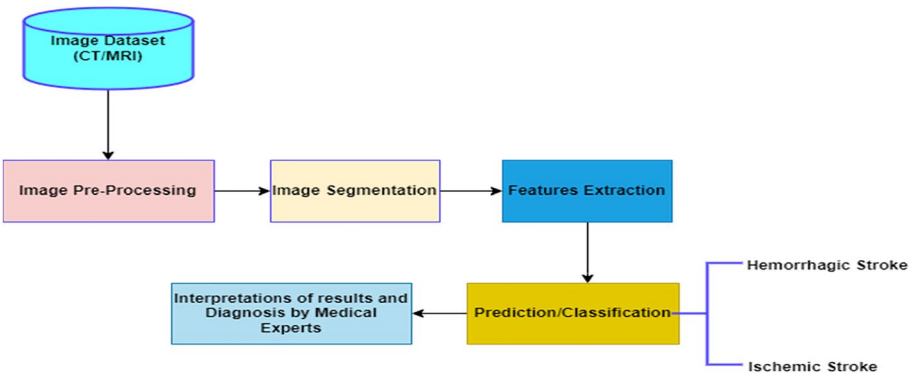


Fig. 9 General process of CAD in stroke diagnosis

Table 2 Summary of ML/DL methods employed in brain stroke classification

S. NO	Authors/Year/Ref	ML/DL Techniques	Imaging Modality/ Samples	Stroke Type	Tool/Library Employed	Performance	Clinical Interpretations
1	Pedro P. Reboucas Filho et al., 2017 [35]	MLP, SVM, kNN, Bayesian and OPF	CT, 420	Ischemic and Hemorrhagic stroke	OpenCV 3.0	Accuracy = 99% Specificity > 92%	Faster and efficient diagnosis of brain stroke detection in clinical practices
2	R. Karthik et al., 2017 [36]	SVM	MRI, 45	Ischemic stroke	MATLAB R2012b	Accuracy = 99.1%	Utilized in ischemic stroke detection in clinical practices
3	Asit Subudhi et al., 2018 [37]	SVM, RF	MRI, 192	Partial anterior circulation syndrome (PACS), Ischemic stroke, Total anterior circulation, and Lacunar Syndrome (LACS) stroke	MATLAB	Sensitivity = 0.92	Implemented to aid neuro-radiologist in efficient ischemic stroke detection
4	U. Rajendra Acharya et al. 2019 [38]	DT, LDA, QDA, kNN, PNN, SVM with tenfold, and SVM with RBF	MRI DWI, 267	Partial anterior circulation syndrome (PACS), Ischemic stroke, Total anterior circulation, and Lacunar Syndrome (LACS) stroke	Not Mentioned	Accuracy = 97.6%	Support in the automatic stroke severity classification
5	Carlos M.J.M. Dourado Jr et al. 2019 [39]	CNN, Bayesian classifier, Multilayer perceptron, kNN, RF, and SVM	CT, 420	Hemorrhagic and Ischemic stroke	Python with Tensorflow and Keras	Accuracy = 100%	IoT-enabled stroke detection systems with less dependent and error-free due to human

Table 2 (continued)

S. NO	Authors/Year/Ref	ML/DL Techniques	Imaging Modality/ Samples	Stroke Type	Tool/Library Employed	Performance	Clinical Interpretations
6	Ajay Patel et al. 2019 [40]	CNN and LSTM (RNN)	NCCT, 1940	Intracranial Hemorrhage	Keras with Theano backend	Area under ROC = 0.96	Treatment and surgical decision in ICH in an emergency
7	King Chung Ho et al. 2019 [41]	CNN, LR, RF, Gradient boosted regression tree, SVM, and step-wise multilinear regression	MRI, 181	Acute ischemic stroke	Python scikit-learn, MATLAB, and Torch7	AUC = 0.765	Utilized in the stroke treatment as a tool for decision support
8	Selladurai Anbumozhi 2019 [42]	k-means clustering and ANFIS	MRI, 75	Ischemic stroke	MATLAB R2014b	Avg accuracy = 99.8% Specificity = 94.0%	Utilized by medical professionals for ischemic stroke detection and classification
9	Anjali Gautam et al. 2019 [43]	cubic SVM, Linear SVM, fine kNN, coarse gaussian SVM, medium kNN, ensemble boosted trees, ensemble bagged trees, and ensemble subspace discriminant	CT, 900	Ischemic and Hemorrhagic stroke	Not mentioned	Accuracy = 84.44%	Utilized by medical professionals in stroke emergencies
10	Francisco F.X. Vasconcelos et al. 2020 [44]	kNN, OPF with euclidean and gaussian distances, SVM with RBF, multilayer perceptron, and Bayes classifier	CT, 473	Ischemic and Hemorrhagic stroke	Python 3.6.5	Accuracy = 98.13%	Assist medical specialists in stroke diagnosis integrated with IoT and edge devices

Table 2 (continued)

S. NO	Authors/Year/Ref	ML/DL Techniques	Imaging Modality/ Samples	Stroke Type	Tool/Library Employed	Performance	Clinical Interpretations
11	Lisa Herzog et al. 2020 [45]	Bayesian CNN	MRI, 511	Ischemic stroke	Python, Keras, and Tensorflow	Accuracy = 95.33%	Automatic classification to help the concerned physician in diagnosis and treatment
12	C. S. S. Anupama et al. 2020 [46]	DCNN	CT, Not Mentioned	Intracerebral hemorrhage	Python-3.6.5, Keras, Tensorflow, Pickle, Numpy, Matplotlib, Pillow, sklearn, and OpenCV	Sensitivity = 94.01%	Utilized in the diagnosis of ICH
13	Anjali Gautam et al. 2021 [47]	CNN	CT, 900	Hemorrhage and Ischemic stroke	MATLAB 2018b	Accuracy = 98.33	Used by physicians in medical practices
14	Yusuf Kenan Cetinoglu et al. 2021 [48]	MobileNetV2 and EfficientNet-B0 CNN	MRI DWI, 421	Ischemic stroke and vascular territorial stroke	Google colaboratory	Accuracy = 96%	Fast and appropriate automatic stroke treatment decisions
15	Yongzhao Xu et al. 2021 [49]	Mask R-CNN and kNN	CT, 356	Hemorrhagic stroke	Python 3	Classification accuracy = 100% Segmentation specificity = 99.93%	Used in clinical practices by healthcare professionals
16	Anjali Gautam et al. 2021 [50]	cubic SVM, Linear SVM, fine kNN, coarse gaussian SVM, weighted kNN, medium kNN, ensemble boosted tree, ensemble bagged tree, and ensemble subspace discriminant classifier	CT, 900	Hemorrhagic and ischemic stroke	Not mentioned	Avg precision = 0.8269	If implemented in CT scanners provide better results to predict stroke at the time of CT screening

Table 2 (continued)

S. NO	Authors/Year/Ref	ML/DL Techniques	Imaging Modality/ Samples	Stroke Type	Tool/Library Employed	Performance	Clinical Interpretations
17	A.S. Neethi et al. 2022 [51]	3D CNN	NCCT, 234	Hemorrhagic and ischemic stroke	Tensorflow, Keras, and OpenCV	F1-score = 0.88	Assist radiologists and doctors in identifying and highlighting abnormalities in the scan
18	Sengul Dogan et al. 2022 [52]	kNN	CT, 2997	Hemorrhagic and ischemic stroke	Matlab 2021b	Accuracy = 98.79%	Used in hospitals, clinics, and emergency services
19	Yung-Ting Chen et al. 2022 [53]	VGG-16, CNN-2, and ResNet-50	CT, 24769	Acute or subacute hemorrhagic stroke	Python (Tensorflow, Keras, and Matplotlib)	Accuracy = 0.9872	Provides radiologists a classification, which can quickly generate a warning to clinicians
20	Burak Tasci et al. 2022 [54]	Resnet101, DenseNet201, SVM, INCA, and IMV	Diffusion MRI, 2010 in dataset-1 and 444 in dataset-2	Ischemic acute stroke	Matlab 2021b	Accuracy = 93.23%	Used to aid in the decision process of physicians to early diagnose Ischemic stroke
21	M.Sait Din et al. 2022 [55]	AlexNet and SVM	CT, 9818	Intraparenchymal hemorrhage, Epidural hematoma, Subdural hematoma, and Subarachnoid hemorrhage	Matlab 2020b	Accuracy = 97.47%	Used in emergency services to provide a prompt diagnosis of a brain hemorrhage

Table 2 (continued)

S. NO	Authors/Year/Ref	ML/DL Techniques	Imaging Modality/ Samples	Stroke Type	Tool/Library Employed	Performance	Clinical Interpretations
22	Miguel Altuve et al. 2022 [56]	ResNet-18	CT, 200	Intracerebral hemorrhage	Matlab R2021a	Avg accuracy = 95.93%	Assist physicians in their medical decision in terms of resource optimization and reduce the time and error in ICH diagnosis
23	B. Deepa et al. 2022 [57]	HSVFC	MRI, 500	Acute and sub-acute stroke	Matlab 2019b	F-score = 0.99	Diagnosis of Brain Stroke at early stages

ROC, Receiver operator characteristic AUC, Area under the curve.

OPF classifier. The proposed model can contribute to a faster, and more efficient diagnosis of brain stroke detection in clinical practices.

In [36], R. Karthik et al. presented an automated approach for the detection, and classification of ischemic stroke using discrete curvelet transformation and support vector machine based on MRI images. Image pre-processing has been done in terms of skull removal and watershed segmentation has been applied for localizing the region of interest. Discrete curvelet transform has been utilized for the extraction of statistical features. A statistical t-test has been conducted to select the significant features. Then classification was done with the help of a support vector machine with Radial Basis Function (RBF) kernel into normal and abnormal classes. The accuracy of the presented method was 99.1%. The proposed method can be utilized in ischemic stroke detection in clinical practices.

To segment and classify ischemic stroke lesions in MR DWI sequences, Asit Subudhi et al. [37], proposed a novel hybrid framework based on Delaunay triangulation and fractional order Darwinian particle swarm optimization (FODPSO) for the detection of ischemic stroke. Firstly, all slices were denoised with wiener filter edge sharpening has been performed. The image segmentation was done with the help of Delaunay triangulation, which works on pixel intensity and characteristic similarity. Then the segmented image was optimized by using FODPSO to get efficient results. Finally, the ischemic stroke classification was done by SVM and Random Forest (RF), and the best result was obtained with a sensitivity of 0.92, and a dice coefficient of 0.93 using a random forest. The limitation of the proposed technique was not able to calculate the lesion's volume. The proposed method can be implemented to aid neuro-radiologist in efficient ischemic stroke detection.

U. Rajendra Acharya et al., in [38], proposed a novel method to identify the severity of ischemic stroke and its types based on MR images. To balance the dataset, adaptive synthetic sampling was employed to reduce the biases of the underlying data distribution. Image pre-processing was done in two steps i.e., histogram equalization and random transformation. For image enhancement ad contrast incrementation, histogram equalization was used. For directional signatures capturing and preserving intensity differentiation, random transform was utilized. Higher-order spectra and bispectrum entropies were utilized to extract the features. Analysis of variance (ANOVA) resulting F-value has been used for features ranking. The classification was done with the named ML techniques; Linear discriminant analysis, Decision tree, Quadratic discriminant analysis, k-NN, Probabilistic neural network, SVM with ten-fold strategy, and SVM classifier with RBF. The proposed method by using SVM with RBF kernel achieved the best results and an accuracy of 97.6%. The shortcoming of the presented study is not the inclusion of the blindfold validation system. The proposed system can support the automatic stroke severity classification.

In [39], Carlos M.J.M. Dourado Jr et al. proposed an IoT-enabled framework for the classification of ischemic stroke, hemorrhagic stroke, and normal brain on CT images. Convolutional Neural Network (CNN), a deep neural network, was utilized for the extraction of features. Stroke classification was done with the use of classification algorithms such as multilayer perceptron, Bayesian classifier, random forest, k-nearest neighbour, and support vector machine. For remote stroke diagnosis and patient monitoring, IoT enabled online system was developed. CNN showed 100% accuracy with almost all the classifiers, whereas less training and testing time was with the use of the Bayesian classifier. Automatic new node identification in IoT configuration in stroke is a limitation of the proposed study. The proposed IoT-enabled stroke detection system can be used as less dependent and error-free due to humans.

To identify and classify ICH, in [40], Ajay Patel et al. presented a novel approach based on 3D non-contrast CT images. The first part of the proposed network consists

of pre-trained CNN, which is used to identify ICH from 2D axial slices. Subsequently, sequential information, based on classified slice level, has been analyzed with the employment of long short-term memory (LSTM) which is a recurrent neural network (RNN). The proposed method attained 0.96 as an area under the receiver operating characteristic curve. Limitations of the proposed work include contextual information loss due to the application of maximum filter on the input image, training of both CNN and LSTM and last limited network configurations. The proposed technique can help in the treatment and surgical decisions in ICH in an emergency.

In [41], King Chung Ho et al. proposed a machine learning strategy to classify time since stroke (TSS) with the use of MR imaging sequences. Imaging features have been generated in three parts; Region of interest (ROI) generation, feature generation related to baseline imaging, and deep feature generation. The hemispheric region in which Tmax > 6 s was selected as the stroke ROI mask. Two baseline features, descriptive and morphological features, were generated with a total of 104 features. Descriptive was statistical and morphological were shapes-related features. Perfusion-weighted imaging (PWI) hidden features have been extracted using a deep autoencoder-based technique. Random Forest, Logistic regression, Support vector machine, Gradient boosted regression tree, and Stepwise multilinear regression ML techniques were employed to classify Time since Stroke (TSS) < 4.5 h or TSS > 4.5 h. Linear regression classifier achieved the highest Area under Curve (AUC) of 0.765 for the proposed model of TSS classification. The proposed ML model was only trained using MR images rather than CT images. Clinical variables such as age are not considered in this study, which may improve TSS classification. The proposed methodology can be utilized in stroke treatment as a tool for decision support.

Selladurai Anbumozhi, [42], proposed a method to diagnose and detect ischemic stroke based on MR images. Noise from the input MR image was removed by using a directional filtering algorithm. Oriented local histogram equalization (OLHE) technique utilized for brain MR image enhancement. Skull removal and brain tissue extraction were done with the help of a morphological opening and closing strategy. Grey Level Co-occurrence Matrix (GLCM), local derivative pattern, and local tetra pattern were utilized for feature extraction. The k-means clustering technique was used for the segmentation and cluster formation and relationship analysis among clusters. Adaptive neuro-fuzzy inference system (ANFIS), for stroke classification, utilized in this study. Average classification accuracy of 99.8% was achieved in this study. This proposed method can be utilized by medical professionals for ischemic stroke detection and classification.

Anjali Gautam et al., [43], proposed a novel method based on feature extraction for the categorization of CT into ischemic, hemorrhagic, and normal images. To extract meaningful features from the images' gradient of gradient, double gradient computation of the image, named the local gradient of gradient pattern was utilized. Nine experiments with distinct classifiers and comparisons with different feature descriptors have been performed to evaluate the proposed method's performance. Nine distinct classifiers i.e., cubic SVM, linear SVM, fine kNN, coarse Gaussian SVM, weighted kNN, medium kNN, ensemble boosted trees, ensemble bagged trees, and ensemble subspace discriminant has been employed to assess the proposed strategy. The best results were attained with the use of fine kNN with both a five-fold and ten-fold cross-validation approach with accuracies of 84.44% and 86.11% respectively. The computational time of the proposed model is higher as compared to other models. The proposed model can be utilized by medical professionals in stroke emergencies.

Francisco F.X. Vasconcelos et al., in [44], presented a novel method to extract brain CT image features using adaptive brain tissue density analysis (adaptive ABTD) for the

classification and detection of stroke. To pre-process the images and removal of the skull, regions between 0–120 Hounsfield unit (HU) were selected and multiple Otsu's thresholds were applied. The inclusion criterion of adaptive ABTD was that it analyzes the image on a pixel basis and checks the pixel's neighbourhood to decide the density region of the examined pixel. Thereafter the image was divided into two hemispheres. And, based on the percentage density of the two hemispheres the feature vector was created. Finally, the presented method was capable enough to embed in the edge devices to extract full features. The model was assessed on different classifiers but the best performance was achieved with kernel size 3 in adaptive ABTD using an SVM classifier with the RBF kernel. The best F1 score and accuracy of the presented method were 97.83% and 98.13% using SVM with RBF kernel. To assist the medical specialist in stroke diagnosis the proposed system can be integrated with IoT and edge devices.

An ischemic stroke diagnosis paradigm using MRI data and Bayesian uncertainty was given by Lisa Herzog et al. in [45]. A Bayesian CNN was used to identify ischemic stroke lesions on 2D MR images, and advanced aggregation techniques were used to correlate image predictions with patient diagnoses. First, a Bayesian model (1Ch-MC) provided uncertainty estimates along with improved prediction accuracy when compared to its non-Bayesian counterpart (1Ch-NoMC). After that, Bayesian CNN was modified to take into account the unique 3D data structure of the MR pictures (3Ch-MC). The most basic baseline aggregation approach combines patient-level diagnosis with image-level forecasts by selecting the patient's image with the highest overall stroke likelihood. The authors went on to point out how applying the suggested neural network-based aggregation approaches improved prediction performance. These techniques took into account the uncertainty in picture predictions and offered a patient-level estimate of uncertainty. An accuracy score of 95.33% was achieved by the proposed method. The proposed system can be used for automatic classification to help the concerned physician in diagnosis and treatment.

In [46], C. S. S. Anupama et al. presented a GrabCut-based synergic deep learning (GC-SDL) model for the segmentation and classification of ICH on CT images. To remove noise and improve the quality of the image Gabor filtering was utilized. To identify the disease-affected portion GrabCut-based segmentation method was used. For feature extraction synergic deep learning (SDL) model was used and, finally, classification was done with the help of the softmax (SM) layer. The presented model was assessed on CT images from a benchmark dataset collected from physionet and attained a specificity of 97.78%. the proposed model can be utilized in the diagnosis of ICH.

To classify brain stroke efficiently into ischemic and hemorrhagic is a crucial task. Anjali Gautam et al., in [47], presented an efficient categorization of brain CT images into hemorrhagic, ischemic, and normal using CNN. The input image to the network was pre-processed using quadtree-based image fusion. Contrasting the region of stroke has been with image fusion. Thereafter, new CNN architecture has been used to classify CT-based images into two (hemorrhagic and ischemic) and three (ischemic, hemorrhagic, and normal). The obtained classification accuracies on dataset-1 were 98.33% and 98.77%, whereas on dataset-2 were 92.22% and 93.33% respectively. The proposed method can be utilized by physicians in medical practices.

Yusuf Kenan Cetinoglu et al., in [48], proposed an approach for the classification and detection of vascular territorial stroke using CNN models. MRI-DWI was used as an imaging modality. MobileNetV2 and EfficientNetB0 were utilized in this study. The transfer learning approach has been used for both stroke detection as well as vascular territorial stroke classification. MobileNetV2 and EfficientNetB0 CNN architectures, pre-trained on the ImageNet dataset were used. The proposed model achieved accuracies of 96% and 93%

using MobileNetV2 and EfficientNetB0 CNN architectures respectively. Limitations of the study include the Trial of ORG 10172 in acute stroke treatment (TOAST) is not conducted and images being obtained with a single MR device. This proposed architecture can be used to automate the stroke detection process and help the physician make a quick and suitable treatment plan.

In [49], Yongzhao Xu et al. introduced a fully automatic and innovative system based on the health of things. The proposed system was capable enough to classify the skull CT images into hemorrhage and normal images. Deep learning methods have been utilized. 14 models have been employed for the classification. Thereafter stroke segmentation was done with the use of mask Recurrent-CNN combined with ML methods. The foremost segmentation result was achieved with mask R-CNN and kNN. The segmentation responsiveness of the proposed model was 4 s, which shows the reliability of the stated method. The accuracies obtained were 100% and 99.73% in cases of classification and segmentation respectively. The presented model can be used in clinical practices by healthcare professionals.

To categorize brain strokes into hemorrhagic, ischemic, and normal, Anjali Gautam et al. [50] proposed a new feature extraction-based method. Local neighbor pattern (LNP), a variant of the local binary pattern, was employed for feature extraction. The reason to use LNP is its extraction capability of quantum information from the neighborhood pixel of the center pixel. Image classification was done with tenfold cross-validation. The proposed model was implemented using nine different classifiers namely; cubic SVM, linear SVM, fine kNN, coarse Gaussian SVM, weighted kNN, medium kNN, ensemble boosted tree, ensemble bagged tree, and ensemble subspace discriminant classifier. The model has been assessed on real CT images. Foremost classification results were obtained with an average precision of 0.8269. If the proposed system is implemented in a CT scanner it can provide superior results to predict stroke at the time of CT screening.

Non-contrast CT images are composed of 3D image slices stack or volumetric images. The 2D CNN-based models are not capable enough to handle this volumetric structure. To address this issue, A.S. Neethi et al., in [51], proposed a 3D fully CNN-based classification model to classify the images in one of the classes i.e., ischemic, hemorrhagic, and normal. The stride slicing augmentation method was used to enhance the CT scan frames and limit the class imbalance problem. A block-wise design has been proposed. In the first phase, dimensionality adjustments with feature retainment were done. And, in the next block combined feature maps were learned with skip connections and bundled convolutions. The presented model achieved an F1 score of 0.88. The limitations include more parameters that have to be trained and more memory required as compared to 2D models. This system can be utilized by radiologists and doctors to identify abnormalities in brain scans.

Sengul Dogan et al., [52], presented an ML-based hand-crafted technique for classifying ischemic and hemorrhagic stroke. The proposal included preprocessing, extracting features, selecting features, and classification. In the preprocessing phase segmentation of the brain area and image resizing were done. A novel P9 function and local phase quantization (LPQ) was utilized for the creation of multilevel features at low and high levels. Moreover, to decrease the time burden of the classifier Chi2 selector was used for selecting the optimal feature vector. Finally, the classification was done by a shallow classifier i.e., kNN with the use of a Bayesian optimizer to choose the best hyperparameters. CT scans were utilized for model assessment and an accuracy score of 98.79% was attained. The limitation of the proposed model is that more CT scan images/classes can be collected. The proposed model can be utilized in hospitals, clinics, and emergency services.

Convolutional neural network-based deep learning models for efficient categorization of brain strokes were employed in the study conducted by Yung-Ting Chen et al.

[53]. Pre-trained Visual Geometry Group (VGG)-16, CNN-2, and Residual Network (ResNet)-50 have been used in this study. And, thereafter unenhanced CT images were classified into acute or sub-acute hemorrhage, normal, acute infarction, and other categories. The proposed study has been compared with the other studies in the diagnosis of hemorrhagic and ischemic stroke. The ResNet-50 and CNN-2 overperformed VGG-16 with an accuracy score of 0.9872, whereas ResNet-50 took more time to produce the output than the other networks. This presented model can be utilized by clinicians and radiologists to quickly issue warnings related to cerebral stroke.

In the direction of deep feature-based methods, Burak Tasci et al., [54], proposed a method to classify the MRI images as acute ischemic stroke and normal. An architecture, using 19 pre-trained CNNs, was utilized to select the best model for feature extraction. Then based on the accuracies achieved Resnet101 and DenseNet201 were chosen for the extraction task. GradCAM (Gradient-based classroom activation mapping), canny edge detection, histogram matching, and locally interpretable model-agnostic explanations (LIME) techniques were utilized for the pre-processing of the datasets. Features were created using Resnet101 and DenseNet201. Most important features have been selected with the iterative neighborhood component analysis (INCA) method. Selected feature classification was done by an SVM classifier. The iterative majority voting (IMV) technique was applied to the obtained results for the highest performance. Classification accuracies obtained by the presented system were 97.56%, 99.32%, and 99.16% on datasets 1, 2, and 3 respectively. Less number of images used and the time burden are the limitations of this study. The presented system can be utilized to aid in the decision-making steps of physicians to diagnose ischemic stroke in the early stages.

M. Sait Din et al., in [55], presented a hybrid of both deep and hand-crafted features-based automatic brain hemorrhage classification models. Deep feature extraction was done with a pre-trained AlexNet deep model. Hand-modeled features have been generated from each exemplar with the implementation of the local phase quantization and local binary pattern. After merging both the extracted features, neighborhood component analysis (NCA), also known as feature selection variation of kNN, was utilized. Finally, the SVM classifier was used to classify the CT images into five cerebral hemorrhage classes namely intraparenchymal hemorrhage, epidural hematoma, subdural hematoma, subarachnoid hemorrhage, and healthy. The proposed model attained an accuracy of 97.47% using ten-fold cross-validation. The unbalanced dataset and AlexNet, which is trained on a big medical database, usage are the limitations of the proposed study. This system can be deployed in emergency services to provide a prompt diagnosis of a cerebral hemorrhage.

The research conducted by Miguel Altuve et al., in [56], utilized ResNet-18 (a deep residual CNN) for the categorization of CT images into normal and intracerebral hemorrhage (ICH). To improve clinical practices and computational cost, raw CT images were utilized without any pre-processing. The transfer learning concept was employed with only 18 layers in the residual network for detection. For proper visualization and understandability by a medical person, the Grad-class activation mapping (Grad-CAM) technique was adopted. The proposed model scored an average accuracy of 95.93%. The limitation of this study was that the validation was done on a smaller and balanced dataset i.e., 200 CT images (100 normal and ICH each). The model can assist the physician in terms of better decision-making in diagnosis with reduced time and error.

In [57], B. Deepa et al. presented a method for the categorization of abnormalities in the brain i.e. tumor and stroke with the use of hybridized ML technique based on MRI images. Feature extraction was performed in the form of texture, intensity, and shape. Feature related to textures was extracted by Neoteric directional-based quantized extreme

pattern (NDQEP). Intensity-based clustering wavelet transforms (ICWT) were utilized for the extraction of intensity features. Feature selection was employed with the usage of the Maximum a posteriori (MAP) based firefly algorithm (FFA). Moreover, the image categorization task was done by the hybridized support vector-based forest classifier (HSVFC). MRI image modality is utilized. The proposed model scored 99.2% accuracy for brain stroke categorization. The proposed model is not suitable for a large database with different pathological conditions is the limitation of the study. The proposed system can be utilized to help radiologists and physicians in the early diagnosis of brain strokes.

2.4.2 Brain stroke lesion segmentation

The section describes brain stroke lesion segmentation investigations performed by various researchers. A total of 13 studies have been discussed. A summary related to stroke lesion segmentation AI-enabled methods is summarised in Table 3, shown below.

In [58], P. Sivakumar et al. introduced a CAD strategy for the segmentation and detection of ischemic stroke diagnosis in cerebral MRI images with adaptive neuro-fuzzy inference system (ANFIS) classifier. The heuristic histogram equalization technique was utilized to enhance the input MRI image. Conversion of images from the spatial domain to multi-resolution was done with discrete curvelet transform, which breaks down the image into discrete and low pass sub-band. The genetic algorithm is used for the selection of features, which are generated from curvelet transform image coefficients. Then, ANFIS is utilized as a classification method for classifying brain abnormalities. Finally, the normalized graph cut algorithm is used for ischemic stroke segmentation. The proposed model achieved 98.76% accuracy. The proposed method can be utilized by neuro-radiologists and physicians in medical practices for ischemic stroke detection and treatment.

Rongzhao Zhang et al., [59], Presented a novel automated technique for the segmentation of acute ischemic stroke using 3D CNN based on diffusion-weighted imaging MR images. DenseNet is utilized for boosting the information and gradient flow in the network. To address the class imbalance issue, the dice objective function employed and 3D fully connected DenseNet for the segmentation of acute ischemic stroke. Multiple experiments were conducted using the DWI and ADC datasets to validate the proposed system. The proposed method also utilized Ischemic Stroke Lesion Segmentation (ISLES) 2015 for validation. The proposed 3D FCDenseNet scored 79.13 as the dice similarity coefficient. Limitations of the proposed method include not being robust if the MR images are corrupted with motion artifacts and not being able to distinguish true lesions and artifacts in some cases. The presented system can be utilized for automatic acute ischemic segmentation in clinical practices.

Anjali Gautam et al., in [60], Proposed a novel semi-automatic method to segment ischemic stroke lesions efficiently from magnetic resonance images. To enhance the stroke lesion and denoise the image, the wavelet-based pre-processing technique is employed for optimal parameter selection. Segmentation has been done with the help of two approaches; random forest and histogram-based thresholding. Insight Segmentation and Registration Toolkit- Segmentation by Assigning Labels to Pixels (ITK-SNAP) [71] software has been utilized as the first segmentation strategy in histogram-based thresholding. To achieve better segmentation results random forest method is used as another segmentation strategy. The random forest method outperformed histogram-based thresholding with an average 0.67 dice coefficient. The proposed model can help doctors in stroke lesion segmentation more accurately.

Table 3 Summary of ML/DL methods employed in brain stroke lesion segmentation

S. No	Authors/Year/Ref	ML/DL Techniques	Imaging Modality/ Samples	Stroke Type	Tool/Library Employed	Performance	Clinical Interpretations
1	P. Sivakumar et al. 2017 [58]	ANFIS	MRI, 100	Ischemic stroke	MATLAB R2014b	Accuracy = 98.76%	Utilized by neuro-radiologists and physicians in medical practices for ischemic stroke detection and treatment
2	Rongzhao Zhang et al. 2018 [59]	3D CNN and DenseNet	DWI and ADC MRI, 242	Acute ischemic stroke	Pytorch	DSC = 79.13	Utilized for automatic acute ischemic segmentation in clinical practices
3	Anjali Gautam et al. 2018 [60]	RF	FLAIR MRI, 28	Sub-acute ischemic stroke	MATLAB 2015a	Avg DSC = 0.67	Help doctors in stroke lesion segmentation more accurately
4	B. Deepa et al. 2019 [61]	IFT	MRI, Not mentioned	Not mentioned	Matlab 2013a	DSC = 0.99	Utilized for the detection of tumors and strokes in clinical practices
5	Liangliang Liu et al. 2019 [62]	Res-CNN	MRI, 30	Acute ischemic stroke	Python and Keras	Avg DSC = 74.20% HD = 2.33 mm	Used as a useful tool for acute ischemic stroke segmentation in clinical practices
6	Mizuho Nishio et al. 2020 [63]	YOLOv3 VGG16	CT, 6114	Acute ischemic stroke	Keras with Tensorflow	Sensitivity = 37.3%	Utilized by radiologists to improve the clinical outcomes of AIS detection

Table 3 (continued)

S. No	Authors/Year/Ref	ML/DL Techniques	Imaging Modality/ Samples	Stroke Type	Tool/Library Employed	Performance	Clinical Interpretations
7	Sanaz Nazari-Farsani et al. 2020 [64]	SVM with linear kernel	T1 and Diffusion-weighted MRI, 192	Acute ischemic stroke	MATLAB R2016b	Mean DSC = 0.50 ± 0.21	Computationally efficient and can be implemented on a normal computer in medical settings
8	Lu Li et al. 2021 [65]	U-Net	CT, 236	Hemorrhagic stroke	Pytorch	DC = 0.8033	Used in clinical practices for hemorrhagic lesion diagnosis
9	Long Zhang et al. 2021 [66]	SFCN	CT and MRI, Not Mentioned	Ischemic stroke	Pytorch 1.2.0	F1-score = 0.75	Help clinicians in the diagnosis of stroke lesions
10	Hyunkwang Shin et al. 2022 [67]	U-Net	T1-weighted MRI, 229	Not mentioned	Keras with Tensorflow	DC = 59.2%	Helps in diagnosing the prognosis of a stroke patient and improves rehabilitation treatment
11	Shuqiang Wang et al. 2022 [68]	U-Net and GAN	T1-weighted MRI, 239	Not mentioned	Pytorch	DC = 0.617	Automated computer-aided brain stroke segmentation
12	Luca Tomasetti et al. 2022 [69]	SVM, DT and RF	CTP, 152	Ischemic core and penumbra	Not mentioned	Avg DC for penumbra = 0.68 Avg DC for core = 0.26	Correct visualization of the ischemic regions will assist in a better treatment plan
13	Qiqi Bao et al. 2022 [70]	U-Net	T1-weighted MRI, 229	Not mentioned	Pytorch	DC = 0.5662	Automated computer-aided segmentation method and stroke rehabilitation

ADC, Apparent diffusion coefficient; CTP, Computed tomography perfusion; DSC, Dice similarity coefficient; HD, Hausdorff distance; DC, Dice coefficient.

B. Deepa et al., in [61], developed o fusion supported segmentation approach for the detection of strokes and tumors on MR images. MRI sequences included T2-weighted, T1-weighted, fluid-attenuated inversion recovery, and diffusion-weighted imaging. An image fusion technique Gradient-based discrete wavelet transform (GDWT), is utilized in this study. All six combinations of the three utilized MRI sequences were used in image fusion. Histogram equalization method employed for image quality improvement. Finally, the intensity factorized thresholding (IFT) method, based on pixel intensities, has been used for the segmentation of stroke and tumor-affected regions in MRIs. The best results were obtained under the tumor and stroke segmentation when T2 weighted and FLAIR, and DWI and FLAIR fusions usage respectively. A dice similarity coefficient of 0.99 was obtained for stroke segmentation with the use of the DWI and FLAIR fusion methods. The proposed model can be utilized for the diagnosis of tumors and strokes in clinical practices.

Liangliang Liu et al., in [62], presented a novel deep learning method, deep CNN (Res-CNN), to segment lesions of acute ischemic stroke with multi-modality MRI sequences. Image pre-processing is done in several steps: Firstly, to convert the Digital Imaging and Communications in Medicine (DICOM) format image into Neuroimaging Informatics Technology Initiative (NIfTI) format Statistical Parametric Mapping (SPM)-12 software is used. Secondly, image registration i.e., T1 and T2 modalities were co-registered, using a method i.e., six-parameter rigid body registration, to the DWI images. Third skull stripping from MRI sequences by MRIcron software. Finally, image smoothening has been done by the Gaussian method. T2 image is fused with the DWI image to generate an enhanced DWI-T2 modality on which the lesion can be easily seen. Due to the data-hunger behaviour of deep learning models, the data augmentation technique has been adopted to generate a large dataset. A novel network i.e., Res-CNN employed for lesion segmentation in acute ischemic stroke. The average dice coefficient and Hausdorff distance of the proposed method were 74.20% and 2.33 mm respectively. Small lesion segmentation is difficult using the proposed method, the main limitation of this study. The proposed model can be used as a useful tool for acute ischemic stroke segmentation in clinical practices.

For the detection of acute ischemic stroke (AIS) based on non-contrast CT images, in [63], Mizuho Nishio et al. developed a two-stage deep learning model. Images have been pre-processed based on the brain window center and window width setup. And, pixel value normalization has been done. Two deep-learning models have been employed in this work. The first one is You Only Look Once v3 (YOLOv3) for the detection of AIS and the second was VGG16 CNN, a classification model, used to reduce false positive rate (FPR). Sensitivities obtained were 37.3%, 33.3%, and 41.3% for the two-stage model, radiologist without software and radiologist with software respectively. The limitations include fewer radiologists being engaged in this study and external validation is not performed. The proposed method can be utilized by radiologists to improve the clinical outcomes of AIS detection.

In the direction of automatic segmentation and classification of acute ischemic stroke (AIS), Sanaz Nazari-Farsani et al., in [64] presented a fully automated method based on MR images. SMP12, running on MATLAB R2016b, is utilized in the image pre-processing phase. Lesion segmentation has been done with the use of a modified t-test known as Crawford-Howell. For the classification of images, SVM is employed with a linear kernel and nested cross-validation is used for parameter optimization. The mean dice similarity coefficient and sensitivity of the proposed method were 0.50 ± 0.21 and 84% respectively. Limitations of the study include lesion masks were not created by optimization settings and a lot of inaccurate segmentations being found in the lesions' borders. The proposed technique is computationally efficient and can be implemented on a normal computer in medical settings.

To segment and detect hemorrhagic strokes from brain CT images, in [65] Lu Li et al. developed a U-Net-based deep learning framework. Input to the network was original and flipped CT slice for maintaining the symmetric structure and contrast enhancement. At the same time augmented data, used for data balance, was also provided as input. U-Net is utilized to detect and segment the images and for the preservation of information on lesion characteristics. U-Net segmentation network utilized GAN generator to make a pixel-level prediction of target classes. The discriminator network took the ground truth and the segmented result of U-Net as input to predict the probable results. The model achieved an accuracy of 0.9859 for location detection and 0.8033 as the dice score for segmentation. The limitation of the presented work is that small lesions cannot be detected. And, the model can be used in clinical practices for hemorrhagic lesion detection.

To detect ischemic stroke lesions from the CT images, Long Zhang et al., [66] presented a symmetric comparison framework. In this study, a label transfer module is used for the normalization of CT images and obtained ground truth with the help of image pairing, calibration, and registration. To locate and identify the ischemic lesion, a Symmetric feature comparison network (SFCN) is employed. The developed SFCN is utilized to extract image block-level features and differentiate them in the features of the image blocks. The local and the global feature map will then be concatenated and sent to the output. Image enhancement is used to amplify the difference between healthy tissue and lesion one. The F1 score by the presented model is 0.75. This method is not effective when used in rare symmetric lesions. The model could be used to help clinicians in diagnosis.

To reduce the trainable parameters in U-Net, Hyunkwang Shin et al. [67] presented an efficient U-Net architecture for the segmentation of chronic stroke lesions. For the reduction of trainable parameters, a depth-wise, convolution-based e-block was utilized to reduce the trainable parameters. For capturing the global features between the encoder and decoder, a global feature attention block (GA-block) is employed. GA-block's main functionality was to restrain long-term contextual information for improved segmentation performance. The proposed model has assessed on T1-weighted MR images Anatomical Tracings of Lesions After Stroke (ATLAS) dataset. The dice coefficient achieved by the model was 59.2% when it is used with GA-block and e-block. The limitation of the presented model is the loss of spatial information about the lesions' location. The proposed model can be utilized in prognosing stroke patients and improving rehabilitation treatment.

In [68], Shuqiang Wang et al. presented a novel approach, a consistent perception generative adversarial network (CPGAN). The presented method is Semi-supervised and, used in brain stroke lesion segmentation. In this work, they employed three neural networks namely segmentation, discriminator, and assistant networks. To improve the segmentation results, down-sampled encoder and up-sampling decoder-based U-Net architecture were utilized in the segmentation network. Further, the segmentation network is composed of U-Net and skip connections with a similarity connection module (SCM). The main purpose of using the SCM was to capture sensitive information and multi-scaled features. To ensure the non-forgetting behavior of classes, due to non-stationary online learning, the discriminator network is assisted by an assistant network. The assistant network is pre-trained and the architecture is also the same as that of the discriminator. Their proposed method is assessed on MR images ATLAS dataset. The experimental outcomes show that the proposed model performed better in segmentation with only two-fifth of the labeled samples outperforming the other fully labeled samples. The proposed system can be used for automated segmentation in clinical practices.

For detecting and segmenting penumbra and ischemic core regions in acute ischemic stroke patients, Luca Tomasetti et al. in [69] proposed a single and two-step method.

Computed tomography perfusion (CTP) scans were utilized in the study. To generate color-coded parametric maps such as CBF, CBV, TTP, MTP, and T_{max}, Syngo.via is used. Superpixel, a simple linear iterative clustering algorithm's modified version, is applied for generating parametric maps' superpixel regions. The synthetic minority over-sampling technique (SMOTE), for handling the classes with a minor number of occurrences, was included in the study. Random forest, decision tree, and support vector machine ML techniques were implemented for ischemic region prediction. Parametric maps and their generated superpixel were utilized as input features. As far as the results are concerned, the best results were obtained with random forest and single-step approach. Average dice coefficients for penumbra and core were 0.68 and 0.26 respectively, whereas average volume differences for penumbra and core were 25.1 ml and 7.8 ml respectively. The limitation of the proposed technique is that the results for the detection of ischemic penumbra and core in non-large vessel occlusion are not good enough. The presented system can be used to correctly visualize ischemic regions and will guide physicians and radiologists to make better treatment decisions.

In [70], Qiqi Boa et al. proposed an MDAN; Mirror Difference Aware Network for brain stroke lesion segmentation. They utilized U-Net, an encoder-decoder-based CNN with a siamese structure in the encoded path extension. Differential feature augmentation (DFA) consisted of two parts: a Siamese contrastive supervision module and a Mirror position-based difference augmentation module. DFA is employed in the encoding phase among multiple layers for the comparison and amplification of features of the provided two input images i.e., the original MR image and the flipped one. For shrinking the distance among features of the normal region and enhancing the distance among lesion-affected region features, the Siamese constructive supervision module is used. The purpose of the Mirror position-based differentiation (MDA) module is to ensure the attention and encouragement of the network toward abnormalities. To embed the flipped and original image features at the decoder end, the Mirror feature fusion (MFF) component is employed. Finally, the presented model is assessed using two datasets namely; Kaggle MRI and ATLAS. The attained dice coefficients were 0.5662 and 0.6897 for ATLAS and Kaggle datasets respectively. And they also enlisted that operations like alignment and targeted warping can be further utilized for non-pathological asymmetries reduction. The proposed system can be utilized for automated computer segmentation and stroke rehabilitation.

2.4.3 Brain stroke classification and segmentation

In this section, AI-based investigations have been presented for both stroke classification and segmentation. The summary of these techniques is summarised in Table 4, shown below.

In [72], Asit Subudhi et al. proposed an automated system to segment and classify ischemic brain stroke using random forest and expectation–maximization based on diffusion-weighted imaging (DWI) sequences of MRIs. Expectation maximization is used for the segmentation of the stroke-affected region of the brain. Then an optimization algorithm, fractional-order Darwinian particle swarm optimization employed to improve the detection accuracy. Morphological and statistical information containing features were extracted using the Gray-level co-occurrence matrix (GLCM). Finally, the stroke classification is done with random forest and SVM based on the extracted features. The proposed method found the best results using the Random Forest classifier. The dice similarity coefficient and accuracy obtained were 0.94 and 93.4% for stroke segmentation and

Table 4 Summary of ML/DL methods employed in both stroke classification and lesion segmentation

S. No	Authors/Year/Ref	ML/DL Techniques	Imaging Modality/ Samples	Stroke Type	Tool/Library Employed	Performance	Clinical Interpretations
1	Asit Subudhi et al. 2020 [72]	SVM, RF	DWI MRI, 192	Ischemic stroke	Not mentioned	DC = 0.94 Accuracy = 93.4%	Used in the ischemic stroke treatment decision-making process
2	Sercan Yalçın et al. 2022 [73]	U-Net	CT, 6650	Ischemic and hemorrhagic stroke	Python (Tensorflow and Keras)	Accuracy = 98.5% Avg IoU = 95.2%	Utilized by medical doctors and healthcare professionals in hospitals for stroke detection

classification respectively. The proposed method can be used in the ischemic stroke treatment decision-making process.

In [73], Sercan Yalçın et al. presented an optimized dimensional U-Net, encoder-decoder-based, deep learning method for classifying and segmenting brain stroke. The proposed model is composed of convolution, pooling, and dimensional transformation blocks. To extract the valuable features from these blocks pre-processing, standardization and normalization techniques have been applied. The study was performed on real CT image datasets. The proposed method achieved 98.5% accuracy on ischemic and hemorrhage classification and an average Intersection over Union (IoU) of 95.2% on segmentation. Based on their experiments' results the method is better than the other competing methods such as ResNet50v2, MobileNetv2, standard U-Net, and ResDenseUNet. Determination of IoU values and minimal boundary box during the segmentation of brain stroke were the limitations of the presented model. This model can be extremely helpful in hospitals for doctors and other healthcare professionals.

3 Discussion

We have performed an extensive study and analyzed the work done in the field of cerebral stroke detection and diagnosis. The year-wise articles selection in brain stroke classification, segmentation, and both classification-segmentation are illustrated in Fig. 10. Various aspects related to stroke classification and segmentation have been explored and examined. In the following section, dataset characterization, performance measures, challenges, future research directions, and limitations of the study have been presented.

3.1 Dataset characterization

The present section describes the characterization and details related to datasets, taken from medical institutions or benchmark standards, used by researchers. The briefing of the benchmark dataset is illustrated in Table 5.

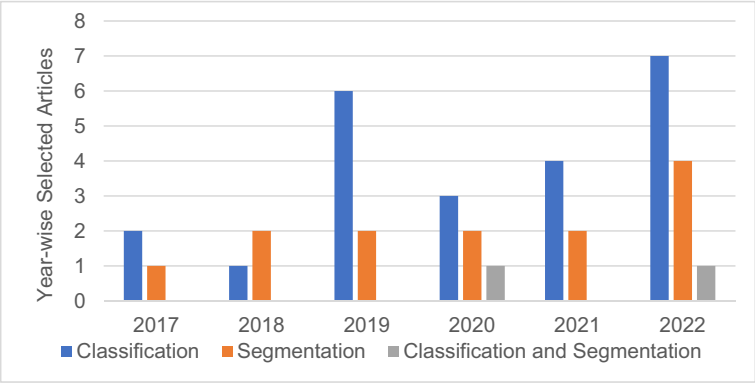


Fig. 10 Illustration of the year-wise studies selection

Table 5 List of open access datasets for brain stroke analysis

S. NO	Dataset	Summary	Ref
1	BraTS	Brain tumor segmentation dataset is a multimodal brain tumor image segmentation benchmark	[74]
2	OpenNeuro	Open platform contains MRI, PET, MEG, EEG, and iEEG data	[75]
3	Physionet	Open access platform provides physiological signals and other imaging modalities	[76]
4	ISLES challenge 2015, 2016, 2017, 2018, 2022	Contains various ischemic stroke lesion segmentation modalities like DWI, ADC, T1, T2, and FLAIR	[77]
5	ATLAS (Anatomical Tracings of Lesions After Stroke)	Contains T1-Weighted MRI scans	[78]
6	ADNI (Alzheimer's Disease Neuroimaging Imitative)	Contains MRI and PET images for analysis	[79]
7	Oasis (Open Access Series of Imaging Studies)	Freely available dataset for the scientific community	[80]

3.2 Performance measures for stroke classification and lesion segmentation

To evaluate and assess the performance of classification and segmentation methods, various standard measures are considered. The listing of the measures is done, shown below in Table 6, and details along with their mathematical equations are covered in this section.

Key Terms used in mathematical notations:

TP : It is the actual positive value predicted positive.
TN : It is the actual negative value predicted negative.
FP : It is the actual negative value predicted as positive.
FN : It is the actual positive value predicted as negative.

3.2.1 Classification accuracy

Classification accuracy is defined as the proportion of the number of correct predictions to the total number of predictions. Mathematically describes [81] as in Eq. (1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where TP, FP, FN, and TN are true positives, false positives, false negatives, and true negatives respectively.

3.2.2 Precision

Precision is defined as the ratio of the number of correctly predicted positive samples to the total positive predicted values. Mathematically described [81] as Eq. (2).

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

where TP and FP represent true positive and false positive respectively.

Table 6 List of performance measures to assess classification and segmentation

Performance Metrics
Classification Accuracy
Precision
Recall or Sensitivity
Specificity
F-Score
Negative Predictive Value
Dice Similarity Coefficient
Average Symmetric Surface Distance
Hausdorff Distance
Intersection over Union

3.2.3 Recall or sensitivity

Recall is defined as the proportion of the number of correctly predicted positive samples to the actual total number of positive samples. Mathematically denoted [81] as Eq. (3).

$$RecallorSensitivity = \frac{TP}{TP + FN} \tag{3}$$

where TP and FN indicate true positive and false negative respectively.

3.2.4 Specificity

Specificity is defined as the proportion of the number of correctly predicted negative samples to the actual total negative predicted samples. Mathematically described [81] as Eq. (4).

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

where TN and FP are true negative and false positive respectively.

3.2.5 F-Score

F-score of a classification model represents its accuracy on a specific dataset and is defined as the harmonic mean of the model's recall and precision. Mathematically depicted [81] as Eq. (5).

$$F - Score = \frac{2 \times precision \times recall}{precision + recall} \tag{5}$$

where precision and recall are described in Eq. (2) and Eq. (3) respectively.

3.2.6 Negative predictive value (NPV)

Negative predictive value is the proportion of the number of correctly predicted negative samples to the total negative predicted samples. Mathematically denoted [81] as Eq. (6).

$$NegativePredictiveValue = \frac{TN}{TN + FN} \tag{6}$$

where FN and TN represent false negative and true negative respectively.

3.2.7 Dice similarity coefficient (DSC)

Dice similarity coefficient or simply dice coefficient is a similarity measurement method between two data sets. It is a very important measure in image segmentation. The mathematical equation of DSC [82] is given in Eq. (7).

$$DSC = \frac{2|X| \cap |Y|}{|X| + |Y|} \tag{7}$$

where X and Y are two volume voxels sets.

3.2.8 Average symmetric surface distance (ASSD)

Average symmetric surface distance (ASSD) is described as the average of all distances measured from segmented region boundary points to the ground truth boundary points. Let 'X' and 'Y' be the two surface point sets then ASSD can be described [83] with the formula in Eq. (8).

$$ASSD(X, Y) = \frac{ASD(X, Y) + ASD(Y, X)}{2} \tag{8}$$

where ASD is the average surface distance given by Eq. (9).

$$ASD(X, Y) = \frac{\sum x \in X^{min} y \in Y^{d(x,y)}}{|X|} \tag{9}$$

as d (x, y) is the Euclidean distance between x and y points.

3.2.9 Hausdorff distance (HD)

Hausdorff distance (HD) is the largest distance between the points of two surface regions and is used to denote outliers. Mathematically denoted [84] in Eq. (10).

$$HD(X, Y) = \max\{max_{x \in X} min_{y \in Y} d(x, y), max_{y \in Y} min_{x \in X} d(y, x)\} \tag{10}$$

where d(.) depicts Euclidean distance.

3.2.10 Intersection over union (IoU)

Intersection over union also known as Jaccard Index is an important measure in segmentation and is determined as a fraction of the area overlapping between predicted and ground truth to the union area between predicted and ground truth. Mathematically depicted [85] as Eq. (11).

$$IoU = \frac{X \cap Y}{X \cup Y} \tag{11}$$

where X and Y are voxel sets for segmentation and ground truth respectively.

4 Challenges and future directions

There are remarkable issues and challenges associated with CAD-based AI-enabled cerebral stroke detection and analysis. Some general issues observed by CAD systems in stroke analysis are the following:

- Benchmark dataset availability is a major challenge in automated brain stroke analysis. The referential databases, where disease-wise neuroimaging modalities along with their examined reports and segmented regions of interest are done by medical professionals are of great importance. Several reviewed studies are performed on primary data which is directly taken from medical institutions. Making such acquired

data in a standard analysis format is also a tiresome task. Hence, there is also a need to develop some standard and medically verified platforms, where researchers can investigate and come up with efficient and advanced techniques in stroke diseases.

- Neuroimaging data annotation is one of the challenging tasks in developing efficient CAD-based stroke analysis systems. There is also a limited collaboration between medical professionals and academic researchers due to ethical issues, which include not revealing the patient's identity. By working together medical specialists and academic researchers can improve the quality of stroke image analysis and provide quality treatment in emergency conditions.
- There is very limited work done on the transient ischemic attack (TIA) also known as mini-stroke. If TIA can be identified timely, which is reversible, then it can prevent some severe ischemia, hemorrhage or even death. Similarly, investigations in the direction of ischemic core detection and segmentation are also very less. In this said areas a lot of research work needs to be carried out so that a qualitative and preventive treatment can be assured.
- Data privacy is also of utmost crucial when we consider medical image analysis. This leads to limits on data availability.
- Data-hungry behavior of deep networks and at the same time, the limited publicly available standard brain imaging datasets lead to problematic situations. Although, some advanced techniques like generative adversarial networks, data augmentation, and transfer learning concepts are very popular recently to address these situations. Still, there is scope for improvement in terms of hyperparameter optimization so that qualitative results can be achieved with less computational time.

5 Limitations of the study

We reviewed a total of 38 studies. Based on that, the following are the limitations of the current review work.

- More search teams can be introduced during the literature search, thereby might explore other researchers' contributions to decision-making CAD-based stroke support systems. Moreover, our research only focuses on the classification and segmentation of brain strokes. Therefore, other stroke aspects are not explored fully.
- Only English-language-published articles have been considered for this work. The inclusion of other languages investigations may improve the quality of the research study.
- Other post-stroke and lesion-related aspects are out of the scope of our research work. However, this may improve the quality of the overall research outcomes.

6 Conclusion

In this paper, an extensive study of different AI-assisted brain stroke classifications and lesion segmentation has been conducted. A summarization of recent AI-assisted techniques in brain stroke and their clinical implications has been presented. Moreover, the identification of publicly available datasets, performance measures in cerebral stroke classification and segmentation, issues identification, key challenges, and future research directions in

stroke medical image analysis are the potential contribution of this study. Digital image processing techniques and artificial intelligence collectively emerge as advanced technological solutions and provide promising results in medical image analysis. Further, our state-of-the-art may be used as an important reference to diagnose cerebral stroke using timely technologies, minimizes risks, and save lives.

Funding No funding has been received from any person or organization to carry out this research.

Data availability Data sharing does not apply to this article as no datasets were generated or analyzed during the current study.

Declarations

Conflicts of interests There is no conflict of interest related to this research.

Competing interests 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. The Central Nervous System: Structure and Function - Per Brodal - Google Books. https://books.google.co.in/books?hl=en&lr=&id=WdATFQ0YUrMC&oi=fnd&pg=PR7&dq=brain+structure+and+nervous+system&ots=TVt9ofJzVV&sig=fjPE2T798Jbr07ksqNluorZXGIw&redir_esc=y#v=onepage&q&f=false. Accessed 3 Nov 2022
2. Calhoun V (2022) Data-driven approaches for identifying links between brain structure and function in health and disease. Dialogues Clin Neurosci 20:87–100. https://doi.org/10.31887/DCNS.2018.20.2/VCALHOUN
3. Casillo SM, Luy DD, Goldschmidt E (2020) A History of the Lobes of the Brain. World Neurosurg 134:353–360. https://doi.org/10.1016/J.WNEU.2019.10.155
4. Bontempi D, Benini S, Signoroni A et al (2020) CEREBRUM: a fast and fully-volumetric Convolutional Encoder-decodeR for weakly-supervised sEgmentation of BRain strUctures from out-of-the-scanner MRI. Med Image Anal 62:101688. https://doi.org/10.1016/J.MEDIA.2020.101688
5. King M, Hernandez-Castillo CR, Poldrack RA et al (2019) (2019) Functional boundaries in the human cerebellum revealed by a multi-domain task battery. Nat Neurosci 228(22):1371–1378. https://doi.org/10.1038/s41593-019-0436-x
6. Sclocco R, Beissner F, Bianciardi M et al (2018) Challenges and opportunities for brainstem neuroimaging with ultrahigh field MRI. Neuroimage 168:412–426. https://doi.org/10.1016/J.NEUROIMAGE.2017.02.052
7. Brain Structure And Function | Brain Injury | British Columbia. https://www.nbia.ca/brain-structure-function/. Accessed 11 Dec 2022
8. Tsao CW, Aday AW, Almarzooq ZI et al (2022) Heart Disease and Stroke Statistics-2022 Update: A Report from the American Heart Association. Circulation 145:E153–E639
9. World Health Organization (2014) The top 10 causes of death. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death. Accessed 17 Oct 2022
10. Inamdar MA, Raghavendra U, Gudigar A et al (2021) A Review on Computer Aided Diagnosis of Acute Brain Stroke. Sensors 21(8507):21–8507. https://doi.org/10.3390/S21248507
11. Vilela P, Rowley HA (2017) Brain ischemia: CT and MRI techniques in acute ischemic stroke. Eur J Radiol 96:162–172. https://doi.org/10.1016/J.EJRAD.2017.08.014
12 Karthik R, Menaka R (2017) Computer-aided detection and characterization of stroke lesion – a short review on the current state-of-the art methods. Imaging Sci J 66:1–22. https://doi.org/10.1080/13682199.2017.1370879
13. Demeestere J, Wouters A, Christensen S, et al (2020) Review of Perfusion Imaging in Acute Ischemic Stroke. Stroke 1017–1024https://doi.org/10.1161/STROKEAHA.119.028337

14. Meijs M, Meijer FJA, Prokop M et al (2020) Image-level detection of arterial occlusions in 4D-CTA of acute stroke patients using deep learning. Med Image Anal 66:101810. https://doi.org/10.1016/J.MEDIA.2020.101810
15. Karthik R, Menaka R, Johnson A, Anand S (2020) Neuroimaging and deep learning for brain stroke detection - A review of recent advancements and future prospects. Comput Methods Programs Biomed 197:105728. https://doi.org/10.1016/J.CMPB.2020.105728
16. Macellari F, Paciaroni M, Agnelli G, Caso V (2014) Neuroimaging in intracerebral hemorrhage. Stroke 45:903–908
17. ISLES: Ischemic Stroke Lesion Segmentation Challenge 2022. http://www.isles-challenge.org/. Accessed 17 Oct 2022
18. Dick S (2019) Artificial Intelligence. Harvard Data Sci Rev 1:. https://doi.org/10.1162/99608F92.92FE150C
19. Alzubi JA, Kumar A, Alzubi OA, Manikandan R (2019) Efficient approaches for prediction of brain tumor using machine learning techniques. Indian J Public Heal Res Dev 10:267–272. https://doi.org/10.5958/0976-5506.2019.00298.5
20. de Bruijne M (2016) Machine learning approaches in medical image analysis: From detection to diagnosis. Med Image Anal 33:94–97. https://doi.org/10.1016/J.MEDIA.2016.06.032
21. Erickson BJ, Korfiatis P, Akkus Z, Kline TL (2017) Machine Learning for Medical Imaging. Radiographics 37:505. https://doi.org/10.1148/RG.2017160130
22. Libbrecht MW (2015) Noble WS (2015) Machine learning applications in genetics and genomics. Nat Rev Genet 166(16):321–332. https://doi.org/10.1038/nrg3920
23. Noble WS (2006) What is a support vector machine? Nat Biotechnol 24:1565–1567
24. Schober P, Vetter TR (2021) Logistic Regression in Medical Research. Anesth Analg 132:365. https://doi.org/10.1213/ANE.0000000000005247
25. Alzubi OA, Alzubi JA, Alweshah M et al (2020) An optimal pruning algorithm of classifier ensembles: dynamic programming approach. Neural Comput Appl 32:16091–16107. https://doi.org/10.1007/S00521-020-04761-6
26. Chen T, Shang C, Su P et al (2021) A Decision Tree-Initialised Neuro-fuzzy Approach for Clinical Decision Support. Artif Intell Med 111:101986. https://doi.org/10.1016/J.ARTMED.2020.101986
27. Wickramasinghe I, Kalutarage H (2021) Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. Soft Comput 25:2277–2293. https://doi.org/10.1007/S00500-020-05297-6/FIGURES/2
28. Thayumanavan M, Ramasamy A (2021) An efficient approach for brain tumor detection and segmentation in MR brain images using random forest classifier, 266–274https://doi.org/10.1177/1063293X211010542
29. Movassagh AA, Alzubi JA, Gheisari M et al (2023) Artificial neural networks training algorithm integrating invasive weed optimization with differential evolutionary model. J Ambient Intell Humaniz Comput 14:6017–6025. https://doi.org/10.1007/s12652-020-02623-6
30. Shahid N, Rappon T, Berta W (2019) Applications of artificial neural networks in health care organizational decision-making: A scoping review. PLoS ONE 14:e0212356
31. Mutasa S, Sun S, Ha R (2021) Understanding artificial intelligence based radiology studies: CNN architecture. Clin Imaging 80:72–76. https://doi.org/10.1016/J.CLINIMAG.2021.06.033
32. Chen M, Shi X, Zhang Y et al (2017) Deep Feature Learning for Medical Image Analysis with Convolutional Autoencoder Neural Network. IEEE Trans Big Data 7:750–758. https://doi.org/10.1109/TBDATA.2017.2717439
33. Yi X, Walia E, Babyn P (2019) Generative adversarial network in medical imaging: A review. Med Image Anal 58:101552. https://doi.org/10.1016/J.MEDIA.2019.101552
34. Alzubi OA, Alzubi JAA, Tedmori S, Rashaideh H, Almomani O (2018) Consensus-based combining method for classifier ensembles. Int Arab J Inf Technol 15(1):76–86
35. Rebouças Filho PP, Sarmento RM, Holanda GB, de Alencar LD (2017) New approach to detect and classify stroke in skull CT images via analysis of brain tissue densities. Comput Methods Programs Biomed 148:27–43. https://doi.org/10.1016/j.cmpb.2017.06.011
36. Karthik R, Menaka R (2017) A multi-scale approach for detection of ischemic stroke from brain MR images using discrete curvelet transformation. Meas J Int Meas Confed 100:223–232. https://doi.org/10.1016/j.measurement.2017.01.001
37. Subudhi A, Acharya UR, Dash M et al (2018) Automated approach for detection of ischemic stroke using Delaunay Triangulation in brain MRI images. Comput Biol Med 103:116–129. https://doi.org/10.1016/j.compbiomed.2018.10.016

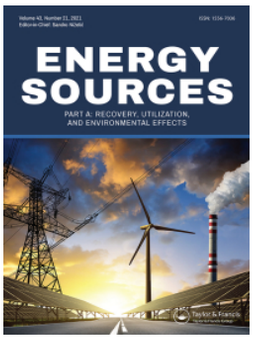
38. Rajendra Acharya U, Meiburger KM, Faust O et al (2019) Automatic detection of ischemic stroke using higher order spectra features in brain MRI images. Cogn Syst Res 58:134–142. https://doi.org/10.1016/j.cogsys.2019.05.005
39. Dourado CMJM, da Silva SPP, da Nóbrega RVM et al (2019) Deep learning IoT system for online stroke detection in skull computed tomography images. Comput Networks 152:25–39. https://doi.org/10.1016/J.COMNET.2019.01.019
40. Patel A, Van De Leemput SC, Prokop M et al (2019) Image Level Training and Prediction: Intracranial Hemorrhage Identification in 3D Non-Contrast CT. IEEE Access 7:92355–92364. https://doi.org/10.1109/ACCESS.2019.2927792
41. Ho KC, Speier W, Zhang H et al (2019) A Machine Learning Approach for Classifying Ischemic Stroke Onset Time from Imaging. IEEE Trans Med Imaging 38:1666–1676. https://doi.org/10.1109/TMI.2019.2901445
42. Anbumozhi S (2020) Computer aided detection and diagnosis methodology for brain stroke using adaptive neuro fuzzy inference system classifier. Int J Imaging Syst Technol 30:196–202. https://doi.org/10.1002/ima.22380
43. Gautam A, Raman B (2020) Local gradient of gradient pattern: a robust image descriptor for the classification of brain strokes from computed tomography images. Pattern Anal Appl 23:797–817. https://doi.org/10.1007/s10044-019-00838-8
44. Vasconcelos FFX, Sarmento RM, Rebouças Filho PP, de Albuquerque VHC (2020) Artificial intelligence techniques empowered edge-cloud architecture for brain CT image analysis. Eng Appl Artif Intell 91:103585. https://doi.org/10.1016/j.engappai.2020.103585
45. Herzog L, Murina E, Dürr O et al (2020) Integrating uncertainty in deep neural networks for MRI based stroke analysis. Med Image Anal 65:101790. https://doi.org/10.1016/j.media.2020.101790
46. Anupama CSS, Sivaram M, Lydia EL et al (2022) Synergic deep learning model–based automated detection and classification of brain intracranial hemorrhage images in wearable networks. Pers Ubiquitous Comput 26:1–10. https://doi.org/10.1007/s00779-020-01492-2
47. Gautam A, Raman B (2021) Towards effective classification of brain hemorrhagic and ischemic stroke using CNN. Biomed Signal Process Control 63:102178. https://doi.org/10.1016/j.bspc.2020.102178
48. Cetinoglu YK, Koska IO, Uluc ME, Gelal MF (2021) Detection and vascular territorial classification of stroke on diffusion-weighted MRI by deep learning. Eur J Radiol 145:110050. https://doi.org/10.1016/j.ejrad.2021.110050
49. Xu Y, Holanda G, Souza LFDF et al (2021) Deep Learning-Enhanced Internet of Medical Things to Analyze Brain CT Scans of Hemorrhagic Stroke Patients: A New Approach. IEEE Sens J 21:24941–24951. https://doi.org/10.1109/JSEN.2020.3032897
50. Gautam A, Raman B (2021) Brain strokes classification by extracting quantum information from CT scans. Multimed Tools Appl 1–17. https://doi.org/10.1007/s11042-021-11342-9
51. Neethi AS, Niyas S, Kannath SK et al (2022) Stroke classification from computed tomography scans using 3D convolutional neural network. Biomed Signal Process Control 76:103720. https://doi.org/10.1016/j.bspc.2022.103720
52. Dogan S, Barua PD, Baygin M et al (2022) Novel multiple pooling and local phase quantization stable feature extraction techniques for automated classification of brain infarcts. Biocybern Biomed Eng 42:815–828. https://doi.org/10.1016/j.bbe.2022.06.004
53. Chen YT, Chen YL, Chen YY et al (2022) Deep Learning-Based Brain Computed Tomography Image Classification with Hyperparameter Optimization through Transfer Learning for Stroke. Diagnostics 12:807. https://doi.org/10.3390/diagnostics12040807
54. Tasci B, Tasci I (2022) Deep feature extraction based brain image classification model using preprocessed images: PDRNet. Biomed Signal Process Control 78:103948. https://doi.org/10.1016/j.bspc.2022.103948
55. Din MS, Gurbuz S, Akbal E et al (2022) Exemplar deep and hand-modeled features based automated and accurate cerebral hemorrhage classification method. Med Eng Phys 105:103819. https://doi.org/10.1016/j.medengphy.2022.103819
56. Altuve M, Pérez A (2022) Intracerebral hemorrhage detection on computed tomography images using a residual neural network. Phys Medica 99:113–119. https://doi.org/10.1016/j.ejmp.2022.05.015
57. Deepa B, Murugappan M, Sumithra MG et al (2022) Pattern Descriptors Orientation and MAP Firefly Algorithm Based Brain Pathology Classification Using Hybridized Machine Learning Algorithm. IEEE Access 10:3848–3863. https://doi.org/10.1109/ACCESS.2021.3100549
58. Sivakumar P, Ganeshkumar P (2017) An efficient automated methodology for detecting and segmenting the ischemic stroke in brain MRI images. Int J Imaging Syst Technol 27:265–272. https://doi.org/10.1002/ima.22231

59. Zhang R, Zhao L, Lou W et al (2018) Automatic Segmentation of Acute Ischemic Stroke From DWI Using 3-D Fully Convolutional DenseNets. IEEE Trans Med Imaging 37:2149–2160. https://doi.org/10.1109/TMI.2018.2821244
60. Gautam A, Raman B (2019) Segmentation of ischemic stroke lesion from 3d mr images using random forest. Multimed Tools Appl 78:6559–6579. https://doi.org/10.1007/s11042-018-6418-2
61. Deepa B, Sumithra MG (2019) An intensity factorized thresholding based segmentation technique with gradient discrete wavelet fusion for diagnosing stroke and tumor in brain MRI. Multidimens Syst Signal Process 30:2081–2112. https://doi.org/10.1007/s11045-019-00642-x
62. Liu L, Chen S, Zhang F et al (2020) Deep convolutional neural network for automatically segmenting acute ischemic stroke lesion in multi-modality MRI. Neural Comput Appl 32:6545–6558. https://doi.org/10.1007/s00521-019-04096-x
63. Nishio M, Koyasu S, Noguchi S et al (2020) Automatic detection of acute ischemic stroke using non-contrast computed tomography and two-stage deep learning model. Comput Methods Programs Biomed 196:105711. https://doi.org/10.1016/j.cmpb.2020.105711
64. Nazari-Farsani S, Nyman M, Karjalainen T et al (2020) Automated segmentation of acute stroke lesions using a data-driven anomaly detection on diffusion weighted MRI. J Neurosci Methods 333:108575. https://doi.org/10.1016/j.jneumeth.2019.108575
65. Li L, Wei M, Liu B et al (2021) Deep Learning for Hemorrhagic Lesion Detection and Segmentation on Brain CT Images. IEEE J Biomed Heal Inform 25:1646–1659. https://doi.org/10.1109/JBHI.2020.3028243
66. Zhang L, Zhu C, Wu YW et al (2021) SFCN: Symmetric feature comparison network for detecting ischemic stroke lesions on CT images. IET Image Process 15:2818–2832. https://doi.org/10.1049/ipr2.12267
67. Shin H, Agyeman R, Rafiq M et al (2022) Automated segmentation of chronic stroke lesion using efficient U-Net architecture. Biocybern Biomed Eng 42:285–294. https://doi.org/10.1016/j.bbe.2022.01.002
68. Wang S, Chen Z, You S et al (2022) Brain stroke lesion segmentation using consistent perception generative adversarial network. Neural Comput Appl 34:8657–8669. https://doi.org/10.1007/s00521-021-06816-8
69. Tomasetti L, Hollesli LJ, Engan K et al (2022) Machine Learning Algorithms Versus Thresholding to Segment Ischemic Regions in Patients with Acute Ischemic Stroke. IEEE J Biomed Heal Informatics 26:660–672. https://doi.org/10.1109/JBHI.2021.3097591
70. Bao Q, Mi S, Gang B et al (2022) MDAN: Mirror Difference Aware Network for Brain Stroke Lesion Segmentation. IEEE J Biomed Heal Inform 26:1628–1639. https://doi.org/10.1109/JBHI.2021.3113460
71. ITK-SNAP Home. http://www.itksnap.org/pmwiki/pmwiki.php. Accessed 1 Aug 2023
72. Subudhi A, Dash M, Sabut S (2020) Automated segmentation and classification of brain stroke using expectation-maximization and random forest classifier. Biocybern Biomed Eng 40:277–289. https://doi.org/10.1016/j.bbe.2019.04.004
73. Yalçın S, Vural H (2022) Brain stroke classification and segmentation using encoder-decoder based deep convolutional neural networks. Comput Biol Med 149:105941. https://doi.org/10.1016/j.compbiomed.2022.105941
74. BraTS - V7 Open Datasets. https://www.v7labs.com/open-datasets/brats. Accessed 22 Oct 2022
75. OpenNeuro. https://openneuro.org/. Accessed 22 Oct 2022
76. PhysioNet. https://physionet.org/. Accessed 22 Oct 2022
77. ISLES: Ischemic Stroke Lesion Segmentation Challenge 2022. https://www.isles-challenge.org/. Accessed 22 Oct 2022
78. Anatomical Tracings of Lesions After Stroke (ATLAS). https://www.fcon_1000.projects.nitrc.org/indi/retro/atlas.html. Accessed 22 Oct 2022
79. ADNI | Alzheimer's Disease Neuroimaging Initiative. https://www.adni.loni.usc.edu/. Accessed 22 Oct 2022
80. OASIS Brains - Open Access Series of Imaging Studies. https://www.oasis-brains.org/. Accessed 22 Oct 2022
81. Naser MZ, Alavi A (2020) Insights into Performance Fitness and Error Metrics for Machine Learning. Archit Struct Constr. https://doi.org/10.1007/s44150-021-00015-8
82. Shamir RR, Duchin Y, Kim J, et al (2019) Continuous Dice Coefficient: a Method for Evaluating Probabilistic Segmentations. https://doi.org/10.48550/arxiv.1906.11031
83. Yeghiazaryan V, Voiculescu I (2018) Family of boundary overlap metrics for the evaluation of medical image segmentation. J Med Imaging 5:1. https://doi.org/10.1117/1.JMI.5.1.015006

84. Karimi D, Salcudean SE (2020) Reducing the Hausdorff Distance in Medical Image Segmentation with Convolutional Neural Networks. IEEE Trans Med Imaging 39:499–513. https://doi.org/10.1109/TMI.2019.2930068
85. Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S (2019) Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. IEEE, pp 658–666

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

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



Energy Sources, Part A: Recovery, Utilization, and Environmental Effects

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ueso20

Attenuation of hotspot phenomenon in PV systems

Swatilekha Ghosh, Santosh Kumar Singh & Vinod Kumar Yadav

To cite this article: Swatilekha Ghosh, Santosh Kumar Singh & Vinod Kumar Yadav (2023) Attenuation of hotspot phenomenon in PV systems, Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, 45:4, 11788-11802, DOI: 10.1080/15567036.2023.2264811

To link to this article: https://doi.org/10.1080/15567036.2023.2264811

Published online: 04 Oct 2023.

Submit your article to this journal

Article views: 23

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=ueso20

ENERGY SOURCES, PART A: RECOVERY, UTILIZATION, AND ENVIRONMENTAL EFFECTS
2023, VOL. 45, NO. 4, 11788–11802
https://doi.org/10.1080/15567036.2023.2264811

Taylor & Francis
Taylor & Francis Group

Check for updates

Attenuation of hotspot phenomenon in PV systems

Swatilekha Ghosh[a], Santosh Kumar Singh[a], and Vinod Kumar Yadav[b]

[a]Department of Environmental Engineering, Delhi Technological University, Delhi, India; [b]Department of Electrical Engineering, Delhi Technological University, Delhi, India

ABSTRACT

PV waste management is a looming problem and a growing concern for environmental engineers and energy policymakers, as the early PV installations are approaching their end-of-life. Hence, improvement of PV reliability and prevention of early failures has now become more important than ever. Notably, the formation of hotspots in PV cells under partial shading and other mismatch conditions remains a dominant cause of early failures in PV modules. In this study, we undertake an experimental investigation to assess the effectiveness of the extant bypass mechanism in preventing hotspot formation in PV modules. The findings reveal several limitations associated with the conventional bypass approach. Although the bypass diode partially fulfills its intended purpose within a string array configuration, its effectiveness is entirely compromised when connected in the configuration of real-world installations, such as in a series-parallel (SP) array. The hotspot temperature in the SP array increased up to 70.7°C under partial shading conditions at the operating point (*i.e.*, at MPP), whereas no hotspot was created (below 46.1°C) in a string array, except in the single-cell shading scenario. To make the bypass circuits effective, in the present work, a novel bypass strategy is proposed.

ARTICLE HISTORY
Received 24 April 2023
Revised 11 September 2023
Accepted 25 September 2023

KEYWORDS
Early failure; PV waste management; reverse breakdown; PV reliability; sustainability

Introduction

Solar PV has become the preferred choice as an energy source of the modern era by policymakers and researchers around the world; however, there are some challenges associated with PV technologies that have started knocking on the door of some countries and going to confront the rest of the world in the near future. One of these challenges pertains to the substantial volume of PV waste that is projected to arise. The PV installation started gaining momentum on a commercial scale since the beginning of the 21st century, and as the PV modules have a useful service life of 25–30 years (Mahmoudi et al. 2019), a large number of early PV installations are nearing their end of life. Additionally, the survey of operational PV plants reveals that about 2% of the PV installations fail within 11–12 years of commissioning (undefined). Hence, the end-of-life PV waste volume together with the PV waste volume due to early failures will be a daunting challenge to manage in the near future. The PV module recycling technology has not reached commercial viability and scale yet and hence presently most of the PV waste ends up in landfills (Mathur, Singh, and Sutherland 2020). As per the estimation of the International Renewable Energy Agency (IRENA), approximately 8 million tonnes of PV waste are expected to be generated by 2030 and 78 million tonnes by 2050 (IRENA and IEA-PVPS 2016). In most developing countries, the major chunk of prevalent electronic waste or rather the complete solid waste volume is managed in nonscientific and non-environment-compliant ways due to a lack of adequate infrastructure (Vats and Singh 2014). Hence, the additional inflow of PV waste is going to throw a tough challenge before policymakers around the world if adequate steps are not taken today.

CONTACT Vinod Kumar Yadav Vinodkumar@dtu.ac.in Department of Electrical Engineering, Delhi Technological University, Delhi, India

© 2023 Taylor & Francis Group, LLC

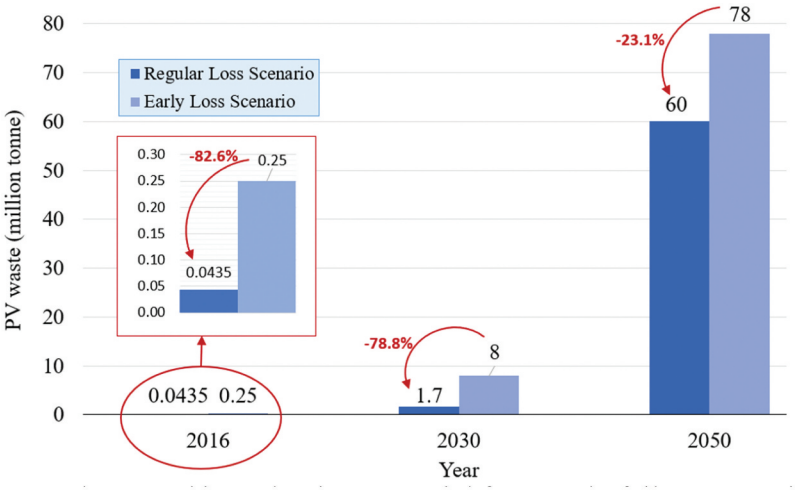


Figure 1. Predicted PV waste volume and its reduction potential from early failure scenario to regular loss scenario.

The silver lining to this impending issue is that the estimated large volume of PV waste volume mentioned above could be reduced to 78.8% and 23.1% of the estimates mentioned above in 2030 and 2050, respectively, if early failures are averted (see Figure 1). Hence, enhancing the reliability of PV modules through the elimination of factors contributing to premature failures presents a promising avenue to address the challenge faced by environmental researchers and policymakers.

One of the dominant causes of early failure of PV modules is the generation of hotspots in the PV cells under mismatch conditions such as partial shading, non-uniform deposition of dust, light-induced degradation, and potential-induced degradation (Palawat et al. 2021; Yadav et al. 2021). There is a whole gamut of factors that influence the nature and intensity of the mismatch losses in PV modules (see Figure 2). Under these mismatch conditions, the nonproductive cells are forced to conduct the same current that the healthy and fully operational cells are producing. When the reverse current flows through the nonproductive cells for a significant duration, a hotspot (up to 400°C (Ghosh, Yadav, and Mukherjee 2020)) is created in those which eventually cause non-reversible damage (Guerrieroa, Tricolib, and Daliento 2019; Stonier et al. 2023), and even fire hazards in extreme cases (Bakır 2023).

One of the solutions proposed by researchers in the past to deal with mismatch power loss is the use of innovative array configurations with different degrees of redundancy in the interconnections

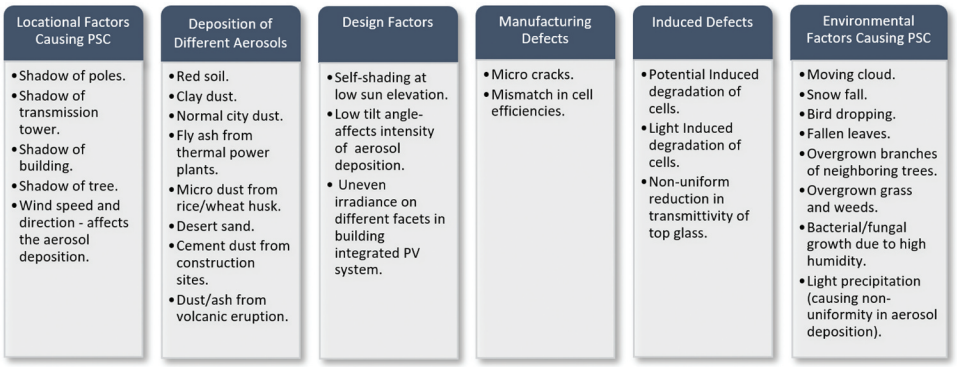


Figure 2. Factors influencing partial shading and mismatch losses.

(Lappalainen and Valkealahti 2017; Palawat et al. 2021), which may be broadly classified into three categories: (a) physical relocation, (b) electrical relocation, and (c) reconfiguration through innovative switching matrices (Janani, Babu, and Vijayakumar 2023). The comparative studies between different array configurations carried out in the past (Janani, Babu, and Vijayakumar 2023; Pachauri et al. 2018; Yadav, Pachauri, and Chauhan 2016; Dhanalakshmi and Rajasekar 2018) reveal that with the increase in the number of cross ties, the susceptibility to mismatch losses decreases. However, the reconfiguration of PV arrays in real applications is extremely difficult to implement, both technically and commercially, especially for utility-scale PV projects. On the other hand, these do not add any value under normal operating conditions; on the contrary, the array with no cross ties, *i.e.*, conventional Series-Parallel (SP) array, is observed to yield the highest fill factor under normal operating conditions (Wang and Hsu 2011). Secondly, the additional cross-ties increase the project cost and reduce the reliability of the PV system as the number of joints increases.

Due to the above factors, in commercial PV plant design SP array is invariably implemented for large-scale and string arrays for small-scale projects (Messenger and Abtahi 2010; Papadopoulou 2013). All innovative configurations presented in the literature practically remain restricted to research scale only.

Another approach to deal with the partial shading and other mismatch conditions taken by researchers in the past is the modification of the bypass circuit. To improve the effectiveness of the bypass mechanism, using multiple bypass diodes was proposed in (Pannebakker, Waal, and Sark 2017), which was demonstrated to effectively mitigate the occurrence of the hotspot phenomenon; however, this solution increases the cost and the complexity of the circuit. More recently several alternative solutions have been proposed based on solid-state switches by the researchers, which can be broadly categorized into three groups: (a) use of a solid-state switch in the bypass path and no device in the series path (Bauwens and Doutreloigne 2014), (b) Use of a bypass diode in bypass path and a solid-state switch in the series path (Guerriero et al. 2017), (c) Use of a solid-state switch in bypass path and a solid-state switch in the series path (Dhimish et al. 2018a). The studies published on the above schemes reported the effectiveness of each one of them; however, the viability was illustrated for the most part on a single module, and the impact of these circuits on the array performance was not quantified. Secondly, a detailed discussion on the strategy for controlling the operations of those solid-state switches was ignored in past studies.

Hence, in the present work, to develop deeper insights into the hotspot prevention mechanism in real-world installations, string and SP array configurations are experimentally studied with a special focus on the behavior of the conventional bypass circuits under mismatch conditions and their efficacy in preventing hotspots in such scenarios. Further, based on the result of the experimental study, a novel bypass-circuit control strategy is proposed in order to increase the effectiveness of the bypass circuits and thereby improve the reliability of the PV modules.

The main contributions of the present article are summarized below.

(a) Critically analyzes the environmental impacts of end-of-life waste generation and early failure of PV modules.
(b) Explores the limitations of the solutions proposed in the past to deal with the impact of partial shading and other mismatch conditions on PV arrays.
(c) Experimentally investigates the effectiveness of the extant bypass mechanisms in a series string and a series-parallel array configuration.
(d) Based on the observations of the experimental study, a novel solution is proposed to increase the effectiveness of the bypass circuits, in order to improve the reliability of the PV modules.

Design of experiment

An experimental setup was developed in the present work to carry out a comparative study of the efficacy of the bypass diodes in the two most commonly used array configurations (*i.e.*,

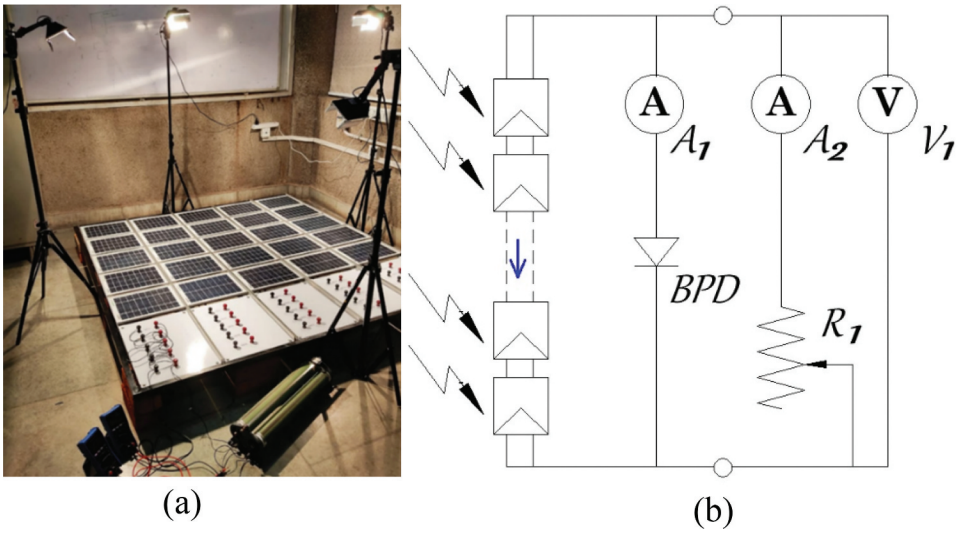


Figure 3. Experimental setup consisting of (a) PV array and measuring instruments, (b) simplified schematic diagram of the experimental setup.

Table 1. Specification of the module used in the experimental study.

Parameter	Specification	Parameter	Specification
Model	EIL 40	Open-circuit voltage (V_{oc})	22.0 V
Maximum power at STC (P_{max})	40.0 W	Current at maximum power (I_{mp})	2.2 A
Short-circuit current (I_{sc})	2.4 A	Voltage at maximum power (V_{mp})	18.0 V

Note: Specified values are at STC (i.e., the irradiance of 1000 W/m^2, AM 1.5, Temperature 25°C).

a string and an SP array) under partial shading conditions (PSCs) in preventing the hotspot formation in the shaded cells. The experimental study was mainly focused on establishing the efficacy of the bypass diode in preventing the hotspot formation in the series and SP array configurations under PSC. For this purpose, the bypass current (*i.e.*, the current through the bypass diode) was measured by disconnecting (desoldering) the bypass diode and connecting an ammeter (A_1) in series with it (see Figure 3), while different PSCs were created by covering a different number of cells. A variable resistive load (R_1), another ammeter (A_2), and a voltmeter (V_1) were connected to the array terminals to establish the power outputs of the arrays as well for the complete operating range, *i.e.*, from the short-circuit voltage (*i.e.*, $V_{SC} = 0$) to the open-circuit voltage (V_{OC}). The bypass current was superimposed on the output of the arrays under different PSCs to generate deeper insights into the behavior of the bypass diodes at different operating points. Additionally, the thermal response of the shaded cells was simultaneously observed using a thermal imaging camera with 640×480 infrared resolution. The modules (refer Table 1 for details specification) were connected first in a 3×1 series string configuration, and different numbers of cells (from 1 to 4) were covered in sequence to create partial shading conditions (PSCs). The same steps were later repeated on a 3×3 array of SP array configuration by connecting three such (*i.e.*, 3×1) strings in parallel. The experiment was conducted under natural solar irradiance to simulate an environment closer to real applications. However, the experiment was conducted in a fixed time slot every day to minimize the variation in the irradiance level and ambient temperature to the extent practically possible. The solar irradiance varied between 550 and 700 W/m^2, during the period of the experimental study. The result of the experimental study is presented and discussed in the following section.

Result and discussion

In the present experimental study, four parameters were measured simultaneously, while different partial shading conditions (PSCs) were simulated by covering a different number of cells, which are (a) bypass current, (b) array current, (c) array voltage, and (d) thermal response of the shaded cells. Additionally, natural solar irradiance and ambient temperature were also recorded for reference to ensure they do not vary too widely to affect the objectives of the study. The power output of the arrays was derived for the complete operating range using the discrete values of the array current and voltage measured at different operating points. The bypass current was also measured for the complete operating range to generate deeper insights about the operation of the bypass circuit vis-à-vis the power output of the arrays. In this section, the results obtained from the above investigation carried out on a 3 × 1 series string and a 3 × 3 SP array are presented and compared to bring out the difference in the functioning of the bypass mechanisms in the above two array configurations. The bypass current superimposed on the performance curves of the 3 × 1 string array, for 1–4 cell shading conditions, is presented in Figures 4–6, respectively. It may be observed from Figure 4 that the bypass diode did not conduct in the single-cell shading scenario in the complete operating range. Only a small leakage current of 0.095 A was recorded in the region close to the short circuit (between 0 and 10 V) through the bypass diode (see Figures 4(a,b)). This implies that the shaded cell had undergone reverse breakdown, and the complete string current (if the small leakage current is ignored) passed through the shaded cell. The infra-red image of the module captured at MPP after a duration of about 180 s validates the above hypothesis. The thermal image shows that a hotspot was formed when a single cell

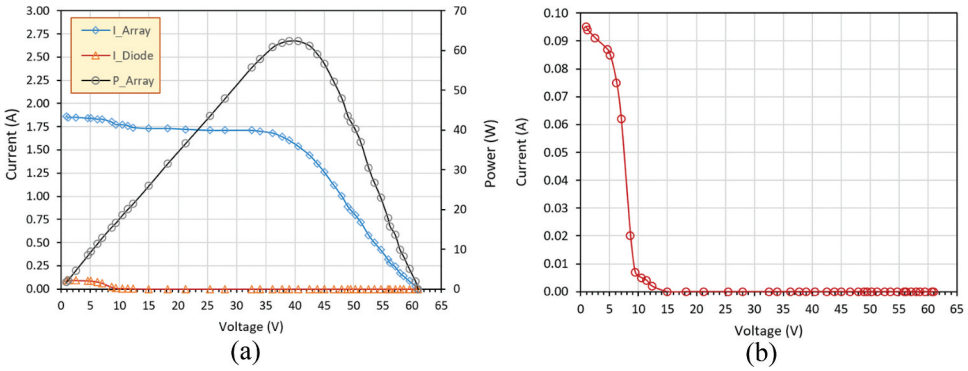


Figure 4. Performance of 3×1 string array under single cell shading condition (a) bypass current superimposed on the electrical performance of the array, (b) bypass current–enlarged plot.

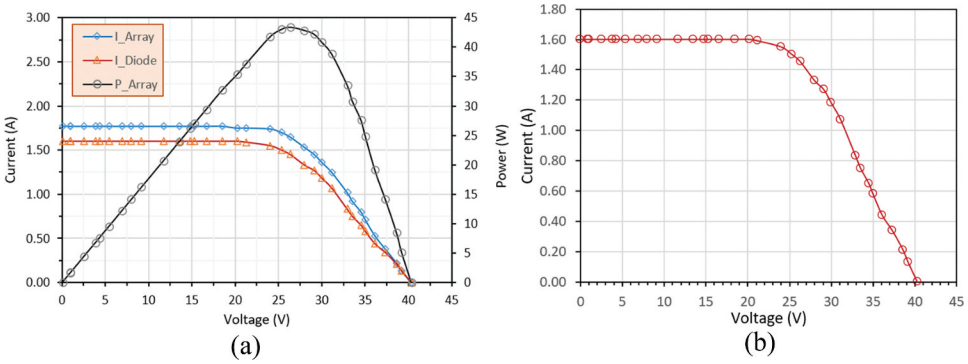


Figure 5. Performance of 3×1 string array under 2-cells shading condition (a) bypass current superimposed on the electrical performance of the array, (b) bypass current-enlarged plot.

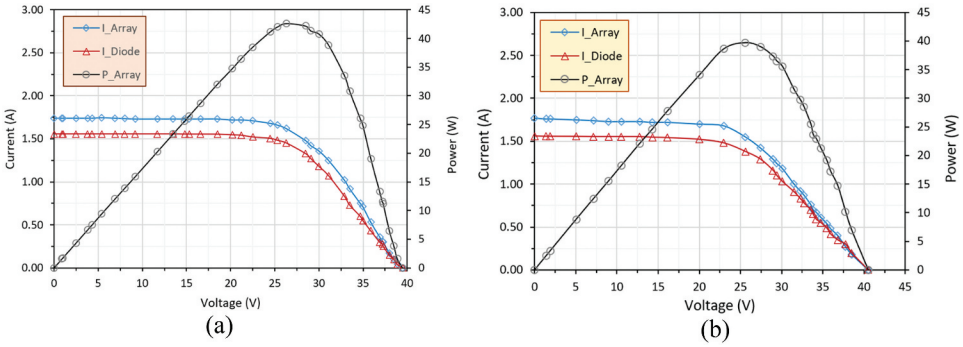


Figure 6. Performance of 3×1 string array inclusive of its bypass circuit under (a) 3-cell shading condition (b) 4-cell shading condition.

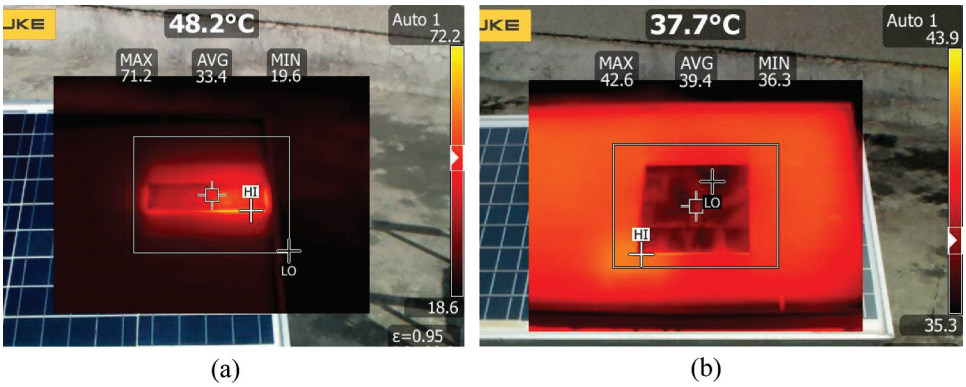


Figure 7. Thermal images of the partially shaded module of a 3×1 string at MPP under (a) single-cell shading, and (b) more than one-cell shading scenarios.

was shaded wherein the temperature increased to 71.2°C (see Figure 7(a)) . However, the temperature decreased when more than one cell was shaded, as the bypass diode started conducting in those scenarios, in the 3 × 1 string array. The temperature dropped to 42.6°C (see Figure 7(b)) when four cells were shaded. In the case of single cell shading, the bypass diode was in nonconductive mode as may be observed from Figure 4, and hence the shaded cell acted as a load to the productive cells and consumed power instead of producing it. As a result, the temperature of the cell quickly increased. This observation is in good agreement with the fault diagnostic studies carried out on utility-scale PV projects in the recent past (Bakır 2023). A similar phenomenon was observed in the case of the 3 × 3 SP array as well for the single-cell shading scenario. The bypass diode may be observed to have remained non-conducting when a single cell was shaded (see Figure 9). Similar to the case of a single string, a small leakage current was conducted through the bypass diode (see Figure 9(b)) as well, near short-circuit region. The value of the leakage current was 0.091 A, whereas the short-circuit current of the array was 5.62 A (see Table 2). An infrared image of the shaded cell was captured in the case of the SP array as well (see Figure 13(a)), which reveals hotspot formation in the shaded cell with a temperature rise of 24.7°C over the normal operating temperature. The above observations imply that in the single-cell shading scenario the bypass diode failed to conduct, and hence reverse breakdown occurred in the shaded cell both in the string and in the SP arrays. The observations delineated in the paragraph above can be attributed to the fact that the bypass diode is basically a passive unidirectional switch and in order to turn bypass diodes to an ON state, a forward bias voltage is required across its terminals which are in opposite polarity of the voltage generated in the PV cells. Unless there is a source of negative polarity voltage in the sub-string, the bypass diode cannot go into the conduction mode. The

Table 2. Bypass and array current at different operating points.

Shading Scenarios	Operating Point	String Array			SP Array		
		MPP	Short-Circuit Point	Peak Cell Temp. (°C)	MPP	Short-Circuit Point	Peak Cell Temp. (°C)
1 cell	String Current (A)	1.6	1.86	71.2	4.731	5.62	68.1
	Bypass current (A)	0	0.095		0	0.091	
	Difference (%)	100	94.9		100	98.4	
2 cells	String Current (A)	1.65	1.77	46.1	3.989	5.743	70.7
	Bypass current (A)	1.452	1.6		0	1.728	
	Difference (%)	12	9.6		100	69.9	
3 cells	String Current (A)	1.62	1.74	44.4	3.319	5.553	67.5
	Bypass current (A)	1.452	1.556		0	1.542	
	Difference (%)	10.4	10.6		100	72.2	
4 cells	String Current (A)	1.55	1.77	42.6	3.137	4.85	64.2
	Bypass current (A)	1.38	1.56		0	1.43	
	Difference (%)	10.9	11.9		100	70.5	

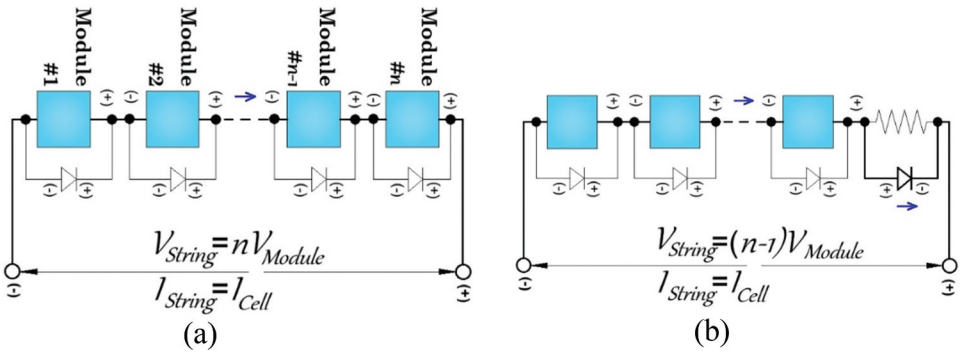


Figure 8. Bypass diode bias voltage polarity. (a) under normal conditions, and (b) under partial shading (on module #n).

polarity of the sub-string voltage can be reversed only if there is a voltage source of opposite polarity to the cells' voltage and of greater amplitude or if the substring becomes a consumer of power (see Figure 8), which is the only feasible possibility in the real-world scenario. This implies that to turn on the bypass diode of the commercial PV modules at least one cell or an equivalent area of cells needs to undergo reverse breakdown and become dissipater of power. Hence, in the case of single-cell shading in both string and SP arrays, hotspots were observed in the shaded cells. For further reading on the physics behind the reverse breakdown mechanisms and hotspot formation in the cells under partial shading, Ghosh, Yadav, and Mukherjee(2020) may be referred.

However, a major shift may be observed in the behavior of the bypass diode in two and more cells' shading scenarios when the string was converted to an SP array. In the case of a string array, the bypass diode was observed to be conducted over the complete operating range under two and more cell shading scenarios (see Figures 8 and 9). A small difference was observed in the array (string) current and the bypass current and this difference widened linearly from the open circuit to MPP and remained almost constant from MPP to the short-circuit region. The difference between the array current and the bypass current is that the string array was between 9.6 and 11.9% at the short-circuit point and between 10.4 and 12% at MPP (see Table 2). This implies that though the bypass diode went

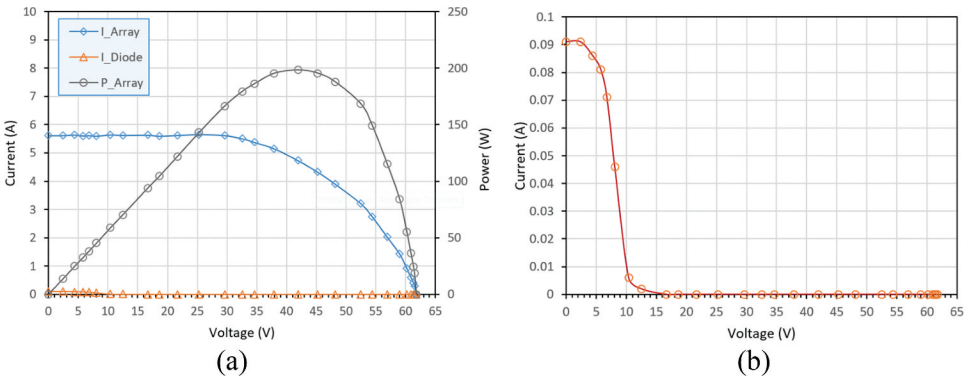


Figure 9. Performance of SP array under single cells shading condition (a) bypass current superimposed on the electrical performance of the array, (b) bypass current-enlarged view.

to conduction mode during shading scenarios other than the one-cell shading scenario, a small fraction of the array current still passed through the strings, especially when the array current was more than the current at MPP (I_{MP}). Though a higher shunt resistance of the PV cells may reduce this stray current, this did not impact the shaded cells detrimentally. Approximately 90% of the array current was conducted through the bypass diode during all these shading scenarios (except single-cell shading scenarios); as a result, the temperature of the shaded cells did not rise and remained at a temperature close to the operating temperature of the other cells of the module (see Figure 7 (b)).

However, the response of the bypass circuit was observed to be in total contrast with the string array when a 3 × 3 SP array was configured using three such strings, in two or more cell shading scenarios. In these scenarios, the bypass diode remained in an OFF state till a point well beyond the global MPP (GMPP) point. The GMPP was observed at 41.8, 46.2, and 47.3 V, respectively, for two cells, three cells, and four-cell shading scenarios, whereas the bypass diode was observed to conduct in the operating range from 0 to 37.81, 38.9, and 38.6 V, respectively (see Figures 10–12), for the above shading scenarios. This implies that at the GMPP the bypass diode remained in an OFF state in all the shading cases. The difference between the array current and the bypass current was much larger compared to the string array. For two cells, three cells, and four cells' shading scenarios, this difference was observed to be 9.9, 72.2, and 70.5%, respectively (see Table 2). This difference is larger than the values recorded for the string array because in the SP array, there were three such parallel strings, and the bypass diode bypassed the current of only one string, *i.e.*, the string subjected to partial shading.

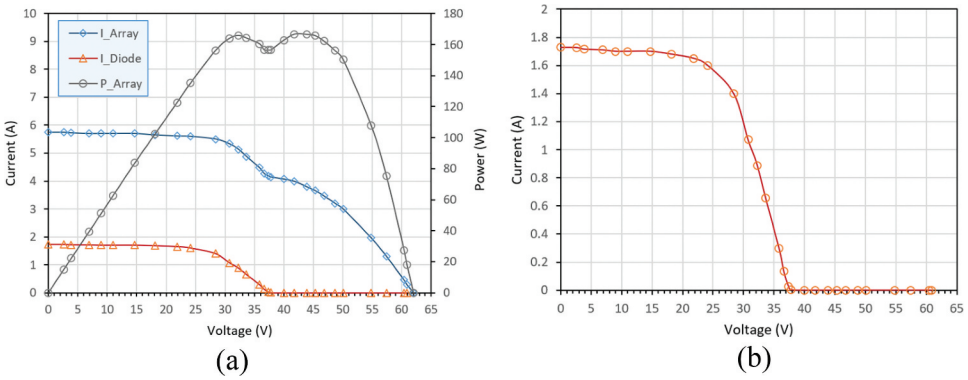


Figure 10. Performance of SP array under two cells shading condition (a) bypass current superimposed on the electrical performance of the array, (b) bypass current-enlarged view.

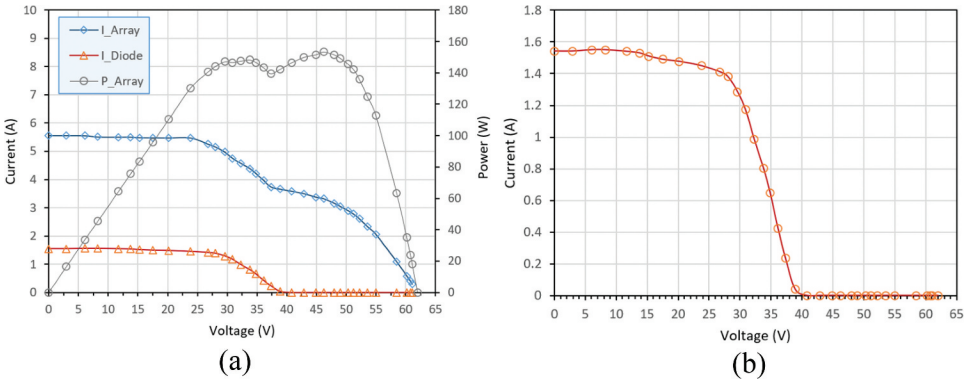


Figure 11. Performance of SP array under three cells shading condition (a) bypass current superimposed on the electrical performance of the array, (b) bypass current-enlarged view.

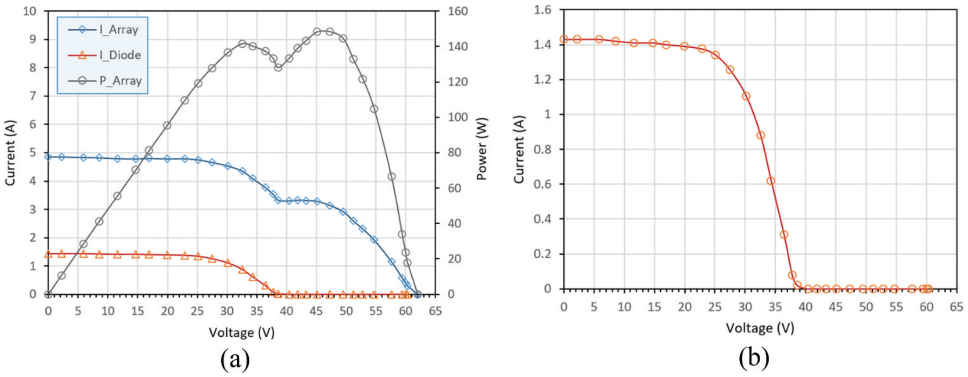


Figure 12. Performance of SP array under four cells shading condition (a) bypass current superimposed on the electrical performance of the array, (b) bypass current-enlarged view.

As the bypass diode did not conduct at GPPP, it is of great interest to study the thermal response of the shaded cells as well. For this purpose, the infrared images were captured under all the shading scenarios by operating the array at GMPP for about 300–350 s. The thermal images (see Figs. 13(a)–13(d)) reveals that the temperature of the shaded cells was increased significantly in all the shading scenarios, whereas in the case of a string array, the hotspot temperature was observed in the case of single cell shading only (compared with Figure 7(a)). It may be observed from Figure 13 that the temperature increased to 70.7, 67.5, and 64.2°C under two-cell, three-cell, and four-cell shading scenarios, respectively.

Proposed strategy for effective protection of modules under PSC

From the results presented in Section III, it is evident that the extant bypass mechanism provided in the commercial PV modules to prevent the hotspot occurrence in the PV cells under partial shading conditions has several limitations. Firstly, the bypass diode does not turn on in the case of single-cell shading in either of the arrays studied in the present work, *i.e.*, string and SP arrays. This appears to have been accepted as the industry standard due to the technical limitations of the used bypass circuit. However, a single-cell shading scenario needs to be viewed with equal importance as individual cells are not replaceable, and hence in case of failure of a single cell the complete module will have to be replaced. Secondly, though in a string array, in the case of

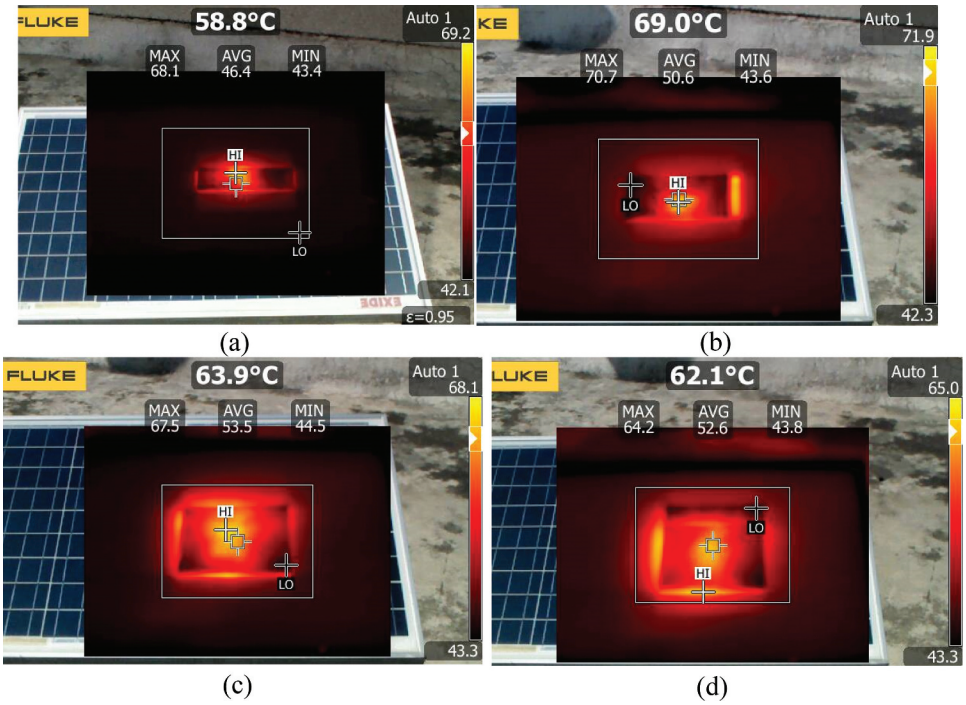


Figure 13. Thermal images of the partially shaded module at MPP in the series-parallel configuration under (a) single-cell shading, (b) two-cell shading, (c) three-cell shading, and (d) four-cell shading scenarios.

two and more cell shading scenarios, the bypass circuit remains in an ON state in the complete operating range, *i.e.*, short circuit to open circuit in a string array, in SP arrays it turns ON only near the short-circuit region and interestingly remains in OFF state at GMPP. Hence, the extant bypass diode does not serve its intended purpose in the SP array and serves its purpose only partially in the string array. Hence, it is of paramount importance to devise an effective bypass circuit and its control strategy to prevent early failures due to PSC and reduce contribution to waste generation.

Along with the bypass circuit, another component of the PV system which is also of paramount importance for the reliable and efficient operation of PV systems is the Maximum Power Point Tracking (MPPT) technology. There is a plethora of MPPT algorithms to execute the above activities which have their own advantages and disadvantages. For example, the conventional techniques such as perturb and observe method (P&O), Incremental conductance (IC), and hill-climbing (HC) work efficiently under uniform irradiation; however, they lose their effectiveness in discriminating the GMPP from local MPPs under partial shading conditions (Kermadi et al. 2020). To make the MPPT algorithms more intelligent and suitable for identifying the GMPPs under dynamic partial shading conditions, soft computing-based algorithms have been proposed, viz. differential evolution (Kumar et al. 2017), flashing fireflies (Sundareswaran, Peddapati, and Palani 2014), gray wolf optimization (Mohanty, Subudhi, and Ray 2016), particle swarm optimization (Ishaque and Salam 2013), artificial bee colony (Sundareswaran et al. 2015), cuckoo search (Ahmed and Salam 2014), and genetic algorithm. These techniques are more efficient in identifying the GMPP as they utilize modern computation power to widen the search space, however, because of the wider search space the tracking speed is lower than the conventional MPPT algorithms, especially under uniform irradiance. Because of the inherent advantages and disadvantages, the above algorithms find applications based on the criticality of the application, the scale of the plant, the dynamic nature of PSC, and the economy of the project. However, the use of all the above MPPT algorithms has been restricted to identifying the

GMPP only and not explored for their potential application in controlling the operation of the bypass circuit to make them effective.

The MPPT and bypass circuit both are equally important for the efficient and reliable operation of PV systems; however, these two components work independently of each other in the present state of the art. This reveals the scope for improvement and makes the MPPT algorithm more versatile and effective from the point of improving reliability. As discussed in Section II, the utility-scale PV plants are generally designed in SP configuration, on the other hand, the experimental results reveal that the bypass diode remains in an OFF state in all the shading scenarios in SP array configuration at GMPP. Hence, the bypass diode can be considered nonfunctional for all practical purposes in the SP array as the array will always operate at GMPP by the MPPT algorithm employed. Secondly, the bypass diode is nonfunctional in all arrays in the case of a single-cell shading scenario. To improve upon these limitations, in the present work, a novel bypass strategy is proposed which includes, firstly, an active switch replacing the passive bypass diode, which can either be a MOSFET-based or IGBT-based. Several researchers have proposed active switch-based bypass circuits either in parallel with the string (*i.e.*, a configuration similar to the extant bypass diode) or in series with the string, and even both in series and parallel with the strings. However, in the past research, adequate attention has not been paid to the gate drive circuit strategy. Several researchers have advocated in favor of not using separate gate drive circuits to provide the gate pulse from the output voltage of the PV module itself to avoid the complexity of the circuit and the cost thereof (Dhanalakshmi and Rajasekar 2018; Papadopoulou 2013). On the other hand, authors of Stonier et al. (2023) and Dhimish et al. (2018b) did use separate gate drive circuits, however, did not use a universally implementable strategy for real-world applications. Kim et al. (2013) proposed a method for hotspot detection which involves frequent interruption in the MPPT and measuring the electrical parameters (shunt capacitance and resistance) of the individual substrings. This causes tracking losses due to frequent interruptions and hence the feasibility of implementation in the real-world application is low. Dhimish et al. (2018b) implemented a gate drive operation strategy based on the observed shading cycle during the day in the chosen site. However, in a real application, the shading patterns are highly unpredictable and are not generally cyclical other than in building integrated installations. Even in those installations, the shading pattern will widely vary over the year.

The present work proposes a bypass activation strategy in synergy with the MPPT algorithm, as merely adding a bypass switch will not serve any real purpose unless it is turned ON precisely at GMPP. However, as discussed earlier in this section, different MPPT algorithms are used based on the dynamism of the PSC, criticality of the project, and accuracy required, which can be broadly classified into two groups, viz. conventional MPPT algorithms and soft computing-based algorithms. Here, two different strategies are proposed that exploit the exclusive capabilities and advantages of the above two groups of MPPT algorithms. The conventional MPPT algorithms such as perturb and observe method (P&O), Incremental conductance (IC), and hill-climbing (HC) are faster than the soft computing-based algorithms because of their limited search space and simplicity of the algorithm. However, they are not effective in discriminating the GMPP from the local MPPs under mismatch conditions and hence often get stuck at local MPPs. Hence, in the case of the conventional MPPT algorithms, the gate of the active bypass switch may be activated based on the following strategy. The average module voltage shall be first calculated by dividing the array voltage at peak power (V_{MPP}) by the number of modules (N) in each string and subsequently this average voltage to be compared with the instantaneous voltages of individual modules, *i.e.*, V_{Module} (see Figure 14). The instantaneous voltage of the modules will reduce in the case of partial shading and other unfavorable environmental conditions and so will under the lower irradiances. Under lower irradiance, the comparator will generate a false positive signal if a fixed reference voltage is compared with the average module voltage; whereas in this case, the $V_{Average}$ will also reduce proportionally. Hence, the instantaneous voltage is compared dynamically with the average V_{MPP} of the modules, instead of a fixed reference voltage. The condition $V_{Module}<V_{Agerage}$ will indicate a mismatch condition and when a mismatch condition is identified and the peak power point is achieved by the MPPT algorithm, the gate pulse for the active bypass switch

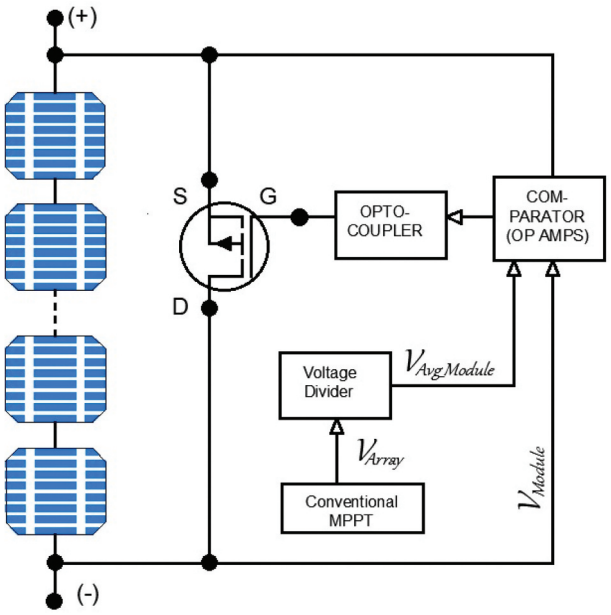


Figure 14. Schematic diagrams of proposed bypass circuit for conventional MPPT technology.

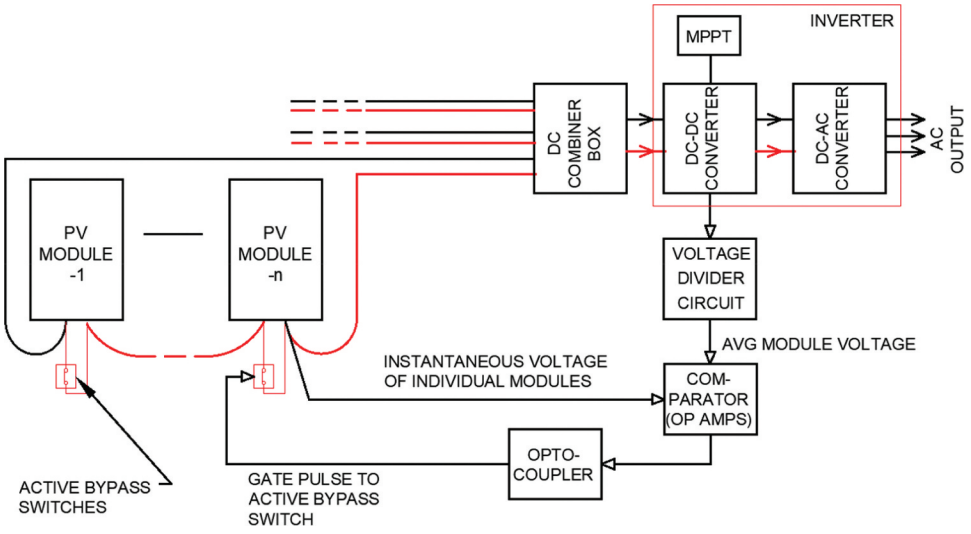


Figure 15. Schematic diagram of proposed bypass mechanism and its control circuit.

may be triggered through an optocoupler as illustrated in Fig. 14). A system level schematic is shown in Figure 15 for implementing the bypass strategy in synergy with the conventional MPPT technology. However, soft computing-based MPPT technology such as differential evolution, flashing fireflies, gray wolf optimization, particle swarm optimization, artificial bee colony, cuckoo search, and genetic algorithm looks for the GMPP in a wider search space and hence is able to make the distinction between the local MPPs and GMPP. Hence, in the case of soft computing-based MPPT technologies, it is easier to identify the mismatch conditions, if any, compared to the conventional MPPT technologies. The proposed bypass strategy for soft-computing-based MPPT algorithms is built on the above unique capability of the MPPT algorithms. A binary output may be obtained from the soft-computing-

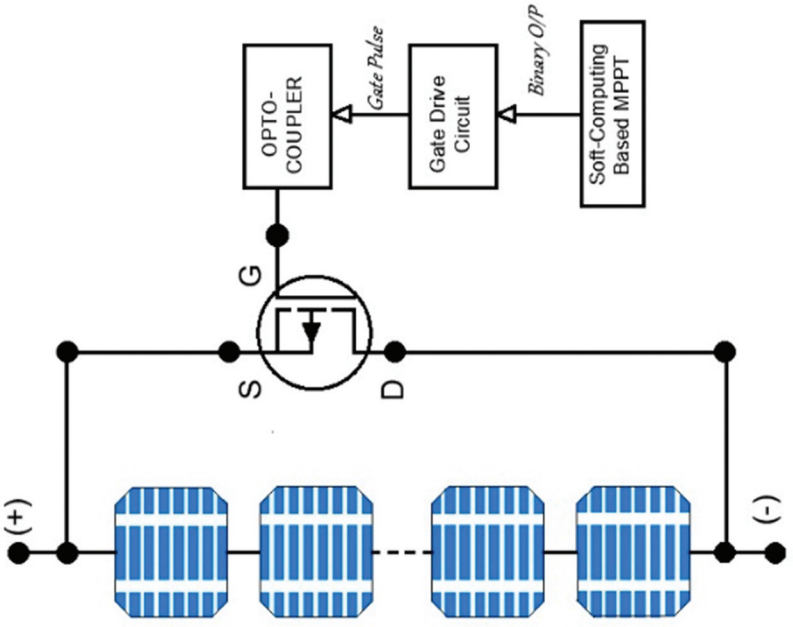


Figure 16. Schematic diagrams of proposed bypass circuit for conventional MPPT technology.

based MPPT algorithms indicating whether there is a mismatch condition that exists based on the condition $N_{MPPs} > 1$. And this binary output can be used to trigger the Gate drive of the active bypass switch through an opto coupler (see Figure 16).

Since both the bypass strategies discussed above shall be triggered by the MPPT technologies used in a particular system, it will ensure that the bypass circuit is positively turned on when it is needed the most, *i.e.*, at the MPP under partial shading and other mismatch conditions. This will help in increasing the shortfall of the existing bypass mechanism as demonstrated through the experimental results in the present work.

Conclusion

The experimental study of the effectiveness of the bypass circuit in a string array and in an SP array revealed that the hot spot prevention mechanism works effectively, in the case of single-string configuration, in all partial shading conditions other than single-cell shading scenarios. The hotspot temperature under the single-cell shading scenario was observed to increase up to 71.2°C as the bypass diode remained non-conductive. However, for two or more cell shading scenarios, the bypass diode went into conduction mode and bypassed 88 to 90.6% of the string current, and as a result, the temperature of the affected cells remained below 46.1°C at MPP.

However, a totally contrasting behavior was observed in the case of the series-parallel (SP) topology. The bypass diode was observed to be ineffective in all shading scenarios at the peak power point. As a result, hotspots were observed under all shading scenarios at the operating point (MPP). The temperature of the affected cells was observed to be between 64.2 and 70.7°C. The bypass diode entered conduction mode at a region far away from the operating point (*i.e.*, MPP). This indicates that the extant bypass mechanism is purely ornamental and serves no real purpose in protecting the cells

under partial shading conditions in the SP array under partial shading conditions. Hence, to improve the effectiveness or rather make the bypass mechanism functional in real installations, a novel bypass-circuit control strategy is proposed.

Future scope of work

The present work mainly focuses on the experimental study of the operation of the extant bypass circuit in the two most widely used array configurations used in real-world scenarios. The study would have looked incomplete without a discussion of the potential solutions to the shortfalls observed in the present state of technology. However, due to the journal's restriction on the length of the article, the experimental validation of the proposed solution is excluded from the present article, which may be taken up as the future scope of work.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Swatilekha Ghosh is a research scholar at Delhi Technological University and pursuing her Ph.D. in the Environmental Engineering Department. Her research interests include Increasing photovoltaic (PV) system reliability, PV waste reduction, and assessment and elimination of ecotoxicity of PV technologies.

Santosh Kumar Singh has held numerous academic and administrative positions in his illustrious career and is the recipient of a plethora of awards and honors for his exemplary contribution to the field of environmental engineering from various national and international bodies. His research interests include environmental engineering, water pollution control, water resource management, and environmental biotechnology. Presently Prof. Singh is gracing the position of the Vice Chancellor of Rajasthan Technological University.

Vinod Kumar Yadav received B.Tech. degree in Electrical Engineering from the Institute of Engineering and Technology (IET), Bareilly, India in 2003, M.Tech. degree in power system engineering from the National Institute of Technology (NIT), Jamshedpur, India in 2005 and a Ph. D. degree in power system Engineering from the Indian Institute of Technology (IIT), Roorkee, India in 2011. Since 2011, he has been associated with various technical universities and involved in teaching electrical engineering and also held various academic and administrative positions. He is a recipient of various awards and recognitions for his exemplary contribution as a researcher and academician. Currently, he is Professor of Electrical Engineering at Delhi Technological University (previously Delhi College of Engineering), Delhi, India. His research interests include the optimization of renewable energy systems, power system policy and restructuring, distributed generation, and smart grid.

References

Ahmed, J., and Z. Salam. 2014. A maximum power point tracking (MPPT) for PV system using cuckoo search with partial shading capability. *Applied Energy* 119:118–30. doi:10.1016/j.apenergy.2013.12.062.

Bakır, H., Detection of faults in photovoltaic modules of SPPS in Turkey; infrared thermographic diagnosis and recommendations, J. Electr. Eng. Technol. 18 (2023) 1945–57

Bauwens, P., and J. Doutreloigne. 2014. Reducing partial shading power loss with an integrated smart bypass. *Solar Energy* 103:134–42. doi:10.1016/j.solener.2014.01.040.

Dhanalakshmi, B., and N. Rajasekar. 2018. Dominance square based array reconfiguration scheme for power loss reduction in solar PhotoVoltaic (PV) systems. *Energy Conversion and Management* 156:84–102. doi:10.1016/j.enconman.2017.10.080.

Dhimish, M., V. Holmes, P. Mather, and M. Sibley. Jun 2018a. Novel hot spot mitigation technique to enhance photovoltaic solar panels output power performance. *Solar Energy Materials and Solar Cells* 179:72–79. doi:10.1016/J.SOLMAT.2018.02.019.

Dhimish, M., V. Holmes, P. Mather, and M. Sibley. 2018b. Novel hot spot mitigation technique to enhance photovoltaic solar panels output power performance, sol. *Energy Mater Solar Cells* 179:72–79. doi:10.1016/j.solmat.2018.02.019.

Ghosh, S., V. K. Yadav, and V. Mukherjee. 2020. A novel hot spot mitigation circuit for improved reliability of PV module. *IEEE Transaction Device Mater Reliability* 20 (1):191–98. doi:10.1109/TDMR.2020.2970163.

Guerrieroa, P., P. Tricolib, and S. Daliento. 2019. A bypass circuit for avoiding the hot spot in PV modules. *Solar Energy181* 181:430–38. doi:10.1016/j.solener.2019.02.010.

Guerriero, P., M. Coppola, I. Spina, I. Matacena, and S. Daliento. 2017. *A voltage divider strategy for reducing the hot spot temperature in partially shaded solar panels*. doi:10.1109/ICCEP.2017.8004791.

International Energy Agency report IEA-PVPS T13-01, review of failures of photovoltaic modules, 2014. ISBN 978-3-906042-16-9.

IRENA and IEA-PVPS, 2016. End-of-life management: Solar photovoltaic panels.

Ishaque, K., and Z. Salam. 2013. A deterministic particle swarm optimization maximum power point tracker for photovoltaic system under partial shading condition. *IEEE Transactions on Industrial Electronics* 60:3195–206. doi:10.1109/TIE.2012.2200223.

Janani, C., B. C. Babu, and K. Vijayakumar. 2023. An optimized reconfiguration technique of photovoltaic array using adaptive-JAYA optimization, Energy sources, part A: Recovery. *Utilization, and Environmental Effects* 45 (2):3777–810. doi:10.1080/15567036.2023.2197853.

Kermadi, M., Z. Salam, A. M. Eltamaly, J. Ahmed, S. Mekhilef, C. Larbes, and E. M. Berkouk. 2020. Recent developments of MPPT techniques for PV systems under partial shading conditions: A critical review and performance evaluation, IET renew. *Power Gener* 14 (17):3401–17. doi:10.1049/iet-rpg.2020.0454.

Kim, K. A., P. T. Krein, G.-S. Seo, B.-H. Cho. 2013. Photovoltaic ac parameter characterization for dynamic partial shading and hot spot detection, in Proc. IEEE Appl. Power Electron. Conf. 109–15. Long Beach, CA, USA.

Kumar, N., I. Hussain, B. Singh, and B. Panigrahi. 2017. Rapid MPPT for uniformly and partial shaded PV System by using JayaDE algorithm in highly fluctuating atmospheric conditions. *IEEE Transaction Industrial Information* 13 (5):2406–16. doi:10.1109/TII.2017.2700327.

Lappalainen, K., and S. Valkealahti. 2017. Effects of PV array layout, electrical configuration and geographic orientation on mismatch losses caused by moving clouds, sol. *Energy* 144:548–55. doi:10.1016/j.solener.2017.01.066.

Mahmoudi, S., N. Huda, Z. Alavi, M. T. Islam, and M. Behnia. 2019. End-of-life photovoltaic modules: A systematic quantitative literature review, Resour. *Conservation Recycl* 146:1–16. doi:10.1016/j.resconrec.2019.03.018.

Mathur, N., S. Singh, and J. W. Sutherland. 2020. Promoting a circular economy in the solar photovoltaic industry using life cycle symbiosis, Resour. *Conservation Recycl* 155:104649. doi:10.1016/j.resconrec.2019.104649.

Messenger, R. A., and A. Abtahi. 2010. *Photovoltaic systems Engineering*. 3rd ed. Boca Raton: CRC Press.

Mohanty, S., B. Subudhi, and P. K. Ray. 2016. A new MPPT design using grey wolf optimization technique for photovoltaic system under partial shading conditions. *IEEE Transaction Sustainable Energy* 7 (1):181–88. doi:10.1109/TSTE.2015.2482120.

Pachauri, R., R. Singh, A. Gehlot, R. Samakaria, and S. Choudhury. 2018. Experimental analysis to extract maximum power from PV array reconfiguration under partial shading conditions. *Engineering Science and Technology, an International Journal* 22 (1):109–30. doi:10.1016/j.jestch.2017.11.013.

Palawat, K. P., V. K. Yadav, R. L. Meena, and S. Ghosh. 2021. *Experimental investigation of performance of PV array topologies under simulated PSCs, applied soft computing and embedded System applications in solar Energy*. CRC Press.

Pannebakker, B. B., A. C. D. Waal, and W. G. J. H. M. V. Sark. 2017. Photovoltaics in the shade: One bypass diode per solar cell revisited, Prog. *Photovoltaic Research Application* 25 (10):836–49. doi:10.1002/pip.2898.

Papadopoulou, E. 2013. *Photovoltaic Industrial systems- an environmental approach*. Berlin, Heidelberg: Springer. 978-3-642-26664-5, Published: 27 February 2013

Stonier, A. A., R. Harish, M. Srinivasan, and D. Sarathkumar. 2023. An extensive critique on fault-tolerant systems and diagnostic techniques intended for solar photovoltaic power generation, Energy sources, part A: Recovery. *Utilization, and Environmental Effects* 45 (1):1856–73. doi:10.1080/15567036.2023.2183282.

Sundareswaran, K., S. Peddapati, and S. Palani. 2014. MPPT of PV systems under partial shaded conditions through a colony of flashing fireflies. *IEEE Transactions on Energy Conversion* 29:463–72.

Sundareswaran, K., P. Sankar, P. S. R. Nayak, S. P. Simon, and S. Palani. 2015. Enhanced energy output from a pv system under partial shaded conditions through artificial bee colony. *IEEE Transaction Sustainable Energy* 6 (1):198–209. doi:10.1109/TSTE.2014.2363521.

Vats, M. C., and S. K. Singh. 2014. E-Waste characteristic and its disposal. *International Journal Ecology Science Environment Engineering* 1:49–61.

Wang, J., and P. C. Hsu. 2011. An investigation on partial shading of PV modules with different connection configurations of PV cells. *Energy* 36 (5):3069–78. doi:10.1016/j.energy.2011.02.052.

Yadav, A. S., R. K. Pachauri, and Y. K. Chauhan. 2016. Comprehensive investigation of PV arrays with puzzle shade dispersion for improved performance, sol. *Solar Energy* 129:256–85. doi:10.1016/j.solener.2016.01.056.

Yadav, A. S., V. K. Yadav, V. Mukherjee, and S. Ghosh. 2021. Performance investigation of different bypass diode topology based sdk-pv arrays under partial shading conditions. In *Innovations in electrical and electronic engineering. Lecture notes in electrical engineering*, 261–70. Singapore 661: Springer. doi:10.1007/978-981-15-4692-1_20.

Kumar Girish, 0000-0002-0448-8918
Delhi Technological, University, India, girishkumar{at}dce.ac.in
Mishra Ravi, 0009-0004-2524-1158
Delhi Technological University, Delhi, India, ravitomeet{at}gmail.com
Khan Tayyab, 0000-0001-6539-7473
Delhi Technological University, Delhi, India, tayyabkhan632{at}gmail.com

Availability analysis of vertical milling centre using Markov approach and Monte Carlo simulations

Keywords

availability, reliability, Markov approach, Monte Carlo, redundancy

Abstract

The purpose of this article is to investigate the availability of vertical machining centre using a Markovian technique and Monte Carlo simulation (MSC). Availability is a critical performance metric for industrial systems. Conventional methodologies focus for steady-state availability evaluation of mechanical systems. The research analyses transient availability assessment for four different system configurations. Monte Carlo simulation modelling is used to compare the results and future scope is suggested to use the developed MCS based algorithms/codes for non-exponential (time dependant) failure and repair time distributions. The research also investigates the influence of active and passive redundancy on availability, indicating that for the vertical machining centre, parallel architecture with standby redundancy outperforms active load sharing. The chapter includes a sensitivity study that modifies the repair rates of the ball screw and sub-assembly to make the component selection process easier for engineers. The authors believe that this chapter will be useful to maintenance and practising engineers because it will assist them in making informed decisions about system availability, developing maintenance/replacement policies, and determining the repair level required to achieve the desired system availability.

1. Introduction

In today's world, innovative technologies are used to increase system productivity while decreasing manufacturing costs. The optimal technologies should be highly dependable, conveniently accessible, and maintainable in order to return the system to its original condition in the event of a manufacturing failure. Lowering the system's load can also improve its availability. The significance of system availability is clear in today's culture, as the globe grows increasingly reliant on modern technology systems that need intricate operations and revolutionary administration.

A balance must be struck between availability and other resources, such as cost, volume, and weight, when building a system for dependability and availability. From home to industrial uses, automated systems have become an increasingly important part of our everyday lives, increasing our reliance on them.

Critical system failures, such as those in air traffic control, nuclear reactors, or hospital patient monitoring, can have devastating repercussions. As a result, enterprises are always working to decrease the likelihood of failure and boost system availability using quantitative analytical methodologies based on industrial engineering and operational

Safety and Reliability of Systems and Processes, Summer Safety and Reliability Seminar 2023.
© Gdynia Maritime University. All rights reserved.
DOI: 10.26408/srsp-2023-09.

research ideas.
This evaluation needs precise knowledge of constant characteristics, such as the failure and repair rates of systems/subsystems prone to random failures caused, among other things, by bad design, a lack of operating skills, or manufacturing technologies.

2. Literature review

In this section, literature review of problem solved using Monte Carlo simulation (MCS) or Markovian approaches (MA) are described.

2.1. Application of Monte Carlo simulation and Markovian approach

Ahmed et al. (Ahmed et al., 1989) suggested a semi-Markov modelling method to assess performance measures in queuing networks during the design of Computer Integrated Manufacturing Systems. This technique lowers complexity by grouping states together and only observing states during transition periods. Mean residence durations are determined using probability distribution functions of conditional state occupancy periods. Bounds of performance metrics are met, resulting in a straightforward and effective computational method. Alexander (Alexander, 2003) proposed the concepts of Monte Carlo simulations and demonstrated their application in determining the reliability of pump systems. Attar et al. (Attar et al., 2017) suggested a simulation-based optimisation technique for multi-objective joint availability redundancy allocation issues in multi-component series-parallel setups using active, cold, and hot backup strategies.
Borgonovo et al. (Borgonovo et al., 2000) introduced a Monte Carlo method for assessing plant maintenance strategies and working routines while adhering to economic limitations. Cadini et al. (Cadini et al., 2017) developed a repair model that combines an extreme weather stochastic model with a genuine cascading failure emulator to measure the effect of extreme weather events on power system reliability/availability. They also presented a Monte Carlo model-based approach for predicting the failure likelihood of degraded components incorporated within an optimum condition-based component replacement strategy, with an application to fatigue degradation. Çekyay and Özekici (Çekyay & Özekici, 2015) investigated the system dependability, mean time to failure, and steady-state availability of coherent systems, as well as the series connection of ***k***-out-of-***n*** standby subsystems with exponentially dispersed component lives. Garg et al. (Garg et al., 2010) used probability analysis and the extra variable method to explain the availability of a crankcase production system in the car business. Ge et al. (Ge et al., 2014) used Monte Carlo modelling to address the reliability evaluation issue of low and high Distribution Generation (DG) penetration levels in active distribution systems and discovered that DG incorporation can enhance system reliability when run actively. Lin and Donaghey (Lin & Donaghey, 1993) suggested using Monte Carlo simulation to determine the minimum cut sets and system reliability based on the reliability block diagram and the life distribution of each component. Based on Monte Carlo simulation and statistical analysis, Maciejewski and Caban (Maciejewski & Caban, 2008) provided an efficient method for predicting the availability of a repairable system using a Beta distribution. Chawla and Kumar (Chawla & Kumar, 2013) proposed a transient availability evaluation for a mechanical system with condition based maintenance strategy using Markov method. Varghese and Kumar (Varghese & Kumar, 2014) also used Markov approach for availability assessment for a mechanical system with opportunistic maintenance scenario.
Okafor et al. (Okafor et al., 2016) used a Markovian method to evaluate the availability of a thermal power station's steam and gas turbine units. Sharma and Kumar (Sharma & Kumar, 2008) presented a paper to examine the need for maintenance practices that reduce both sudden sporadic failures and operation and maintenance costs. Kumar et al. (Kumar et al., 2013) suggested a semi-Markov process-based analytical approach for steady-state availability evaluation.
Sharma and Vishwakarma (Sharma & Vishwakarma, 2014) analyzed the performance measures of complex repairable systems involving reduced states using time-homogeneous Markov processes. Lognathan et al. (Lognathan et al., 2015) attempted steady-state availability evaluation using semi-Markov analytical approach for a vertical milling centre. Kumar et al. (Kumar et al., 2019) suggested simulative approach for reliability and availability assessment for repairable mechanical systems.

2.2. Gaps in literature

The literature on evaluating the availability of mechanical systems was examined, but the findings derived from the different methods proposed by researchers were having several limitations. Due to the assumptions made in the modelling the results were significantly different from what were anticipated.
Majority of the research work was focussed for steady-state availability evaluation. In some of the papers non-exponential distributions were approximated to exponential distribution so that Markov approach can be applied. Monte Carlo simulation approach was applied to simple mechanical systems with limited number of states. Monte Carlo simulation is given rare focus in spite of being one of the most powerful approaches in reliability and availability evaluation which can handle time-dependent failure and repair rates.

2.3. Problem statement

This chapter suggests that effective maintenance planning and management can return failed systems to their original functioning state by performing numerous fixes and component changes in the least amount of time. The authors suggest a quantitative method for assessing system availability that employs a stochastic modelling namely Markov model. The Markov model creates transition states, which are then solved by algorithms. The results of these findings are confirmed using Monte Carlo simulation, a technique that was not widely used for availability assessment of complex systems.
This chapter is an extension of the work of Loganathan et al. (Loganathan et al., 2016) wherein the availability evaluation was carried using semi-Markov approach in a case study of vertical milling centre (VMC). However, only active redundancy is considered in their work. Moreover, the work was done for steady-state assessment. In the present work, system is evaluated both in active and standby redundancy by using Markovian approach which is more of a transient system approach and result obtained are more real life compliant. Also, the use of Monte Carlo simulation which is being used here to back the result by further relaxing norms and assumptions.
The chapter approach is defined in the next section with the methodology being employed. A stepped account of work is presented below:

- modelling a series configuration and parallel configuration (passive redundancy) system as an example by using Markovian approach and results are compared with MCS,
- modelling case study system of VMC in active and passive redundancy using Markov model and MCS,
- sensitivity analysis to acknowledge dependence on repair rates, and also identifying critical components to availability.

3. Methodology

In this section, methodology used for availability evaluation for three different cases is presented using Markovian approach and Monte Carlo simulation. Following 4 cases are considered:

- two components in parallel configuration (passive redundancy) – Markov approach and MCS,
- two components in series configuration – Markov approach and MCS,
- VMC (3 components) – series-parallel case (active redundancy) – Markov approach,
- VMC (3 components) – series-parallel case (passive redundancy) – Markov approach and MCS.

3.1. Two components in parallel configuration-passive redundancy

Availability evaluation for system containing two components in parallel configuration (Figure 1) is discussed.

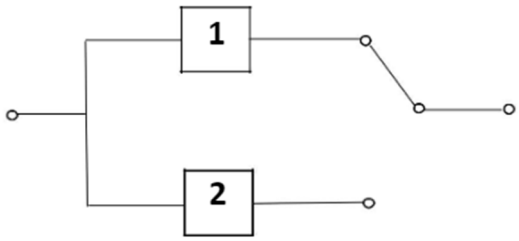


Figure 1. A system containing two components in parallel connection.

3.1.1. Markovian approach

System availability evaluation using Markov model is performed as following.
Step i: First all the possible states are identified to develop the model.
Step ii: Next the Markov model transition diagram is constructed.

Step iii: Transition rates are gathered at the component level.
Step iv: Rate equations are formulated in the form of Ordinary differential equations.
Step v: State probability solutions are evaluated using MATLAB and used for availability evaluation.
Table 1 presents all the possible states, transitions among states and system level status in various states. Using Table 1, a state transition diagram is constructed for Markov model and the same is given in Figure 2. Bold line in Figure 2 represents transition from state when one of the components fails, while dotted line is used to represent transition from one state to another when one of the failed components is repaired to original working condition.

Table 1. Development of system states (2-component in passive redundancy)

State	Sub-Assembly 1	Sub-Assembly 2	System Response	Transition to State
1	O	S	W	3
2	S	O	W	4
3	F	O	W	2, 5
4	O	F	W	1, 5
5	F	F	NW	4, 3

O – operating, S – standby, F – failed, W – working, NW – non-working.

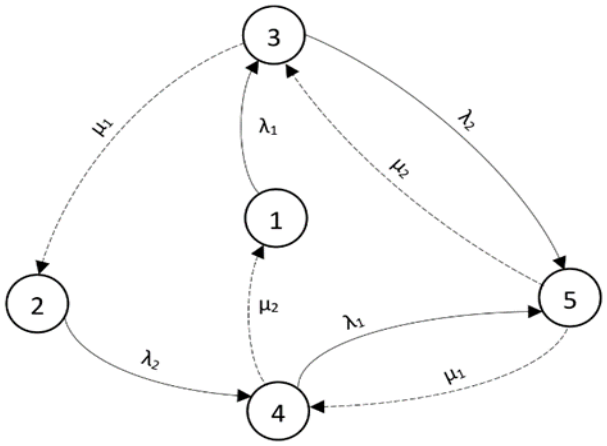


Figure 2. Transition state diagram for system containing two components in parallel arrangement (passive redundancy).

For the system sub-assemblies the failure time and the repair time distributions are Weibull and Lognormal respectively. The Weibull and Lognormal parameters values are given in Table 2. As Markov approach is based on constant failure rate (λ) and constant repair rate (μ). The equations (1), (2) and (3) are used to evaluate distributions parameters.

$$\lambda = \frac{1}{\text{MTTF}} \text{ and } \mu = \frac{1}{\text{MTTR}}. \quad (1)$$

The values of MTTF and MTTR are evaluated from the following results:

$$\text{MTTF} = \theta \lceil (1 + 1/\beta) \quad (2)$$

$$\text{MTTR} = e^{(\mu + \frac{\sigma^2}{2})}. \quad (3)$$

The data presented in Table 2 is taken from (Loganathan et al., 2016).

Table 2. Weibull and Lognormal distribution parameters (Loganathan et al., 2016)

Distribution	CDF	Parameters
$W(\beta_{13}, \theta_{13})$	1→3	$\beta_{13} = 1.5$; $\theta_{13} = 2438$
$W(\beta_{24}, \theta_{24})$	2→4	$\beta_{24} = 1.5$; $\theta_{24} = 2438$
$W(\beta_{35}, \theta_{35})$	3→5	$\beta_{35} = 1.5$; $\theta_{35} = 2438$
$Ln(\mu_{32}, \sigma_{32})$	3→2	$\mu_{32} = 3.4$; $\sigma_{32} = 1.79$
$W(\beta_{45}, \theta_{45})$	4→5	$\beta_{45} = 1.5$; $\theta_{45} = 2438$
$Ln(\mu_{41}, \sigma_{41})$	4→1	$\mu_{41} = 1.5$; $\sigma_{41} = 1.79$
$Ln(\mu_{53}, \sigma_{53})$	5→3	$\mu_{53} = 3.4$; $\sigma_{53} = 1.79$
$Ln(\mu_{54}, \sigma_{54})$	5→4	$\mu_{54} = 3.4$; $\sigma_{54} = 1.79$

The failure rate and the repair rate are evaluated using equations (1), (2) and (3). The same are given here:

λ_1 = failure rate of Sub-Assembly 1 = 0.00061648,
λ_2 = failure rate of Sub-Assembly 2 = 0.00061648,
μ_1 = repair rate of Sub-Assembly 1 = 0.00672,
μ_2 = repair rate of Sub-Assembly 2 = 0.00672.

State transition equations for the given system are derived as following:

$$\frac{dy_1}{dt} = -\lambda_1 y_1 + \mu_2 y_2 \quad (4)$$

$$\frac{dy_2}{dt} = -\lambda_2 y_2 + \mu_1 y_3 \quad (5)$$

$$\frac{dy_3}{dt} = -(\mu_1 + \lambda_2) y_3 + \lambda_1 y_1 + \mu_2 y_5 \quad (6)$$

$$\frac{dy_4}{dt} = -(\lambda_1 + \mu_2) y_4 + \mu_1 y_5 + \lambda_2 y_2 \quad (7)$$

$$\frac{dy_5}{dt} = -(\mu_1 + \mu_2) y_5 + \lambda_2 y_3 + \lambda_1 y_4. \quad (8)$$

System availability (series configuration) is given as:

$$A(t) = P(1) + P(2) + P(3) + P(4). \quad (9)$$

3.1.2. Monte Carlo simulation

MCS approach is presented here which will be used to compare the result obtained from MA. Steps involved are as follows:

- step i: identification of working/ non-working states,
- step ii: developing an algorithm for implementation in MATLAB,
- step iii: generation of MATLAB codes based on algorithm developed.

Algorithm steps

(i) Identify the working states and non-working states at system level.
(ii) Start by generating a random number and use them to generate TTF (time to failure) and TTR (time to repair).
(iii) First develop TTF value for component 1 and then similarly for component 2.
(iv) Compare the values of two and decide which component fails first, which leads to transition from working states to non-working states.
(v) Store the down time of machine in some variable, say DT.
(vi) Similarly, repeat the steps till all possible cases are accounted for and total down time (TDT) is measured.
(vii) Putting the values in Availability equation as given below, Availability of system is calculated.
(viii) Running the algorithm in MATLAB for a number of iterations, till value of Availability converges towards a constant value.

Formulae used in MCS evaluation are given below. These are derived from Billinton and Allan (Billinton & Allan, 2007).

$$\text{TTF} = -\left(\frac{1}{\lambda}\right) \cdot \log(\text{YF}) \quad (10)$$

$$\text{TTR} = -\left(\frac{1}{\mu}\right) \cdot \log(\text{YR}) \quad (11)$$

where:
λ = corresponding failure rate,
μ = corresponding repair rate,
YF/YR = random variables in interval (0,1).

In a parallel system, as described in Markovian approach above, the system is down only when both the components in parallel fails simultaneously. The system is functional whenever one of the components is operating. As can be seen from the Figure 3 that system is down when the component failures coincide during the mission time, rest of the time it is up and working.

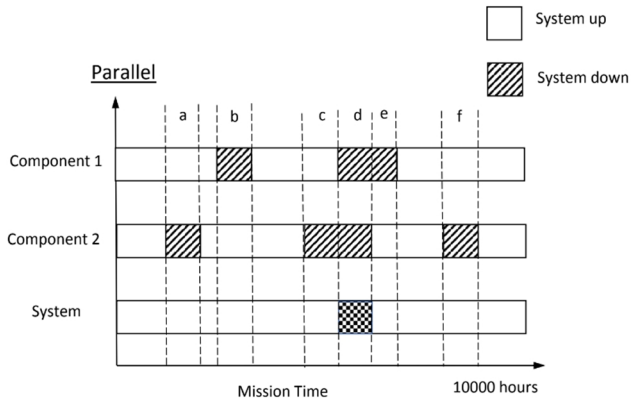


Figure 3. Systems working/non-working states (two component in passive redundancy).

Using MCS approach availability is evaluated as:

$$A(t) = \frac{\text{total uptime}}{\text{mission time}}. \quad (12)$$

3.2. Two components in series configuration

System availability evaluation using Markov model is performed, with similar methodology followed for parallel configuration.

3.2.1. Markovian approach

Table 3 shows the transition possible along with the definition of state and system response. All these represent the possible states and their transitions to one another.

Table 3. Development of system states (2 component in series configuration)

State	Sub-Assembly 1	Sub-Assembly 2	System Response	Transition to State
1	O	O	O	3, 2
2	O	F	F	1
3	F	O	F	1

Using Table 3 a Markov model diagram is constructed as given in Figure 4.
Table 4 shows the distribution used along with the

transition of state and the parameters associated. The data presented in Table 4 is derived taken from Loganathan et al. (Loganathan et al., 2016).

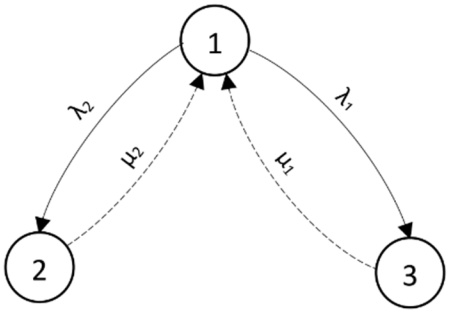


Figure 4. Transition state diagram for system with two components in series connection.

Table 4. Weibull and Lognormal distribution parameters (Loganathan et al., 2016)

Distribution	CDF	Parameters
$W(\beta_{12}, \theta_{12})$	1→2	$\beta_{12} = 2.72$; $\theta_{12} = 3315$
$W(\beta_{13}, \theta_{13})$	1→3	$\beta_{13} = 1.9$; $\theta_{13} = 1828$
$Ln(\mu_{21}, \sigma_{21})$	2→1	$\mu_{21} = 2.88$; $\sigma_{21} = 1.55$
$Ln(\mu_{31}, \sigma_{31})$	3→1	$\mu_{31} = 3.4$; $\sigma_{31} = 1.79$

For this work, the failure rate and the repair rate for the components are evaluated using the equations (1), (2) and (3). The rates are given here:

λ_1 = failure rate of Sub-Assembly 1 = 0.0003391,
λ_2 = failure rate of Sub-Assembly 2 = 0.00061648,
μ_1 = repair rate of Sub-Assembly 1 = 0.01688,
μ_2 = repair rate of Sub-Assembly 2 = 0.00672.

State transition equations for the given system are as following:

$$\frac{dy_1}{dt} = -(\lambda_1 + \lambda_2)\, y_1 + \mu_2 y_2 + \mu_1 y_3 \quad (13)$$

$$\frac{dy_2}{dt} = -(\mu_2) y_2 + \lambda_2 y_1 \quad (14)$$

$$\frac{dy_3}{dt} = -(\mu_1) y_3 + \lambda_1 y_1. \quad (15)$$

System availability (series configuration) is given as:

$$A(t) = P(1). \quad (16)$$

MCS approach is presented below which will be used to compare the result obtained from MA. Steps followed are similar to previous sections. TTF and TTR are calculated using equations (10) and (11).

The working states are easily identifiable in series system. The only possibility being both the component up and working. Even when one of them fails the whole system goes to non-working state and system is down. Refer Figure 5, whenever either component fails, the system is down.

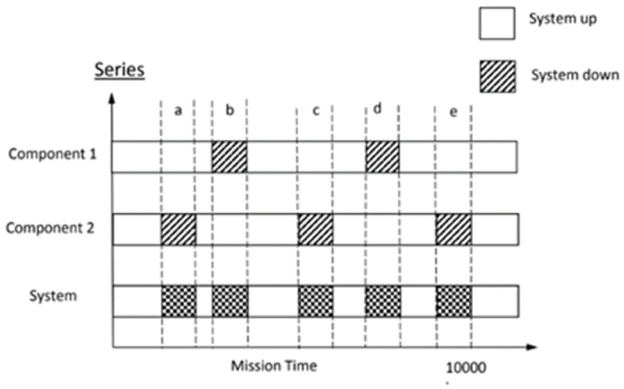


Figure 5. Systems working/non-working states (two components in series).

The availability is evaluated using the equation (12).

3.3. Vertical milling centre (3 components in series-parallel arrangement)

In this section, availability evaluation of the vertical milling centre (Figure 6) is done taking both cases of active and passive redundancy. First availability assessment is done MA and in the next MCS is employed to compare the results. However, MCS is carried out for passive redundancy only as dynamic variation in failure rate in case of active redundancy couldn't be easily estimated.

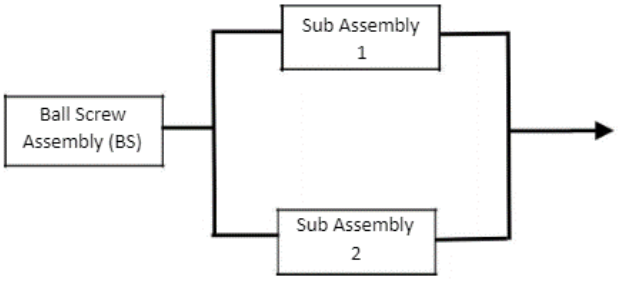


Figure 6. Vertical milling centre system.

3.3.1. Availability modelling – active redundancy (Markovian approach)

System availability evaluation using Markov Model is performed. Table 5 presents possible

states, its transitions and system response. Using Table 5, a state transition diagram for Markov model is generated.

Table 5. Development of system states (vertical milling centre system – active redundancy)

State	Ball Screw	Sub-Assembly 1	Sub-Assembly 2	System Response	Transition to state
1	O	O	O	W	2, 3, 5
2	O	O	F	W	1, 4, 6
3	O	F	O	W	1, 4, 7
4	O	F	F	NW	2, 3
5	F	O	O	NW	1
6	F	O	F	NW	2
7	F	F	O	NW	3

The same is shown in Figure 7. Bold line in Figure 7 represents transition from state when one of the components fails, while dotted line is used to represent transition from one state to another when one of the failed components is repaired to original working conditions.

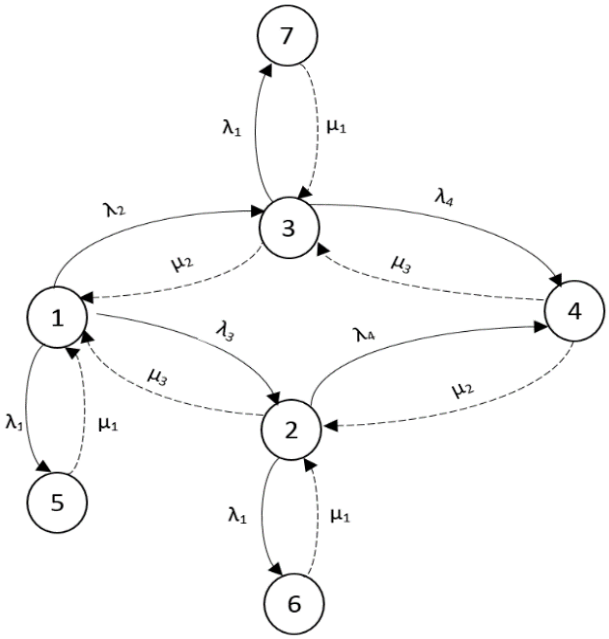


Figure 7. Markov model transition diagram for mechanical subsystem of twin-spindle VMC in load sharing (active redundancy).

Table 6 shows the distribution used along with the transition of state and the parameters associated. The data presented in Table 6 is derived from Loganathan et al. (Loganathan et al., 2016).

Table 6. Weibull and Lognormal distribution parameters (Loganathan et al., 2016)

Distribution	CDF	Parameters
$W(\beta_{12}, \theta_{12})$	1→2	$\beta_{12} = 1.5$; $\theta_{12} = 2438$
$W(\beta_{13}, \theta_{13})$	1→3	$\beta_{13} = 1.5$; $\theta_{13} = 2438$
$W(\beta_{15}, \theta_{15})$	1→5	$\beta_{15} = 2.72$; $\theta_{15} = 3315$
$Ln(\mu_{21}, \sigma_{21})$	2→1	$\mu_{21} = 3.4$; $\sigma_{21} = 1.79$
$W(\beta_{24}, \theta_{24})$	2→4	$\beta_{24} = 1.9$; $\theta_{24} = 1828$
$W(\beta_{26}, \theta_{26})$	2→6	$\beta_{26} = 2.72$; $\theta_{26} = 3315$
$Ln(\mu_{31}, \sigma_{31})$	3→1	$\mu_{31} = 3.4$; $\sigma_{31} = 1.79$
$W(\beta_{34}, \theta_{34})$	3→4	$\beta_{34} = 1.9$ $\theta_{34} = 1828$
$W(\beta_{37}, \theta_{37})$	3→7	$\beta_{37} = 2.72$ $\theta_{37} = 3315$
$Ln(\mu_{42}, \sigma_{42})$	4→2	$\mu_{42} = 3.4$; $\sigma_{42} = 1.79$
$Ln(\mu_{43}, \sigma_{43})$	4→3	$\mu_{43} = 3.4$; $\sigma_{43} = 1.79$
$Ln(\mu_{51}, \sigma_{51})$	5→1	$\mu_{51} = 2.88$; $\sigma_{51} = 1.55$
$Ln(\mu_{62}, \sigma_{62})$	6→2	$\mu_{62} = 2.88$; $\sigma_{62} = 1.55$
$Ln(\mu_{73}, \sigma_{73})$	7→3	$\mu_{73} = 2.88$; $\sigma_{73} = 1.55$

The failure rate and the repair rate for the components are evaluated and given here:

λ_1 = failure rate of Ball Screw = 0.0003391,
λ_2 = failure rate of Sub-Assembly 1 with load sharing = 0.0004544,
λ_3 = failure rate of Sub-Assembly 2 with load sharing = 0.0004544,
λ_4 = failure rate of Sub-Assemblies without load sharing = 0.00061648,
μ_1 = repair rate of Ball Screw = 0.01688,
μ_2 = repair rate of Sub-Assembly 1 = 0.00672,
μ_3 = repair rate of Sub-Assembly 2 = 0.00672.

State transition equations for the given system are derived as following:

$$\frac{dy_1}{dt} = -(\lambda_2 + \lambda_3 + \lambda_1)\, y_1 + \mu_1 y_5 + \mu_3 y_2 + \mu_2 y_3 \quad (17)$$

$$\frac{dy_2}{dt} = -(\lambda_1 + \lambda_4 + \mu_3)\, y_2 + \lambda_3 y_1 + \mu_2 y_4 + \mu_1 y_6 \quad (18)$$

$$\frac{dy_3}{dt} = -(\lambda_1 + \mu_2 + \lambda_4) y_3 + \mu_1 y_7 + \lambda_2 y_1 + \mu_3 y_4 \quad (19)$$

$$\frac{dy_4}{dt} = -(\mu_3 + \mu_2) y_4 + \lambda_4 y_3 + \lambda_4 y_2 \quad (20)$$

$$\frac{dy_5}{dt} = -(\mu_1) y_5 + \lambda_1 y_1 \quad (21)$$

$$\frac{dy_6}{dt} = -(\mu_1) y_6 + \lambda_1 y_2 \quad (22)$$

$$\frac{dy_7}{dt} = -(\mu_1) y_7 + \lambda_1 y_3. \quad (23)$$

Availability of system is defined as:

$$A(t) = P(1) + P(2) + P(3). \quad (24)$$

3.3.2. Availability modelling – passive redundancy (Markovian approach)

System availability evaluation using Markov Model is performed. Table 7 presents possible states, its transitions and system response.

Table 7. Development of system states (VMC – passive redundancy)

State	Ball Screw	Sub-Assembly 1	Sub-Assembly 2	System Response	Transition to state
1	O	O	S	W	2, 7
2	F	O	S	NW	1
3	O	O	F	W	9, 4, 1
4	F	O	F	NW	3
5	O	S	O	W	3, 6
6	F	S	O	NW	5
7	O	F	O	W	5, 8, 9
8	F	F	O	NW	7
9	O	F	F	NW	7, 3

Using Table 7, a state transition diagram for Markov model is generated. The same is shown in Figure 8. Bold line in Figure 8 represent transition from state when one of the component fails, while dotted line is used to represent transition from one state to another when one of the failed component is repaired to original working condition.

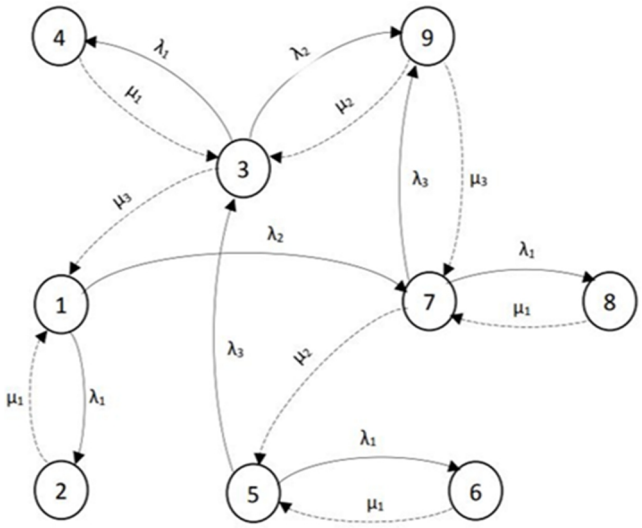


Figure 8. Markov model transition diagram for twin-spindle VMC (passive redundancy).

Table 8 shows the distribution used along with the transition of state and the parameters associated. The data presented in Table 8 is derived from Loganathan et al. (Loganathan et al., 2016).

Table 8. Weibull and Lognormal distribution parameters (Loganathan et al., 2016)

Distribution	CDF	Parameters
$W(\beta_{12}, \theta_{12})$	1→2	$\beta_{12} = 2.72$; $\theta_{12} = 3315$
$W(\beta_{17}, \theta_{17})$	1→7	$\beta_{17} = 1.5$; $\theta_{17} = 2438$
$Ln(\mu_{21}, \sigma_{21})$	2→1	$\mu_{21} = 2.88$; $\sigma_{21} = 1.55$
$Ln(\mu_{31}, \sigma_{31})$	3→1	$\mu_{31} = 3.4$; $\sigma_{31} = 1.79$
$W(\beta_{34}, \theta_{34})$	3→4	$\beta_{34} = 2.72$; $\theta_{34} = 3315$
$W(\beta_{39}, \theta_{39})$	3→9	$\beta_{39} = 1.5$; $\theta_{39} = 2438$
$Ln(\mu_{43}, \sigma_{43})$	4→3	$\mu_{43} = 2.88$; $\sigma_{43} = 1.55$
$W(\mu_{53}, \sigma_{53})$	5→3	$\beta_{53} = 1.5$; $\theta_{53} = 2438$
$W(\beta_{56}, \theta_{56})$	5→6	$\beta_{56} = 2.72$; $\theta_{56} = 3315$
$Ln(\mu_{65}, \sigma_{65})$	6→5	$\mu_{65} = 2.88$; $\sigma_{65} = 1.55$
$Ln(\mu_{75}, \sigma_{75})$	7→5	$\mu_{75} = 3.4$; $\sigma_{75} = 1.79$
$W(\beta_{78}, \theta_{78})$	7→8	$\beta_{78} = 2.72$; $\theta_{78} = 3315$
$W(\beta_{79}, \theta_{79})$	7→9	$\beta_{79} = 1.5$; $\theta_{79} = 2438$
$Ln(\mu_{87}, \sigma_{87})$	8→7	$\mu_{87} = 2.88$; $\sigma_{87} = 1.55$
$Ln(\mu_{93}, \sigma_{93})$	9→3	$\mu_{93} = 3.4$; $\sigma_{93} = 1.79$
$Ln(\mu_{97}, \sigma_{97})$	9→7	$\mu_{97} = 3.4$; $\sigma_{97} = 1.79$

The failure rate and the repair rate for the components are evaluated and given here:

λ_1 = failure rate of Ball Screw = 0.0003391,
λ_2 = failure rate of Sub Assembly 1 = 0.00061648,
λ_3 = failure rate of Sub Assembly 2 = 0.00061648,
μ_1 = repair rate of Ball Screw = 0.01688,
μ_2 = repair rate of Sub Assembly 1 = 0.00672,
μ_3 = repair rate of Sub Assembly 2 = 0.00672.

State transition equations for the given system are derived as following:

$$\frac{dy_1}{dt} = -(\lambda_1 + \lambda_2)\, y_1 + \mu_1 y_2 + \mu_3 y_3 \quad (25)$$

$$\frac{dy_2}{dt} = -\mu_1 y_2 + \lambda_1 y_1 \quad (26)$$

$$\frac{dy_3}{dt} = -(\lambda_1 + \lambda_2 + \mu_3) y_3 + \mu_2 y_9 + \lambda_3 y_5 + \mu_1 y_4 \quad (27)$$

$$\frac{dy_4}{dt} = -\mu_1 y_4 + \lambda_1 y_3 \quad (28)$$

$$\frac{dy_5}{dt} = -(\lambda_1 + \lambda_3)\, y_5 + \mu_1 y_6 + \mu_2 y_7 \quad (29)$$

$$\frac{dy_6}{dt} = -(\mu_1) y_6 + \lambda_1 y_5 \quad (30)$$

$$\frac{dy_7}{dt} = -(\lambda_1 + \lambda_3 + \mu_2)\, y_7 + \lambda_2 y_1 + \mu_1 y_8 + \mu_3 y_9 \quad (31)$$

$$\frac{dy_8}{dt} = -\mu_1 y_8 + \lambda_1 y_7 \quad (32)$$

$$\frac{dy_9}{dt} = -(\mu_2 + \mu_3)\, y_9 + \lambda_2 y_3 + \lambda_3 y_7. \quad (33)$$

Availability of system is defined as:

$$A(t) = P(1) + P(3) + P(5) + P(7). \quad (34)$$

3.3.3. Availability modelling – passive redundancy (Monte Carlo simulation)

MCS approach is presented to compare the results obtained from MA. Steps followed are similar to previous sections. TTF and TTR are calculated using equations (10) and (11).

The working states are easily identifiable in series-parallel combination system. The possibility being both the subsystems/subassembly in series being up and working. Even when one of the subsystem/subassembly fails the whole system goes to non-working state and system is down. As seen in the Figure 9, whenever either subsystem fails, the system is down.

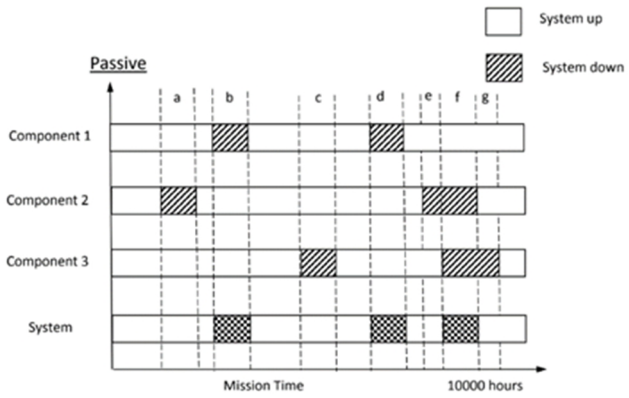


Figure 9. Possible states in series-parallel combination system.

4. Results and analysis

In this section, the result are obtained from study of 4 different systems using Markov model and Monte Carlo simulation and the same are discussed.

4.1. Transient availability results

Based on the Markov model, availability evaluation is done for the 4 different system models presented in previous section. The set of state equations are solved using MATLAB software (ODE 45 function) at time interval of 10,000 hrs and these state probability values are used to evaluate system availability. Also the result obtained using MCS approach except the case of VMC active redundancy case. All the results are presented in the Table 9. Also the results are presented in graphical form in Figure 10 to Figure 16. From the results under Table 9, it is clear that the Markov and MCS availability results are close. Therefore, MCS algorithm are giving the results comparable with closed form results of the Markovian approach.

Table 9. Transient availability results for 10,000 hrs

System type	System availability (Markov approach)	System availability (MCS approach)
two components in parallel configuration (passive redundancy)	0.9962	0.9937
two components in series configuration	0.8994	0.9051
vertical milling centre (3 components series-parallel case – active redundancy)	0.9751	NA
vertical milling centre (3 components series-parallel case – passive redundancy)	0.9766	0.9731

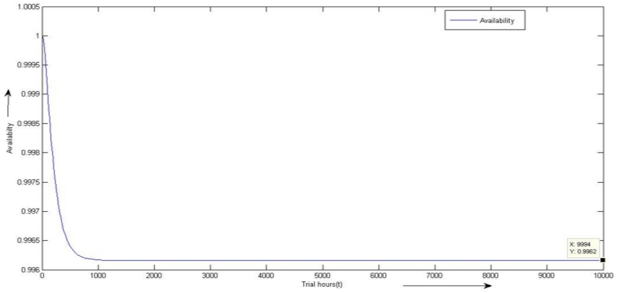


Figure 10. Availability Markov result of two components in parallel configuration (passive redundancy).

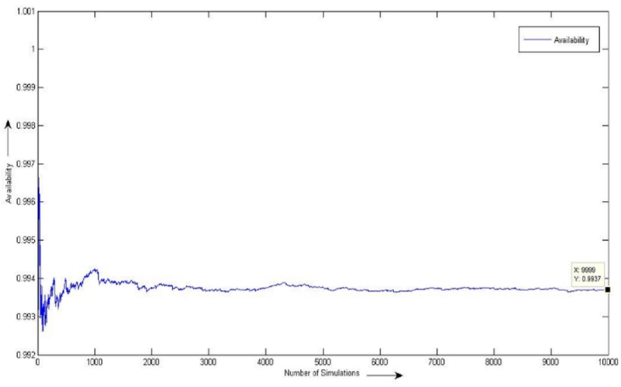


Figure 11. Availability MCS result of two components in parallel configuration (passive redundancy).

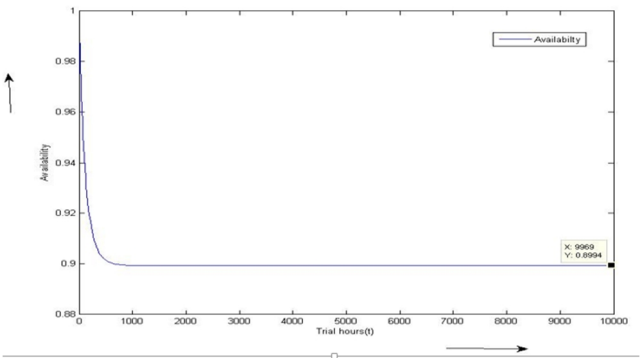


Figure 12. Availability Markov result of two components in series configuration.

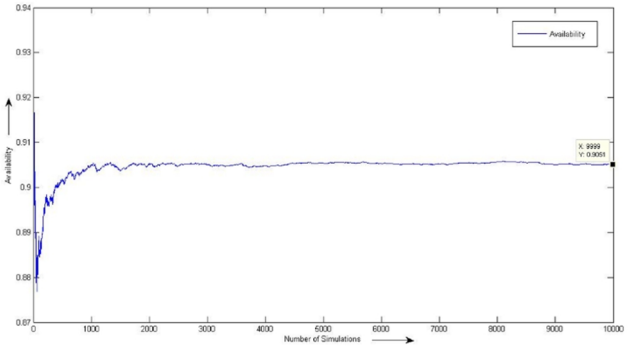


Figure 13. Availability MCS result of two components in series configuration.

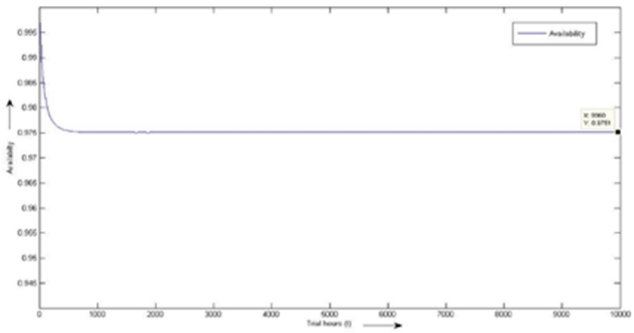


Figure 14. Availability Markov result of vertical milling centre (3 components series – parallel case – active redundancy).

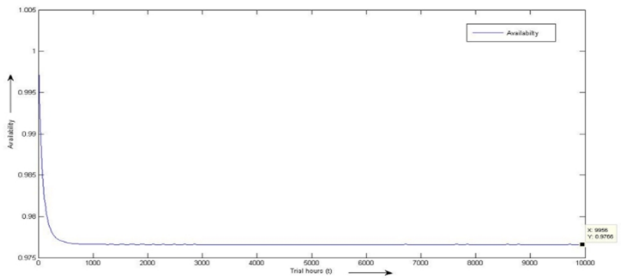


Figure 15. Availability Markov result of vertical milling centre (3 components – series –parallel case – passive redundancy).

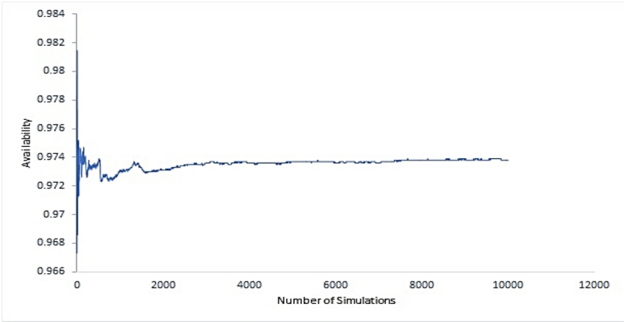


Figure 16. Availability MCS result of vertical milling centre (3 components) – series-parallel case (passive redundancy).

4.2. Sensitivity analysis – active redundancy

Sensitivity analysis is done by varying the value of failure and repair rate. At different value of failure rate and repair rate availability value changes. Here as only repair rate is in our control we vary the same. Sensitivity analysis is done as under taking λ_1 = 0.0003391, λ_2 = 0.0004544, λ_3 = 0.0004544, λ_4 = 0.00061648, μ_2 = 0.00672, μ_3 = 0.00672 as constant and changing the value of μ_1, the availability assessment is done and results presented in Table 10.

Table 10. Variation of availability with change in repair rate of ball screw assembly

μ_1	0.01688	0.02	0.04	0.08
Availability	0.9751	0.9781	0.9863	0.9904

Similarly, taking λ_1 = 0.0003391, λ_2 = 0.0004544, λ_3 = 0.0004544, λ_4 = 0.00061648, μ_1 = 0.01688, μ_3 = 0.00672 as constant and changing the value of μ_2, the availability values are evaluated and shown in Table 11.

Table 11. Variation of availability with change in repair rate of sub assembly

μ_2	0.00672	0.008	0.02	0.04
Availability	0.9751	0.9759	0.9785	0.9794

On the similar lines, taking λ_1 = 0.0003391, λ_2 = 0.0004544, λ_3 = 0.0004544, λ_4 = 0.00061648, μ_1 = 0.01S688, μ_2 = 0.00672 as constant and changing the value of μ_3 availability values are obtained. The same are shown in Table 12.

Table 12. Variation of availability with change in repair rate of sub assembly 2

μ_3	0.00672	0.008	0.02	0.04
Availability	0.9751	0.9759	0.9785	0.9794

4.3. Sensitivity analysis – passive redundancy

Sensitivity analysis is done by varying the value of failure and repair rate. At different value of failure rate and repair rate availability value changes. Here as only repair rate is can be controlled, we vary the same. Sensitivity analysis is done as under taking λ_1 = 0.0003391, λ_2 = 0.0004544, λ_3 = 0.0004544, μ_2 = 0.00672, μ_3 = 0.00672 as constant and changing the value of μ_1 the availability is assessed. The same are shown in Table 13.

Table 13. Variation of availability with change in repair rate of ball screw assembly

μ_1	0.01688	0.02	0.04	0.08
Availability	0.9766	0.9796	0.9878	0.992

Similarly, taking λ_1 = 0.0003391, λ_2 = 0.0004544, λ_3 = 0.0004544, μ_1 = 0.01688, μ_3 = 0.00672 as constant and changing the value of μ_2 the availability values are evaluated and the same are given in Table 14.

Table 14. Variation of availability with change in repair rate of sub assembly 1

μ_2	0.00672	0.008	0.02	0.04
Availability	0.9766	0.9772	0.979	0.9797

On the similar lines, taking λ_1 = 0.0003391, λ_2 = 0.0004544, λ_3 = 0.0004544, μ_1 = 0.01688, μ_2 = 0.00672 as constant and changing the value of μ_3 the availability values are obtained. The same are presented in Table 15.

Table 15. Variation of availability with change in repair rate of sub assembly 2

μ_3	0.00672	0.008	0.02	0.04
Availability	0.9766	0.9772	0.979	0.9797

5. Discussion

The availability evaluation of a Vertical Milling Centre is done using Markov model considering constant failure and repair rates during period of operation and further Monte Carlo simulation is used for the same system to compare the results. Standby redundancy is used to further increase the availability of the system. It was found out that out of 10,000 hours of operation, system in passive (standby) redundancy would be available for 9,766 hours. While in active (load sharing) redundancy it will be available for 9,751 hours.

Based on above results it can be easily pointed out that using passive redundancy over active redundancy will result in an increase of 15 hours of available life, or an increase of 0.15% of available life over active redundancy system. Further, Monte Carlo simulation method is used for the same system under passive redundancy to compare the results obtained from Markovian approach. Next, sensitivity Analysis is performed under active redundancy and it can be seen an increase in repair rate of the Ball screw and Sub-assemblies in VMC, availability increases.
Table 10 shows that by increasing the repair rate of Ball Screw from 0.01688 to 0.08, availability increases by 1.5691%. Table 11 shows that by increasing the repair rate of Sub-Assembly-1 from 0.00672 to 0.04, availability increases by 0.4409%. Table 12 shows that by increasing the repair rate of Sub Assembly-2 from 0.00672 to 0.04, availability increases by 0.4409%. Again, sensitivity analysis is performed under passive redundancy and it can be seen an increase in repair rate of the Ball screw and Sub-assemblies in VMC, availability increases. Table 13 shows that by increasing the repair rate of Ball Screw from 0.01688 to 0.08, availability increases by 1.577%. Table 14 shows that by increasing the repair rate of Sub-Assembly-1 from 0.00672 to 0.04, availability increases by 0.31%. Table 15 shows that by increasing the repair rate of Sub-Assembly-2 from 0.00672 to 0.04, availability increases by 0.31%.
Hence, it can be easily pointed out, by increasing repair rate of ball screw sub assembly, availability gain for overall system is higher and hence recommended.

6. Limitations

Assumptions taken during availability analysis restrict the result to specified condition wherein results are found to be valid and in conformance to actual observed. These limitations need to be relaxed to form a much universally acceptable result. Some of these limitations as present in the proposed models are given below:

- in this work, Markovian approach is limited to constant repair and failure rates which needs to be relaxed,
- the model assumes that their cannot be any failure in standby components, but in real scenario the probability of standby components to fail is > 0%,

- constants and factors should be multiplied to availability obtained as in present context smooth operation of machine without any kind of shock or dynamic loading is considered for evaluation of failure rates (which in real world would be greater than as obtained),
- further the initial wear in period effects are neglected, which can substantially decrease availability.

7. Conclusion

In order to make the approach more effective and applicable to the real system, the following aspects are proposed under the future scope.

- Costs are not considered in this study for the availability analysis. It will be relevant if it is connected to the cost of the various maintenance activities and an analysis is performed in terms of availability benefits vs. cost of replacement.
- Due to the limitations of the Markovian method, an exponential distribution is considered for failure and repair time. For a more practical analysis, this premise can be dropped and suitable non-exponential distributions such as Weibull for failure time and Log-normal for restoration time may be considered.
- Integration of availability evaluation with decision-making algorithms that take various variables into consideration.
- Extension of use of developed MCS algorithms and codes for systems with non-exponential (time dependent) failure/repair time distributions just by replacing the TTF and TTR calculation in terms of non-exponentials distribution parameters.
- Monte Carlo simulation approach can be used for availability evaluation and validation of MA for the system in case of active redundancy where changes in failure rates due to load sharing occurs.

Acknowledgment

Authors would like to offer their sincere acknowledgment to authors of article “Availability evaluation of manufacturing systems using semi-Markov model”, i.e. Loganathan et al. (2016).

References

Ahmed, S.B., Alam, M. & Gupta, D. 1989. Performance modelling and evaluation of flexible manufacturing systems using a semi-Markov approach. *International Journal of Computer Integrated Manufacturing* 2(5), 275–280.

Alexander, D.C. 2003. *Application of Monte Carlo Simulation to System Reliability Analysis*. Texas A&M University, 91–94.

Attar, A., Raissi, S. & Khalili-Damghani, K. 2017. A simulation based optimization approach for free distributed repairable multi state availability redundancy allocation problems. *Journal of Reliability Engineering and System Safety* 157, 177–191.

Billinton, R. & Allan, R. N. 2007. *Reliability evaluation of engineering systems: concepts and techniques*. 2nd Edition, Springer, New Delhi.

Borgonovo, E., Marseguerra, M. & Zio E. 2000. A Monte Carlo methodological approach to plant availability modelling with maintenance ageing and obsolescence, *Journal of Reliability Engineering and System Safety* 67(1), 61–73.

Cadini, F., Agliardi, G.I. & Zio, E. 2017. A modelling and simulation framework for reliability/availability assessment of a power transmission grid subject to cascading failure under extreme weather conditions. *Journal of Applied Energy* 185(1), 267–279.

Çekyay, B. & Özekici, S. 2015. Reliability, MTTF and steady-state availability analysis of systems with exponential lifetimes. *Applied Mathematical Modelling* 39(1), 284–296.

Chawla, R. & Kumar, G. 2013. Condition bases maintenance modelling for availability analysis of a repairable mechanical system. *International Journal of Innovations in Engineering & Technology* 2(2), 371–379.

Garg, S., Singh, J. & Singh, D.V. 2010. Availability analysis of crank-case manufacturing in a two-wheeler automobile industry. *Applied Mathematical Modelling*, 34(6), 1672–1683.

Ge, S., Xu, L., Liu, H. & Zhao, M. 2014. Reliability assessment of active distribution system using Monte Carlo simulation method. *Journal of Applied Mathematics* 2014(1), 421347.

Kumar, G., Jain, V. & Gandhi, O.P. 2013. Availability analysis of a repairable mechanical system using analytical semi-Markov approach. *Quality Engineering* 25(2), 97–107.

Kumar, G., Jain, V. & Soni, U. 2019. Modeling and simulation of repairable mechanical systems reliability and availability. *International Journal of System Assurance Engineering and Management* 10, 1221–1233.

Lin, J.Y. & Donaghey, C.E. 1993. A Monte Carlo simulation to determine minimal cut sets and system reliability. *Proceedings of Annual Reliability and Maintainability Symposium*, IEEE, Atlanta, GA, USA, 246–249.

Loganathan, M.K., Kumar, G. & Gandhi, O.P. 2016. Availability evaluation of manufacturing systems using semi-Markov model. *International Journal of Computer Integrated Manufacturing* 29(7), 720–735.

Maciejewski, H. & Caban, D. 2008. Estimation of Repairable System Availability Within Fixed Time Horizon, *Journal of Reliability Engineering & System Safety* 93(1), 100–106.

Okafor, C.E., Atikpakpa, A. & Okonkwo, U.C. 2016. Availability assessment of steam and gas turbine unit of a thermal power station using Markovian approach. *Archives of Current Research International* 6(4), 1–17.

Sharma, R.K. & Kumar, S. 2008 Performance modelling in critical engineering systems using ram analysis. *Reliability Engineering & System Safety* 93(6), 913–919.

Sharma, S.P. & Vishwakarma Y. 2014. Application of Markov process in performance analysis of feeding system of sugar industry. *Journal of Industrial Mathematics* 2014(5), 1–9.

Varghese, J.P. & Kumar, G. 2014. Availability analysis with opportunistic maintenance of a two component deteriorating system. *International Journal of Materials, Mechanics and Manufacturing Manufacturing* 2(2), 155–160.

Cracking the Figurative Code: A Survey of Metaphor Detection Techniques

Vrinda Kohli[1], Himanshu Nandanwar[2], Rahul Katarya[2]
Manipal University Jaipur, India[1]
Delhi Technological University, India[2]
Corresponding author: Himanshu Nandanwar, Email: himanshunandanwar9cm0@gmail.com

Metaphor Detection is a crucial area of study in computational linguistics and natural language processing, as it enables the understanding and communication of abstract ideas through the use of concrete imagery. This survey paper aims to provide an overview of the current state-of-the-art approaches that tackle this issue and analyze trends in the domain across the years. The survey recapitulates the existing methodologies for metaphor detection, highlighting their key contributions and limitations. The methods are assigned three broad categories: feature-engineering-based, traditional deep learning-based, and transformer-based approaches. An analysis of the strengths and weaknesses of each category is showcased. Furthermore, the paper explores the annotated corpora that have been developed to facilitate the development and evaluation of metaphor detection models. By providing a comprehensive overview of the work already done and the research gaps present in pre-existing literature, this survey paper hopes to help future research endeavors, and thus contribute to the advancement of metaphor detection methodologies.

Keywords: Metaphor Detection, Natural Language Processing, Linguistic Analysis, Computational Linguistics, Lexical Semantics

2023. In Satyasai Jagannath Nanda & Rajendra Prasad Yadav (eds.), *Data Science and Intelligent Computing Techniques*, 347–359. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.56155/978-81-955020-2-8-31

1 Introduction

Roughly 12% of the words used in a natural language document are used metaphorically [1]. Metaphors are linguistic tools that present comparisons between two seemingly unrelated ideas through shared traits. They act as a means to describe abstract concepts through vivid imagery. A metaphor is defined by a stark difference in its literal and contextual meanings (Fig 1). For example, in the phrase "I am a forest fire" [2], the speaker does not mean that she is an actual forest fire, but instead uses the phrase to convey the raging intensity of her emotions, displaying a vast disparity between the literal and contextual sense of the expression "forest fire".

Automated Metaphor Detection involves identifying a metaphorical word (or token) in a given text sequence by a machine learning model. This demands a deeper understanding of the often subtle, figurative language used which requires computational models to go beyond surface-level interpretations and delve into the underlying semantic layers of the sentence to capture relevant contextual information. Consequently, the detection of metaphors warrants sophisticated approaches that can encompass the intricacies in the interplay between language, context, and figurative expressions to achieve reliable and insightful results. This task also shows importance in other natural language processing tasks such as machine translation [3], sentiment analysis or opinion mining [4], dialogue systems [5], and machine reading comprehension [6].

The pre-existing techniques for metaphor detection can be broadly classified into three categories. Feature-based methodologies deal with extracting metaphor-specific features from the corpus to identify the need. Traditional Deep Learning-based approaches employ various RNN and hybrid architectures to model the sequential nature of sentences. Lastly, transformer-based approaches use attention-equipped encoder-decoder-style pre-trained architectures (BERT, RoBERTa, etc.) to capture semantic and syntactic relationships from the input text.

Thenceforth, the study of metaphor detection holds considerable implications for understanding language, cognition, and communication. By examining the existing literature, this survey paper attempts to shed light on research gaps. This paves the way for further advancements in the field for developing robust and context-aware models that show generalization across different languages, cultures, and domains. Through this paper, we hope to provide a comprehensive resource for researchers interested in the field of automated metaphor detection.

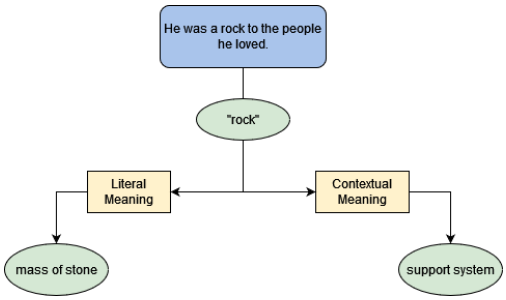


Figure 1. Metaphors have different literal and contextual meanings

2 Literature Review

2.1 Previous Studies

The techniques employed for metaphor detection (MD) have witnessed various trends over the years. In the earlier years of research about this problem, a lot of focus was given to hand-crafted metaphor-centric features. [7] used word concreteness and abstractness as defining features, while [8] used feature norms. Imageability [9], bag-of-words features [10] and sparse distributional features [11] have also been used as linguistic features for machine learning models.

Next came techniques utilizing Neural architectures, such as BiLSTM [12], CNN-hybrids [13], and Graph Neural Networks [14] [15]. These methods popularized the use of word embeddings such as GloVe [16] and Elmo [17] vectors for metaphor detection. [18] further integrates linguistic theory conventions Metaphor Identification Procedure (MIP) [19] and Selectional Preference Violation (SPV) [20] by modeling them as neural architectures.

Transformer-based approaches typically model linguistic rules and other contextual information by using BERT or RoBERTa encoder modules, using those in conjunction with techniques such as context denoising [21], self-supervised learning [22], reading comprehension [23] and parse-tree alterations [14].

A detailed survey covering the specifications of all three approaches can be found in Table 1, and Table 2 demonstrates the quantifiable results obtained by these models. The survey involved studies done over the past six years and presented findings from ten prominent papers in the field.

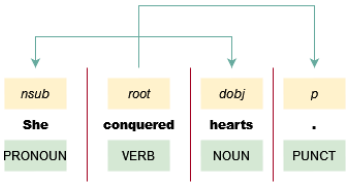


Figure 2. Metaphors with verb-noun direct object relation

2.2 Publicly Available Datasets

There are primarily three datasets on which experimentation about MD tasks is performed.

VUA: The VU Amsterdam Metaphor Corpus (VUA) [24] dataset is the largest publicly available dataset annotated for metaphor detection tasks. It is sampled from the British National Corpus across four genres (Academic, News, Conversation, and Fiction), and consists of 117 fragments. It has over 2000 unique verbs, and the metaphors are distributed with a natural likelihood (~10%).

MOH-X: MOH-X [25] is a verb metaphor detection dataset that has data points sampled from WordNet [26] example sentences. Each sentence has only a singular metaphor tagged in it. The average sentence length is 8 tokens and 48.69% of the words are metaphorical.

TroFi: TroFi [27] is a single target verb metaphor detection dataset that is comprised of sentences from 1987-1989 Wall Street Journal Corpus Release-1. The average length for this dataset is 28.3 tokens per sentence, which is the longest among the three datasets explored. The percentage distribution of metaphors in the dataset amounts to 43.54%.

Table 1. Existing Methodologies

Model	Year & Ref	Category	Contribution	Methodology	Limitations	Advantages
BiLSTM	2018 [12]	Traditional DL approach	Utilization of BiLSTM models with ELMo embeddings for MD.	Tokens concatenated with their ELMo embeddings are encoded using a BiLSTM module. The detection task is modeled in two ways: the classification task is done by using a feedforward neural network, and the sequence labeling task applies an attention layer for computing attention weight per token for weighted classification.	BiLSTM encoder struggles to capture metaphors with long-range dependencies, indirect metaphors, and personification -related metaphors.	Infers that predicting metaphor labels of context words helps predict the target word and that contextualized word vectors improve model performance
Disc	2019 [28]	Feature Engineering approach	Usage of broader discourse-based features to train gradient boosting classifiers for MD task	The GloVe embeddings, doc2vec vectors, skip-thought vectors, and ELMo embeddings are obtained and their concatenation is used as a feature vector for input to a gradient boosting algorithm (XGBoost)	Conversation-based metaphors are harder to detect and this approach has an a-priori need for broader context beyond sentence level.	Competitive results without neural architectures or manually-engineered metaphor-specific features. The usage of paragraph-level context vastly improves detection performance.
DeepMet	2020 [23]	Transformer based approach	Reading comprehension paradigm for	MD is considered to be a reading	Faces difficulties in detecting	Demonstrates that FGPOS features provide

			MD at a token level.	comprehension task, based on context and query words. It involves inputting global and local text contexts, query features, POS features, and FGPOS features into a Siamese architecture with two separate BERT encoders for local and global features. The encoders share weights and an average pooled vector is used as input to the metaphor discrimination module. Cross-validation introduces a metaphor preference parameter.	metaphors triggered by multiple words since the queries are answered one word at a time. Downsampling via average pooling may lead to the loss of relevant information.	more information than standard POS features. The metaphor preference parameter models real-world scenarios in dealing with imbalanced datasets.
WSD-GCN	2020 [14]	Traditional DL approach	Leverages Graph Convolution Networks (GCN) with dependency parse trees and a multi-task framework for exploiting the similarity of MD and word sense	A BiLSTM is used to obtain a feature vector from GLoVe, ELMo, and index embeddings of the sentence, which is then inputted into a GCN module. The GCN and BiLSTM vectors are	The usage of dependency parse trees imposes a reliance on the dataset structure for successful generalization of the approach. A lack of cross-dataset evaluation	The GCN approach successfully identifies relevant context words based on their importance. The multi-task approach handles the issue of knowledge transfer between

			disambiguation (WSD) tasks.	aggregated via calculated control vectors that filter out irrelevant information. A dense network with a Softmax layer is used for MD. Owing to the multitask approach. Two encoders are trained alternatively and simultaneously for WSD and MD to share knowledge between the two tasks.	leaves the question of generalizability unanswered. This technique is hard to apply to batch optimization due to the complicated tree-related structure.	two tasks when the dataset is only annotated for one of the two.
MWE-GCN	2020 [15]	Traditional DL-based approach	Introduces a multiword expression aware model for metaphor identification	The Dependency parse tree information is treated as an undirected graph. The adjacency matrix of this graph is linearly combined with attention-based matrices, providing fully connected weighted graph matrices to determine relation strength between nodes. These matrices are inputted to different	No comparison with the standard VUA dataset, which is considerably vast in its information and generalization strength is not evaluated. The complex tree-related structure makes this approach less amenable to batch optimization.	Demonstrates that the knowledge of Multiword Expressions can significantly boost the performance of MD methods

				Graph Convolution Networks, the outputs from which are linearly combined. The same process is followed for token-level relations between multiword expression components present in the sentence. The GCN outputs of both architectures are concatenated and passed through another GCN to obtain results.		
MelBERT	2021 [1]	Transformer based approach	Uses contextualized word representations and linguistic theories, namely Metaphor Identification Protocol (MIP) and Selectional Preference Violation (SPV) for MD	SPV and MIP are modelled using two RoBERTa backboned encoders and a combined prediction score is obtained post late-stage interaction.	Borderline or implicit metaphors are much harder to identify. The syntactic structure isn't utilized as context words across subsentences lose their relation.	Since late interactions are utilized between the two lingual rules, the sentence vectors can be reused, leading to an amortized cost of encoding. A good level of generalization is achieved across datasets as exhibited in Zero Shot experimentation.
CATE	2021 [22]	Transformer based approach	Introduces a semi-supervised self-training strategy for collecting large-scale	A BERT model is finetuned using pre-existing labeled data. A Target-based Generating	When the available training data size is high, the net gain from self-training drops.	Significant improvement when small-scale datasets are used due to self-supervised data

			candidate instances from generated unlabeled corpus, and a contrastive objective for capturing MIP is defined.	Strategy is used to create a large-scale, relevant unlabeled corpus. The fine-tuned model pseudo-labels this corpus, and this data is then used to augment the training data. The fine-tuned model is updated iteratively using a self-training strategy.	Model accuracy drops when words from multiword expressions are utilized in their literal sense.	augmentation. Self-training leads to a more diverse dataset, bringing about better MD in underrepresented genres. The contrastive objective quantifies the contrast between literal and contextual meanings, upholding MIP without a bulky architecture.
CIA*	2022 [29]	Feature Engineering Approach	Lightweight algorithm for Direct Object related metaphors (Fig 2) specific to the cybersecurity domain	Bing API is queried for the top 50 websites related to a selected verb, relevant sentences are extracted and added to the corpus which is then parsed to obtain collocated nouns. The synsets and hyponyms for these nouns are obtained via WordNet. If the main synset is not present in the collocated noun list, the word is predicted to be a metaphor.	Only a particular style of metaphor is evaluated, constricting the extent of evaluation.	Comparable results without bulky deep learning architecture. The development of a real-world corpus is simple enough to be extended for usage across multiple domain-specific tasks. This approach can identify multiple metaphorical instances present in a sentence successfully.
Frame-BERT	2023 [31]	Transformer based approach	Explainable and interpretable	Two RoBERTa encoders are used: the	Features such as Frame Elements,	Usage of FrameNet embeddings

			metaphor detection by incorporating FrameNet embeddings.	conceptual encoder processes the FrameNet embeddings and the sentence encoder models MIP and SPV. The outputs from both encoders are concatenated to obtain input for the classification module.	Lexical Units, and context graphs need to be explored.	brings up performance by 1.2% owing to their ability to capture deep-level semantics.
RoPPT	2023 [21]	Transformer based approach	A target-oriented parse tree structure is utilized for MD by extracting semantically relevant neighbors of a target word.	The original parse tree is reshaped by rooting the tree at the target word. Context Denoising is performed by pruning the tree based on the distance between the root and leaves. Two RoBERTa-based encoders are used for encoding, one for the target word, and the other for the input sentence, followed by a classification module.	The usage of average pooling may lead to a loss of fine-grained details. Performance is lower than expected for shorter sentences.	The modified tree structure allows the model to focus on only relevant information about the target word. Irrelevant parts are ignored despite their position in the input sentence. Demonstrates the robustness of context denoising mechanism over long sentences.

Table 2. Results on Various Metrics

Ref	Model	VUA				TroFi				MOH-X			
		P	R	F1	Acc	P	R	F1	Acc	P	R	F1	Acc
[1]	MelBERT	80.1	76.9	78.5	-	53.4	74.1	62.0	-	79.3	79.7	79.2	-
[12]	BiLSTM	68.2	71.3	69.7	81.4	70.7	71.6	71.1	74.6	79.4	73.5	75.6	77.2
[14]	WSD-GCN	74.8	75.5	75.1	93.8	73.1	73.6	73.2	76.4	79.7	80.5	79.6	79.9
[15]	MWE-GCN	-	-	-	-	73.78	71.81	72.78	73.45	79.98	80.40	80.19	80.47
[21]	RoPPT	80.0	78.2	79.1	-	54.2	76.2	63.3	-	77.0	83.5	80.1	-
[22]	CATE	79.3	78.8	79.0	94.8	74.4	74.8	74.5	77.7	85.7	84.6	84.7	85.2
[23]	DeepMet	75.6	78.3	76.9	91.6	72.1	80.6	76.1	77.0	93.3	90.3	91.8	92.3
[28]	Disc	58.9	77.1	66.8	-	-	-	-	-	-	-	-	-
[29]	CIA*	-	-	-	-	72	66	68	69	-	-	-	-
[30]	Frame-BERT	82.7	75.3	78.8	-	70.7	78.2	74.2	-	83.2	84.2	83.8	-

3 Research Gap

After a thorough analysis of existing works, as shown in Table 1, we have identified the challenges and limitations of prior approaches as follows:

3.1 Low Generalizability

On average, the proposed approaches rarely discuss the generalizability across datasets, barring a few exceptions [1] [31]. Probing-based studies done in [32] demonstrate that large gaps exist between the in-distribution and out-of-distribution performances of transformer-based methods for MD tasks, presumably due to annotation bias present across the datasets. This implies that the generalizability across datasets of such approaches is lower than expected.

3.2 Heavy Dependency on Dataset

Upon analyzing trends across various methods, one common denoting factor is that these techniques are highly dataset-specific, which poses a challenge for generalization on real-world data which is usually much more diverse in its linguistic styles, cultural references, and domain-specific terminologies. There is a need to develop methods that do not depend this heavily on their training corpus

3.3 LLM-centric approaches

[14] shows competitive results in MD tasks by leveraging its similarity to Word Sense Disambiguation (WSD) [33]. It is shown in [34] the successful usage of LLMs for solving the WSD task. Thus, cross-domain knowledge can be utilized to apply similar techniques for LLM-centric approaches for MD.

4 Discussion

There are primarily three categories of methodologies discussed in this survey, each having its inherent drawbacks and benefits. Even though all methods show a certain level of sensitivity towards the corpus quality, these effects are vastly pronounced in Feature Engineering methods. These methods are only as good as the hand-crafted features utilized by them and the process of extracting corpus-specific

features implies a lack of generalization capability across unseen data. Thus, rarely used metaphors are difficult to identify [1].

Traditional deep learning-based approaches often lack interpretability. Due to the shallow nature of the neural architectures used, the entire extent of context information across different hierarchical levels is not obtained [23].

Transformer-based methodologies were proposed to primarily tackle the limitations induced by the shallowness of these methods. Due to their superior ability to encode metaphorical knowledge [32] these show state-of-the-art performance on MD tasks (Table 2). Out of all surveyed methods, CATE [22] seems to give the most accurate predictions on VUA and TroFi, owing to its semi-supervised self-training mechanism. On the other hand, DeepMet [23] outperforms all other approaches on MOH-X.

5 Conclusion

Summing up, several approaches to broaching automated detection of metaphors in natural language corpora were discussed in this paper. We have discussed the linguistic aspects of metaphor and how they get modeled as computational tasks. Understanding and recognizing metaphors rigorously through computational techniques is bound to bring significant progress in the aligned natural language processing tasks and provide insight into human cognition.

As the field continues to advance, researchers should focus on developing robust and context-aware models that tackle the prevalent issues with prior techniques, integrating up-and-coming innovations within them. A possible course of action for the authors would be to explore and apply themselves to the research gaps and look into LLM-based methodologies for metaphor detection.

In conclusion, by providing a thorough understanding of the current landscape, challenges, and limitations of the current methods for metaphor detection, this paper hopes to facilitate future research endeavors and foster collaborative efforts for the development of advanced metaphor detection techniques.

References

[1] Choi, Minjin, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. "MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories." *arXiv preprint arXiv:2104.13615* (2021).

[2] Mitski: A Burning Hill [Song] on *Puberty 2*. Dead Oceans.

[3] Shi, Chunqi, Toru Ishida, and Donghui Lin. "Translation agent: a new metaphor for machine translation." *New Generation Computing* 32 (2014): 163-186.

[4] Cambria, Erik, Soujanya Poria, Alexander Gelbukh, and Mike Thelwall. "Sentiment analysis is a big suitcase." *IEEE Intelligent Systems* 32, no. 6 (2017): 74-80.

[5] Dybala, Pawel, and Kohichi Sayama. "Humor, emotions and communication: Human-like issues of human-computer interactions." In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 34, no. 34. 2012.

[6] Tu, Ming, Guangtao Wang, Jing Huang, Yun Tang, Xiaodong He, and Bowen Zhou. "Multi-hop reading comprehension across multiple documents by reasoning over heterogeneous graphs." *arXiv preprint arXiv:1905.07374* (2019).

[7] Turney, Peter, Yair Neuman, Dan Assaf, and Yohai Cohen. "Literal and metaphorical sense identification through concrete and abstract context." In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 680-690. 2011.

[8] Bulat, Luana, Stephen Clark, and Ekaterina Shutova. "Modelling metaphor with attribute-based semantics." In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 523-528. 2017.

[9] Broadwell, George Aaron, Umit Boz, Ignacio Cases, Tomek Strzalkowski, Laurie Feldman, Sarah Taylor, Samira Shaikh, Ting Liu, Kit Cho, and Nick Webb. "Using imageability and topic chaining to locate metaphors in linguistic corpora." In *Social Computing, Behavioral-Cultural Modeling and Prediction: 6th International Conference, SBP 2013, Washington, DC, USA, April 2-5, 2013. Proceedings 6*, pp. 102-110. Springer Berlin Heidelberg, 2013.

[10] Köper, Maximilian, and Sabine Schulte im Walde. "Distinguishing literal and non-literal usage of German particle verbs." In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pp. 353-362. 2016.

[11] Leong, Chee Wee, Beata Beigman Klebanov, and Ekaterina Shutova. "A report on the 2018 VUA metaphor detection shared task." In *Proceedings of the Workshop on Figurative Language Processing*, pp. 56-66. 2018.

[12] Gao, Ge, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. "Neural metaphor detection in context." *arXiv preprint arXiv:1808.09653* (2018).

[13] Wu, Chuhan, Fangzhao Wu, Yubo Chen, Sixing Wu, Zhigang Yuan, and Yongfeng Huang. "Neural metaphor detecting with CNN-LSTM model." In *Proceedings of the workshop on figurative language processing*, pp. 110-114. 2018.

[14] Le, Duong, My Thai, and Thien Nguyen. "Multi-task learning for metaphor detection with graph convolutional neural networks and word sense disambiguation." In *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 05, pp. 8139-8146. 2020.

[15] Rohanian, Omid, Marek Rei, Shiva Taslimipoor, and Le Ha. "Verbal multiword expressions for identification of metaphor." ACL, 2020.

[16] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532-1543. 2014.

[17] Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. "Deep Contextualized Word Representations." In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2227–2237. New Orleans, Louisiana: Association for Computational Linguistics, 2018.

[18] Mao, Rui, Chenghua Lin, and Frank Guerin. "End-to-end sequential metaphor identification inspired by linguistic theories." In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 3888-3898. 2019.

[19] Group, Pragglejaz. "MIP: A method for identifying metaphorically used words in discourse." *Metaphor and symbol* 22, no. 1 (2007): 1-39.

[20] Wilks, Yorick. "A preferential, pattern-seeking, semantics for natural language inference." *Artificial intelligence* 6, no. 1 (1975): 53-74.

[21] Wang, Shun, Yucheng Li, Chenghua Lin, Loïc Barrault, and Frank Guerin. "Metaphor Detection with Effective Context Denoising." *arXiv preprint arXiv:2302.05611* (2023).

[22] Lin, Zhenxi, Qianli Ma, Jiangyue Yan, and Jieyu Chen. "CATE: A contrastive pre-trained model for metaphor detection with semi-supervised learning." In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3888-3898. 2021.

[23] Su, Chuandong, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiqun Chen. "DeepMet: A reading comprehension paradigm for token-level metaphor detection." In *Proceedings of the second workshop on figurative language processing*, pp. 30-39. 2020.

[24] Steen, Gerard, Aletta G. Dorst, J. Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. "A method for linguistic metaphor identification." *Amsterdam: Benjamins* (2010).

[25] Mohammad, Saif, Ekaterina Shutova, and Peter Turney. "Metaphor as a medium for emotion: An empirical study." In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pp. 23-33. 2016.

[26] Kilgarriff, Adam. "Wordnet: An electronic lexical database." (2000): 706-708.

[27] Birke, Julia, and Anoop Sarkar. "A clustering approach for nearly unsupervised recognition of nonliteral language." In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 329-336. 2006.

[28] Mu, Jesse, Helen Yannakoudakis, and Ekaterina Shutova. "Learning outside the box: Discourse-level features improve metaphor identification." *arXiv preprint arXiv:1904.02246* (2019).

[29] Hilton, Kelsey, Akbar Siami Namin, and Keith S. Jones. "Metaphor identification in cybersecurity texts: a lightweight linguistic approach." *SN Applied Sciences* 4, no. 2 (2022): 60.

[30] Li, Yucheng, Shun Wang, Chenghua Lin, Frank Guerin, and Loïc Barrault. "FrameBERT: Conceptual metaphor detection with frame embedding learning." *arXiv preprint arXiv:2302.04834* (2023).

[31] Zhang, Shenglong, and Ying Liu. "Metaphor detection via linguistics enhanced Siamese network." In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 4149-4159. 2022.

[32] Aghazadeh, Ehsan, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. "Metaphors in pre-trained language models: Probing and generalization across datasets and languages." *arXiv preprint arXiv:2203.14139* (2022).

[33] Agirre, Eneko, and German Rigau. "Word sense disambiguation using conceptual density." *arXiv preprint cmp-lg/9606007* (1996).

[34] Laba, Yurii, Volodymyr Mudryi, Dmytro Chaplynskyi, Mariana Romanyshyn, and Oles Dobosevych. "Contextual embeddings for Ukrainian: A large language model approach to word sense disambiguation." In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pp. 11-19. 2023.

Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-17250-4

DAC-BiNet: Twitter crime detection using deep attention convolutional bi-directional aquila optimal network

Monika[1] · **Aruna Bhat**[1]

Received: 15 July 2022 / Revised: 23 May 2023 / Accepted: 22 September 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract
Recently, the crime rates in social media are rising rapidly and it produces serious problems to the users. From the past few years, twitter platform has been gaining higher attention due to its effective services. But, the growth of crimes in twitter generates several issues and it redirect users to malicious or phishing websites. Thus, in order to identify various illegal activities on twitter, the proposed study introduces a robust Deep Attention Convolutional Bi-directional Aquila Optimal Network (DAC-BiNet) model. For effectively processing the proposed network, the raw inputs are pre-processed initially to reduce the noises through diverse steps like segmentation of sentences, punctuation removal, tokenization, removal of stop words, acronym and slang correction, lemmatization, lower casing, stemming, removal of hashtag and URLs. From the pre-processed data, the significant features are extracted using ITF-IDF (Improved Term Frequency-Improved Document Frequency), Feature hashing and Glove Modelling methods. Then, the feature dimensionality issue is solved by grouping the features into clusters by Possibilistic Fuzzy LDA (Latent Dirichlet Allocation) based clustering method. Finally, classification stage is initiated to identify the crime tweets in the given inputs through proposed DAC-BiNet model. The proposed study used python platform for simulation and the efficiency of proposed model is measured by evaluating various performance matrices and comparing with other existing techniques. The obtained simulation results proves that proposed model gains enhanced accuracy of 98.23%, precision of 83.86%, recall of 90.05%, specificity of 98.86%, and F1 score of 86.84%.

Keywords Twitter crime detection · Pre-processing · Feature extraction · Deep Attention Convolutional Bi-directional Aquila Optimal Network (DAC-BiNet) · Aquila optimization

✉ Aruna Bhat
aruna.bhat@dtu.ac.in

Monika
monika.siwaliya@gmail.com

[1] Department of Computer Science and Engineering, Delhi Technological University, Delhi, India

Published online: 17 October 2023

1 Introduction

Nowadays, the rate of crime and criminals are increasing rapidly, and security has become a major concern. The term 'crime' refers to an offensive action that creates psychological or physical harm to person or property and punished by criminal law [1, 2]. With the advances in technologies, several crimes are active in social media platforms such as Twitter, Facebook, Instagram [3, 4] etc. Social sites are evolving as a large source of sentiment rich data to express people's opinions, views, and emotions in the form of reviews, tweets, blog posts etc. Twitter is one of the most popular social sites which acts as a communication tool all over the world in the form of short texts (tweets) [5].

Twitter is the trendy micro-blogging site having 217 M (million) active users according to the quick statistics (https://www.omnicoreagency.com/twitter-statistics/). Twitter is being used by huge number of users to share their views on diverse social aspects. Till date, twitter crimes are increasing daily due to the lack of predictive criminal factors [6, 7]. Large quantity of fake accounts has been generated by fraudsters in order to imitate the real twitter users with the wrong intent of spreading spam messages. Online crimes have severe impacts which greatly affects the economic growth and quality of life. Detecting the twitter crime is very challenging and can optimize the spread of online criminal events [8, 9]. Therefore, it is necessary to develop an effective crime pattern detection which can be beneficial in identifying the spam tweets.

Data Mining (DM) is widely utilized to extract knowledge from the large volume of data and discovers hidden data patterns using AI (Artificial Intelligence) techniques. Data mining can be used to solve the issues on credit card thefts, crime detection, fake news detection etc. DM techniques can be used for discovering patterns from huge twitter data and used for fetching and finding the criminal online data [10–12]. Some of the poplar DM techniques such as clustering, rule mining, classification is significantly used for observing the illegal data and effectively detects the crime data. Text Mining (TM) is the most important type which uses NLP (Natural Language Processing) techniques to convert the unstructured text data into a structured form [13–15].

Crime detection aims to identify the crimes in social media before causing severe harm to the society. Computer vision (CV) is the sub-field of AI which trains the machine to understand and realize the visual world by analysing various applications [16]. With the technological proliferation, twitter crimes are ubiquitous and automated tool is needed to be designed for the effective classification of crime as well as non-crime tweets. The advanced AI models can automatically classify the crime tweets since the crime information is significantly spreading with the bad motive over Twitter) [17–19]. The crime tweet detection helps to identify the crime posts over online and useful in managing the crime information reported in better manner. AI based crime detection minimizes the time taken to resolve the crime also benefits both police officials and victims in detecting the crime.

Recently, the online crimes are increasing with higher rate and has serious impact with the social-economic activities. Crime events also have direct consequence on country's social constructs, economy and global repute. More effective security measures are required to detect and minimize the online frauds and crime activities. With the recent technological advances, the traditional approaches need to be replaced with different Deep learning based models to predict the twitter crime events. The automated detection model helps to analyse the crime against different applications such as election results, crime on women, box-office results prediction, crime against children, cybercrimes, cyberbullying, etc. Thus, the DL based effective crime detection can benefits the law enforcers and

legislatures to build quick and informed models that can eliminate the online crime and facilitate development of nation. The major contribution of this proposed twitter crime detection model is characterized as,

- To provide robust twitter crime detection, an effective Deep Attention Convolutional Bi-directional Aquila Optimal Network (DAC-BiNet) model is introduced.
- To minimize the noise in the online tweet data using pre-processing techniques followed with multiple feature extraction procedures.
- To minimize the computational complexity using Possibilistic Fuzzy LDA which minimizes the original feature set and forms a cluster based model.
- To evaluate the detection accuracy by comparing the proposed DL model with other conventional deep learning classifiers.

The rest of this study is organised as follows, Section 2 deals with the literature survey related on existing crime detection studies, Section 3 presents the proposed methodology, Section 4 describes the results and discussions of proposed DAC-BiNet model and Section 5 represents the overall conclusion of this paper and also it points out the future direction for further research studies.

2 Related work

2.1 Some of the recent existing research studies carried out on crime detection using different techniques are described below,

Lal et al. [20] introduced the classification and analysis of online crime tweets using text-mining approaches. This analysis focus on the crime contents and messages posted on Twitter. The tweet information focusses on the following hashtags such as #rape #crime #violence #murder #fraud #cyber crime #harassment #robbery etc. Initially, the data cleaning was done using different steps such as stop words removal, segmentation of sentence, hashtag expansion, punctuation removal, expanding slang and acronym, and TF-IDF. The data which was collected manually was labelled as crime and non-crime tweets. The experimentation was done with WEKA tool. The classification was conducted and analysed with accuracy on different classifiers such as RF (98.1%), ZeroR (61.5%), NB (97.83%), J48 (97.0%) respectively. The major limitation was high time complexity with RF classifier as 7.24 sec.

Alghamdi and Murtaza [21] developed an intelligent model for the detection of suspicious tweets on Arabic language. The Twitter streaming (TS) API (Application Programming Interface) was used to gather the Arabic tweet information. Tweet labelling was done manually as not-suspicious (label 0) and suspicious (label 1). The filtering process was conducted to enhance the quality of the proposed model. Different supervised classifiers were modelled for the classification of tweets. The classifiers KNN, SVM, LDA and DT were processed with BoW (Bag-of-Words) based feature extraction whereas the LSTM and ANN classifiers use word embedding procedure. The accuracy was higher with SVM (86.72%) classifier. The drawbacks were limited number of data, and the binary tweet categorization.

Santhiya et al. [22] developed the automatic categorization of crime tweets to analyse the crime events against the children and women. The classification framework has been

analysed based on the perspective of big data using dual models such as Apache Spark and Hadoop MapReduce. The diverse types of crime events can be classified based on the hierarchical lexical domain in distributed and parallel manner. Pre-processing was performed for filtering the tweet data. Next, computation of similarity was presented with document and entity based similarity procedures. The classification of tweets was processed using MapReduce Naïve Bayes (MNB) classifier and the final presentation layer was utilized for the geo-location prediction and spatial visualization. The accuracy obtained with the training set was (0.1153) and with the testing set was (0.1372). The major limitation was lower classification accuracy.

Boukabous and Mostafa [23] presented the prediction of crime tweet using the BERT model. The labelling of the crime dataset was done using lexicon-method as crime-related and normal contents. The labelled dataset contents were used to train the BERT model. The tagging of words was completed with different word based on dictionary categories as positive, incremental, negative, inverse words and decremental. For each sentence, the sentiment score was evaluated, and the positive score illustrates the normal tweet whereas the negative sore indicates the crime tweet. The prediction and analysis of sentiments were processed with the integration of lexicon (BERT) and DL classifier model. The hybrid classification obtained F1-score (94.92%) and accuracy (94.91%). The drawback was high cost with the hardware specs.

Hissah and Hmood [24] presented the classification and detection of crimes with the techniques of text mining focusing on twitter Arabic posts. The data was collected from the twitter platform and manual labelling was done. The tweet pre-processing was represented with different approaches such as tokenization, filtering, normalization and stemming. The extraction of Arabic tweets was performed with TF-IDF based approach. The classification of tweets has been analysed on four different classifiers such as CNB (88.17%), SVM (91.55%), DT (82.46%) and KNN (78.06%). The highest classification was evaluated with SVM whereas KNN with the lowest value. The classifier CNB (Complement NB) was the fastest classifier with speed and processing time. The limitation was analysis has not considered the temporal and spatial information. Table 1 shows the comparison of state-of-the-art methods.

2.2 Problem statement

The crime detection technology provides high security to the user's data and also it avoid the presented rumors in social media platform. In recent years, twitter is becoming more famous because of the good microblogging service. The twitter data also carries several crimes which makes serious impacts for the society. In existing, several techniques are utilized for crime detection in twitter data. However, the traditional methods are not significant due to the manual recognition of crime tweets. While detecting crime data, the existing methods faced higher computation complexity problem because of an incapability of handling larger inputs. In general, the social media dataset contains large number of user reviews, which contains both real user data and crime data. Thus, analysing crimes in such datasets are becoming a challenging issue. Some of the other challenges recognized with the detection of crime tweets are number of massive tweets, presence of noisy tweets, and large twitter handles. Thus, to solve the mentioned issues in the existing methods, the proposed study develop a new crime detection model in twitter environment.

Table 1 State-of-the-art methods for crime data detection and its limitations

Author	Techniques used	Dataset used	Performance achieved	Limitations
Lal et al. [20]	Naïve Bayes, J48, Random forest, ZeroR	Manually gathered from Twitter interface	Accuracy of Naïve bayes is 97.83%, random forest is 98.1%, J48 is 97% and ZeroR is 61.5%	• Obtained higher computational complexity • Failed to handle more number of tweets • Higher time complexity
Alghamdi and Murtaza [21]	KNN, SVM, LDA and DT	Tweet dataset	86.72% of accuracy obtained by SVM	• Processed only in low number of tweets • Failed to work on multi-class classification
Santhiya et al. [22]	Multinomial Naïve Bayes classifier	Tweets collected from the year of Jan 2016 to Dec 2017 in Twitter Search API	Training accuracy as 0.1153 and testing accuracy as 0.1372	• Classification accuracy is gets affected
Boukabous and Mostafa [23]	BERT model	Collected tweets data from twitter platform	Achieved accuracy as 94.91% and F1-score as 94.92%	• Requires high cost with hardware specs
Hissah and Hmood [24]	SVM, KNN, DT and CNB	Data collected from Twitter API and Topsy	Accuracy obtained by SVM is 91.55%, KNN is 78.06%, DT is 82.46% and CNB is 88.17%	• The developed technique cannot fetch spatial and temporal attributes

3 Proposed methodology

A Deep Attention Convolutional Bi-directional Aquila Optimal Network (DAC-BiNet) based crime tweet detection is presented in this proposed work to ensure more secure and safest online services. The steps used for the detection of crime tweets are Data Acquisition, Pre-processing, Multi-Feature Extraction, Clustering and Crime tweet detection. Initially, the tweet data is collected from the online source. The pre-processing step is needed to be performed before the detection process as the online tweet data is in unstructured form. The steps used in the pre-processing are Segmentation of sentences, punctuation removal, tokenization, removal of stop words, acronym and slang correction, lemmatization, lower casing, stemming, removal of hashtag and URLs. Feature extraction aims to transform the arbitrary form of data (images or text) into numerical form. Here, the tweet features are extracted using multi-feature approaches such as ITF-IDF (Improved TF-IDF), Feature hashing and Glove Modelling. The best feature extraction process is analysed, and the best feature extracted outcome is fed to the clustering model. Next, Clustering of unstructured tweets contents are processed with Possibilistic Fuzzy LDA (Latent Dirichlet Allocation). This clustering approach convert the original set of features into a reduced set which increases the ability of predictive models. Finally, the detection and classification of crime tweets can be performed using Deep Attention Convolutional Bi-directional Aquila Optimal Network (DAC-BiNet). Here, the loss in the deep learning model is minimized using Aquila Optimization (AO) Algorithm. However, this optimization algorithm is used to learn the parameters from the given input data and improve the efficacy of the deep learning classifier model. Thus, this proposed DAC-BiNet model can be benefited significantly with the use of optimisation algorithm for tuning the network parameters and performs accurate classification of twitter crime data. Figure 1 illustrates the work flow of proposed DAC-BiNet model.

3.1 Pre-processing

The pre-processing process is performed initially to remove the noises in the given twitter data and it helps to convert the unstructured data in to structured form. For pre-processing, the proposed work utilized different steps like Segmentation of sentences, punctuation removal, tokenization, removal of stop words, acronym and slang correction, lemmatization, lower casing, stemming, removal of hashtag and URLs.

Segmentation of sentences: It is the process of analysing the longer processing units consisting of one or more words. In this task, the sentence boundaries among words in varied sentences are identified. This step can partition apart the sentences whenever it identifies a punctuation mark. Thus, it is also referred as sentence boundary recognition, sentence boundary disambiguation or sentence boundary detection. An example of segmentation of sentences is given in Table 2.

Punctuation removal: This step is important to separate the given input data in to phrases, paragraphs and sentences. The main intention of this step is to remove commas, full stops, exclamation marks, colons, question marks etc. from the twitter data.

Tokenization: Tokenization is the significant step in NLP, which transfers the input sentences into understandable bits of data that a model can process with. For the purpose of computational linguistics, the words that detected are frequently said to as tokens. By performing tokenization, the input text are splitted into a set of meaningful tokens or pieces.

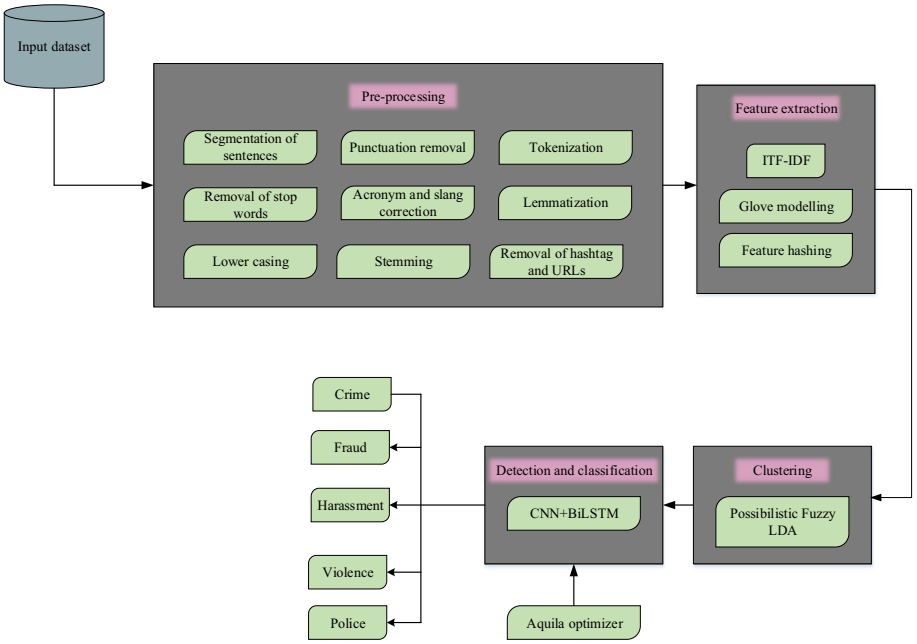


Fig. 1 Structure of proposed DAC-BiNet model

Table 2 Segmentation of sentences: An example

Before sentence segmentation	After sentence segmentation
"I am going to the shop. I like to buy a new pen. I usually use fountain pen"	I am going to the shop I like to buy a new pen I usually use fountain pen

For instance, the group of text are converted into words or sentences. An example of tokenization is given as, input tweet- 'I like pink colour dress', after the process of tokenization, it is converted into 'I', 'like', 'pink', 'color'.

Removal of stop words: In the stop words removal step, the low level information of data are removed and it highly concentrates the data which provide essential information. In general, the articles and pronouns presented in the input data are categorized as stop words. An example of stop words are 'the', 'what', 'is' etc. These terms are not enough to provide higher information to detect crime on twitter data and hence such type of words are removed in this step.

Acronym and slang correction: This stage expands the similar slangs and acronyms presented in the twitter data. An example is described as, 'bf' is changed in to 'beautiful flowers' and the slang such as 2 day was expanded in to today.

Lemmatization: Based on the word's meaning in the twitter data, this lemmatization step finds the word's lemma. This step aids to describe the morphological analysis of the twitter word and it eliminates the inflectional endings.

Lower casing: This step transforms the twitter words in to lower case, for example BOOK in to book.

Stemming: The stemming step removes the word in to its root stem and it eliminates prefix or suffix from a word. This process makes the system to run with higher speed and is the important process in natural language understanding (NLU) and NLP.

Removal of hashtag and URL: In this step, the available hashes in the twitter data is removed and also it eliminates the URL from the input data. For example, #TheProudFamily is converted in to The Proud Family. For URL removal, the given example is https://www.computerhope.com and such type of URLs are removed in the given twitter data.

Depending the above steps, the input twitter data is pre-processed in the proposed work and the noisy data are gets eliminated. Also, this stage generates the twitter based data as structured manner and it make ease the process of further crime detection stages.

3.2 Feature extraction

From the pre-processed data, the essential features are extracted to reduce the computational complexity of proposed model. The essential features exhibit useful information for crime detection. For this purpose, the proposed work uses ITF-IDF, Feature hashing and Glove modelling.

3.2.1 Improved Term Frequency-Improved Document Frequency (ITF-IDF)

The ITF-IDF method is the fundamental text based feature extraction approach and it extracts various necessary features. Because of the lower computations, the ITF-IDF method can easily fetches the rich information from pre-processed data. In existing, different methods were utilized for extracting the text features. However, they requires high computation and thereby enhances the overall processing time. By using ITF-IDF, the importance of each words in a given text is identified effortlessly. Also, the lower computation of ITF-IDF can reduces the processing time and helps to enhance the classification results. Thus, the proposed study prefers ITF-IDF method for feature extraction purpose. Inverse document frequency weight feature is another name of ITF-IDF method. The repeated features in the input twitter data is mentioned as TF and the weight value is computed to diminish the larger text length. Hence, a weight evaluation method is employed to initiate the procedure of adding feature weights. The feature weight of each twitter data is given as,

$$\omega_i = \frac{TF_i \times \log_2\left(\frac{Z}{z_i} + 0.01\right)}{\sqrt{\sum\left(TF_i \times \log_2\left(\frac{Z}{z_i} + 0.01\right)\right)^2}} \tag{1}$$

Where, the feature weight of pre-processed twitter data is signified as ω_i, total volume of training sample is indicated as Z, amount of data in the text feature t_i is considered as z_j and the volume of occurrences which presents in feature item t_i is mentioned as TF_i. If the volume of instances that a word presented in the given tweet is high and the volume of instances that a word available in another tweet is low, then the proposed TF-IDF is robust to categorize the different type of information. The performance of classification is not excellent by using TF-IDF method because of the undetermination of location information of the tweet data. Moreover, the TF-IDF method has no ability to extract semantic information. Thus, the precision and recall performance is reduced. Basically, the keywords are referred as verb, adjectives, phrases or noun. But, conjunctions, quantifier and function words are also said as stop words. Hence, extracting such words are cannot provide better

outcomes because of the unnecessary content. Thus, in this type of words, filtering is introduced to raise the performances of recall and precision. Let us assume, the length of word, word's position and speech part are included in to weight formulation formula for computing the feature items. By proposing an improved TF-IDF, the performance of conventional TF-IDF is enhanced. It is mentioned as,

$$\omega_i = \frac{TF_i \times \log_2\left(\frac{Z}{z_i} + 0.01\right)}{\sqrt{\sum\left(TF_i \times \log_2\left(\frac{Z}{z_i} + 0.01\right)\right)^2}} \times \gamma \times \sum_{i=1}^{n} \lambda \times \xi \tag{2}$$

Where, ω_i represents the tweet data's feature weight, the parameter of feature length is given as γ, the volume of instances that presented in position i is indicated as λ and the parameter of word's speech location is represented as ξ.

3.2.2 Feature hashing

In feature hashing, a hash function is applied to each of the features in the presented data. The feature hashing approach transforms different length of unstructured text data in to similar length of numeric tweets features. The hashing features has the ability to minimize the data dimensionality and it helps to explore a features as high speed by changing string compared hash value. The features extracted from this feature hashing process is increased because all the n-gram is mapped to a tweet based feature.

3.2.3 Glove modelling

The Glove modelling is one of the word embedding model which mentions the words of tweets as a matrix of numeric vectors or values. The word embedding method generates same vector representation for each words which contain same meanings. Here, all the words in twitter data are transformed to single vector and the mapping process is mainly performed in low-dimensional space. Also, it is enabled based on the data size. The Glove modelling technique encodes the word's meaning depending the proportion of co-occurrence possibilities. This Glove modelling powerfully produce word vectors for the tasks of word similarity. In this method, the semantic features are gets extracted.

By using ITF-IDF, feature hashing and Glove modelling approaches, 33,485 amount of features are extracted. From that, the features which are highly suitable for crime detection is analysed and the optimal features are selected for the further process. In the proposed work, the features from ITF-IDF approach is more effective hence such features are subjected to the next stage.

3.3 Possibilistic Fuzzy LDA based clustering for feature reduction

The preceding feature extraction stage generates several number of extracted features which supported for crime detection. However, the high dimension of features may leads misclassification error and thereby influences the performance of proposed classifier. Therefore, it is important to minimize the feature size for making superior classification outcomes. Thus, the large dimensionality of features extracted from the feature extraction techniques is reduced in this clustering method. In order to make reduced feature set, the proposed work used

possibilistic fuzzy LDA approach. Here, LDA is employed for identifying the top features that sufficient for crime detection. Then, the possibilistic fuzzy c-means (PFCM) is used to group the selected features for minimizing the feature size based on the membership grade. The utilized PFCM method develops an objective function for performing clustering process which is discussed in the following descriptions.

The LDA is a basic probabilistic topic method, in which each document is mentioned as a random combination of latent topics. In LDA, each of the latent topic is represented as a distribution over constant group of words and is utilized to detect the design of underlying latent topic depending the given data. In general, the words are created in two phases for all the document. For each word in the extracted feature set, a distribution over topics is randomly selected in the initial phase. In LDA method, a word is an individual data from a vocabulary index and is represented as $\{1, \ldots\ldots, X\}$, $y = \left(y_1, y_2 \ldots\ldots..y_m\right)$ be a sequence of m words and a group of N documents are indicated as $f = \left(y_1, y_2 \ldots\ldots..y_N\right)$. Depending the 3-level Bayesian graphical approach, the LDA method is represented. Here, 3 of the nodes involves random variables and the edges are feasible dependencies among the variables. At corpus generation, the parameters such as ϕ and λ are analysed. For LDA's generative process, a joint distribution over random variable is mentioned and the probability density function of random variable at q-dimension is evaluated as,

$$p(\zeta/\phi) = \frac{\Gamma\left(\sum_{i=1}^{q} \phi_i\right)}{\prod_{i=1}^{q} \Gamma\left(\phi_i\right)} \zeta_1^{\phi_1 - 1} \ldots\ldots\ldots \zeta_q^{\phi_q - 1} \tag{3}$$

The joint distribution of combination of topic is computed as,

$$p(\zeta, a, b|\phi, \lambda) = p(\zeta, |\phi)\Pi_{m=1}^{M}\, p\left(a_m|\zeta\right) p\left(b_m|a_m, a\right) \tag{4}$$

The probability of corpus is formulated as,

$$p(f|\phi, \lambda) = \Pi_{d=1}^{N} \int p\left(\zeta_m|\phi\right) \times \left(\Pi_{m=1}^{m_d} \Sigma_{a_{dm}} p\left(a_{dm}|\zeta_d\right) p\left(b_{dm}|a_{dm}, \lambda\right)\right) d\zeta_d \tag{5}$$

Where, ϕ mentions the parameter of dirichlet, the variables of document level topic is indicated as ζ, per-word topic assignment is considered as a, acquired word is represented as b, topics are signified as λ, total amount of words are mentioned as M and the document is indicated as N. The positive and negative key words are extracted along with separate weight values and such key words are collected in dictionary. The positive and negative weight values are obtained by coordinating the testing tweet data with the dictionary during the phase of testing. After attaining the weight values of positive and negative keywords, the process of clustering is performed to diminish the feature size. The clustering process is done by using PFCM and here each of the tweet data is considered as the part of cluster to a level represented by the membership grade. The objective function of PFCM is given as,

$$G_{PFCM}(p, t, v) = \sum_{i=1}^{h} \sum_{j=1}^{m} \left(p_{ij}^{n} + t^{m}\right) d^2\left(a_j, v_i\right) \tag{6}$$

The constraints are mentioned as,

$$\sum_{i=1}^{h} \lambda_{ij} = 1, \forall j \in \{1, \ldots\ldots, m\} \tag{7}$$

$$\sum_{j=1}^{h} t_{ij} = 1, \forall i \in \{1, \ldots, h\} \tag{8}$$

Where, the objective function is referred as G_{PFCM}, partition matrix is denoted as p, typicality matrix is indicated as t and the vector of cluster centroids is represented as v. For each iteration, the objective function is evaluated in which the membership degree and cluster centre is computed as the following Eq. (9), (10) and (11).

$$\lambda_{ij} = \left[\sum_{q=1}^{h} \left(\frac{da_j, v_i}{da_j, v_q} \right)^{\frac{2}{n-1}} \right]^{-1}, 1 \le i \le h, 1 \le j \le m \tag{9}$$

$$t_{ij} = \left[\sum_{q=1}^{m} \left(\frac{da_j, v_i}{da_j, v_q} \right)^{\frac{2}{m-1}} \right]^{-1}, 1 \le i \le c1 \le j \le n \tag{10}$$

$$v_i = \frac{\sum_{q=1}^{m} \left(u_{iq}^{n} + t_{iq}^{m} \right) a_q}{\sum_{q=1}^{m} \left(u_{iq}^{n} + t_{iq}^{m} \right)}, 1 \le i \le h \tag{11}$$

Where, m mentions the amount of data points, the amount of cluster centroid is mentioned as h and the coordinates are denoted as (a_j, v_i), which is utilized to compute the interval among data points and cluster center. For each cluster, PFCM generates memberships and possibilities with usual cluster centres and prototypes. The performance of clustering is depends upon the objective function utilized for clustering. The powerful objective function is developed by considering the two requirements. At first, the interval among the cluster should be minimized. Secondly, the interval among the data points that assigned in the clusters should be diminished. The objective function provided in the PFCM is enhanced by driven prototype learning and is mentioned as the parameter η. This parameter is gets updated for each of the iteration and η is formulated as,

$$\eta = \exp\left(-\min_{i \ne q} \frac{\| v_i - v_q \|}{a} \right) \tag{12}$$

Where, the sample variance is considered as a and is expressed as,

$$a = \frac{\sum_{j=1}^{m} \| a_j - \overline{a} \|^2}{m} \tag{13}$$

$$\text{Where}, \overline{a} = \frac{\sum_{j=m}^{m} a_j}{m} \tag{14}$$

The similar value of η is represented by applying a weight parameter. This weight function affords improved classification results and is evaluated as,

$$\omega_{ji} = \exp\left(-\frac{\|a_j - v_i\|^2}{\left(\sum_{j=1}^{m}\|a_j - \overline{v}\|^2\right) \times h/m}\right) \tag{15}$$

Where, the weight function of j^{th} point with i^{th} class is signified as ω_{ji}. Based on this process, the clustering is done in which the feature sets are gets reduced and it helps to generate the classification results as high. Here, the features are clustered into five cluster groups and each cluster groups are different in size. Based on the rules of PFCM, the clustering is formed and reduces the feature dimensionality problem.

3.4 DAC-BiNet based twitter crime detection and classification

In this final step, the crime data that presented in the given inputs are detected and classified by the proposed classifier. The clustered feature sets from the preceding step are fed to the input of proposed DAC-BiNet classification model. In this, the proposed work hybridizes CNN and Bi-LSTM approaches to attain higher detection results. Initially, the CNN method receives the input of clustered feature set and the layers presented in CNN extracts the most important features and transmit it to the Bi-LSTM method for crime detection. The proposed DAC-BiNet model consists of six layers such as input layer, convolutional layer, pooling layer, Bi-LSTM layer, attention layer and output layer. Figure 2 illustrates the architecture of proposed DAC-BiNet. Table 3 mentions the hyper parameter settings of proposed work.

The feature set obtained from the clustering method is subjected to the input of input layer. This input layer send the feature set to the convolutional layer. The convolutional layer extracts the significant features but its dimension is too high which may affect the detection performance. Thus, the feature dimension is minimized by applying a max-pooling layer. This max-pooling layer also reduces the over-fitting issue and it allows to generate optimal results.

$$O_c = x\left(a_c * w_c + b_c\right) \tag{16}$$

Where, O_c mentions the output value after the process of convolutional layer, the RELU activation function is represented as x and is defined as $x(a) = \max(0, a)$, input vector is indicated as a_c, convolutional kernel's weight is mentioned as w_c and the bias

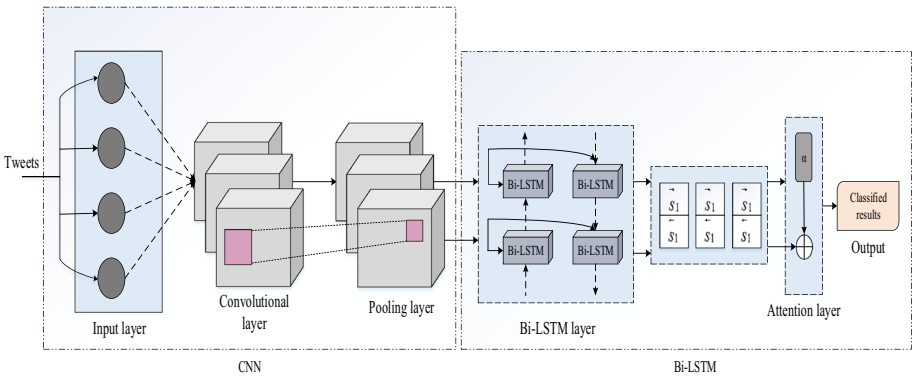


Fig. 2 Proposed DAC-BiNet model

Table 3 Hyper parameters

Parameters	Values
Learning rate	0.1
Kernel size	5
Number of epoch	100
Activation	Relu
Pool size	2
Batch size	128

of convolutional kernel is represented as b_c. The reduced features from the CNN method transmits the features to the Bi-LSTM Layer. This Bi-LSTM layer process the input features with different length by forward and backward LSTMs. The Bi-LSTM layer flows the information in both forward and backward directions with the aid of two hidden states. The main intention of this bidirectional encoding is to encode the crime data in to two directions. The memory cell's output is evaluated as,

$$\vec{s} = \left\{\vec{s}_1, \vec{s}_2, \ldots .., \vec{s}_c\right\} \tag{17}$$

$$\overleftarrow{s} = \left\{\overleftarrow{s}_1, \overleftarrow{s}_2, \ldots .., \overleftarrow{s}_c\right\} \tag{18}$$

$$s_c = \vec{s}_c \oplus \overleftarrow{s}_c \tag{19}$$

Where, $\vec{s}$ mentions the hidden state of forward direction, $\overleftarrow{s}$ represents the hidden state of backward direction and the addition operation is indicated as $\oplus$. The Output layer is updated by perform concatenation between $\vec{s}$ and $\overleftarrow{s}$. For twitter text classification, identifying the relationship among the words are highly important to classify the presented crimes. Thus, the hidden sequences are transmitted to the attention layer. This attention layer detects the importance of each words and also identifies the relationship between each of the words. Based on the higher weight values, the significant features are detected. Each word's attention weights are evaluated as,

$$A_w = \chi_w^T\left(M_w s_c + b_w\right) \tag{20}$$

$$A_w = Soft\ \max \frac{\exp\left(A_w\right)}{\sum_w \exp\left(A_w\right)} \tag{21}$$

$$h = \sum_w a_w s_w \tag{22}$$

Where, the weight matrix is signified as M_w, the bias term is represented as b_w, the transposed weight vector is given as χ_w^T, a_w is mentioned as the attention weights through softmax function and h represents the weighted sum of hidden representation. The attention layer passes the output to the output layer in which sigmoid function is utilized to predict the results. It is expressed as,

$$y = \begin{cases} 0 \ d[0, 0.5] \\ 1 \ Otherwise \end{cases} \tag{23}$$

Where, d denotes the estimated probability and the classified result is represented as y. If no crimes are available, then it is represented as 0 and 1 is represented as presented crimes. If the crime is analysed, then the classifier determines the different classes of crimes available in the tweet. In the proposed classification stage, the presented loss function degrades the detection performance. This loss function is minimized by updating the weight parameters through AO algorithm [25]. The loss function of classification stage is mentioned as,

$$L_f = \frac{1}{S} \sum_{i=1}^{S} \left(a_i - p_i^{'}\right)^2 \tag{24}$$

Where, S indicates the maximum iterations, a_i is mentioned as the actual value and the estimated value is given as $p_i^{'}$. The crimes on twitter data are analyzed based on the fitness function and is given as,

$$Fitness = Minimize\left[L_f\right] \tag{25}$$

The objective function of AO approach is to reduce the loss function by updating the weight parameters for effective crime detection. This optimizer is population based algorithm which is inspired from the Aquila's hunting behaviour. The feature initialization process of AO approach is represented as,

$$Tweet\ features = \begin{bmatrix} k_{1,1} & \cdots & k_{1,j} & \mathrm{k}_{1,\mathrm{L}-1} & k_{1,L} \\ k_{2,1} & \cdots & .k_{2,j} & \cdots & k_{2,L} \\ \cdots & \cdots & .k_{i,j} & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ k_{N-1,1} & \cdots & .k_{N-1,j} & \cdots & k_{N-1,L} \\ k_{N,1} & \cdots & k_{N,j} & k_{N,L-1} & k_{N,L} \end{bmatrix} \tag{26}$$

The above Eq. (26) mentions the current feature solution in which k_j represents the text feature's position in j^{th} solution, total amount of features is denoted as N and the size of problem is represented as L. After initialization process, the features are randomly generated as,

$$F_{ij} = rand \times \left(u^b{}_j - l^b{}_j\right) + LB_j, i = 1, 2, \ldots, N \ j = 1, 2, \ldots, L \tag{27}$$

Where, $u^b{}_j$ represents the upper bound in position j, $l^b{}_j$ is indicated as lower bound in position j and the random number is represented as $rand$. In order to obtain the optimal results, the AO algorithm performs four varied stages such as expanded exploration, narrowed exploration, expanded exploitation and narrowed exploitation.

Expanded exploration: In this stage, the search agent explores the significant features and selects optimal features for efficient crime detection. This process is defined as,

$$K(v+1) = K_{best}(v) \times \left(1 - \frac{l}{V}\right) + \left(K_P(v) - K_{best}(v) \times rand\right) \tag{28}$$

Where, $K(v+1)$ represents the upcoming iteration in v^{th} solution, the best solution that provides the useful feature's position is mentioned as $K_{best}(v)$, the term $\left(1-\frac{l}{V}\right)$ is applied to manage large exploration process. The random distribution function ranging between 0 and 1 is denoted as *rand*, v mentions the current iteration number and the maximum iteration is indicated as V.

$$K_P(v)=\frac{1}{N}\sum_{i=1}^{N}K_i(v),\forall_j=1,2,\ldots,L \tag{29}$$

Where, N is represented as the total number of features, problem size is denoted as L and v signifies the current iteration.

Narrowed exploration: In narrowed exploration, the needed feature locations are initially targeted by the search agent. Then, the search agent circles the essential features and then it chooses the feature for crime detection. The position updation of search agent is formulated as,

$$K_2(v+1)=K_{best}(v)\times\xi(L)+K_r(v)+(a-b)\times rand \tag{30}$$

Where, $K_2(v+1)$ indicates the next iteration of v^{th} solution, the feature position is mentioned as L, $\xi(L)$ denotes the function of levy flight distribution and the random solution is represented as $K_r(v)$. The computation of levy flight distribution is expressed as,

$$\xi(L)=u\times\frac{h\times\beta}{|\eta|^{\frac{1}{\chi}}} \tag{31}$$

Where, u signifies the constant value of 0.7, the random numbers ranging from 0 to 1 is represented as h and β. The constant value β is evaluated as,

$$\beta=\left[\frac{\gamma(1+\chi)\times\sin\left(\frac{\pi\chi}{2}\right)}{\gamma\left(\frac{1+\chi}{2}\right)\times\chi\times2^{\gamma-1/2}}\right] \tag{32}$$

Expanded exploitation: The search agent expands the process of exploitation by vertically descending to discover the feature characteristics. And then it starts to select the features that suitable for detecting the crimes in given twitter data. The search agents can chooses the features based on the size and quality. Moreover, the features that satisfy the fitness function are selected and is given as,

$$K_3(v+1)=\left(K_{best}(v)-K_P(v)\right)\times\varpi-rand+\left(\left(u^b-l^b\right)\times rand+l^b\right)\times\vartheta \tag{33}$$

Where, $K_3(v+1)$ is mentioned as upcoming iteration of solution v, the random function is indicated as *rand* and is ranges from 0 to 1, ϖ and ϑ represents the parameters and the upper bound and lower bound are mentioned as u^b and l^b.

Narrowed exploitation: Depending the effectiveness of the feature, the search agent can selects the useful features for classification. In this, the search agent chooses the features with last location. This behaviour is defined as,

$$K_4(v+1)=Qualify\ function\times K_{best}(v)-\left(J_1\times K(v)\times rand\right)-J_2\times\xi(L)+rand\times J_1 \tag{34}$$

Where, $K_4(v+1)$ represents the searching process of last iteration, the term *Qualify function* helps to control the strategy of exploring process, the varied feature characteristics are considered as J_1, reducing value is denoted as J_2 and the current solution is signified as $K(v)$. The quality function is computed as,

$$Qualify\ function(v) = v^{\frac{2\times rand-1}{(1-V)^2}} \tag{35}$$

$$J_1 = 2 \times rand - 1 \tag{36}$$

$$J_2 = 2 \times \left(1 - \frac{v}{V}\right) \tag{37}$$

Where, $Qualifyfunction(v)$ represents the quality function's range in iteration v and *rand* denotes the random number from the interval of 0 to 1. For each iteration, the fitness function is evaluated and the obtained solution in each of the iteration is compared with each other. Then, the optimal solution is attained by analysing the best fitness function. Hence the presented crimes from the twitter data are efficiently detected in the proposed DAC-BiNet model. The Pseudocode of AO approach is illustrated in Table 4.

From the proposed DAC-BiNet classification model, the varied classes of crimes on twitter like fraud, harassment, police, violence and crimes are detected from the provided input data. With the aid of Aquila optimization, the presented loss function is reduced and it makes the classification to obtain better performance.

4 Results and discussions

In this section, the proposed DAC-BiNet crime detection model is examined using Twitter user data dataset. In the proposed work, the experimental setup is performed through python tool and the performance of proposed techniques are measured by evaluating several performance metrics against other existing techniques. By using Twitter user data dataset (https://data.world/data-society/twitter-user-data), the simulation analysis of proposed DAC-BiNet model is performed. This dataset involves 20,000 rows, which includes random tweets, images, location information, user name and several account profiles. Also, the utilized dataset contains five varied types of classes such as crime, fraud, harassment, violence and police. Various existing works utilized this dataset and attains better outcomes. Thus, the proposed study prefer this dataset for detecting crimes on twitter data.

4.1 Performance metrics

The major performance metrics utilized to compute the efficiency of the proposed model are accuracy, specificity, recall, precision, F1-score and processing time. The determination of classification accuracy exhibit the effectiveness of the model. The attained results of proposed model is compared with some of the other deep learning approaches like CNN, Bi-LSTM, DNN (Deep Neural Network) and DBN (Deep Belief Network).

- Accuracy

Table 4 Pseudocode of AO algorithm

Algorithm : AO approach

Initialize the features
Initialize parameters like $\varpi, \chi, ...$
if (condition is not satisfied) **do**
Evaluate the fitness function using equation (25)
$K_{best}(v)$= Compute the best solution based on the fitness function
for $(i = 1,2,....,N)$do
Update the current solution $K_P(v)$
Update q, r, J_1, J_2 and $\xi(L)$
if $rand \leq 0.5$ then
Update the solution using equation (28)
if fitness $(K_1(v+1)) < fitness(K(v))$ **then**
$K(v) = K_1(v+1)$
end if
else
Update the solution using equation (30)
if fitness $(K_2(v+1)) < fitnessK(v))$ **then**
$K(v) = K_2(v+1)$
end if
else
if $rand \leq 0.7$ **then**
Update the solution using equation (33)
if fitness $(K_3(v+1)) < fitness(K(v))$ **then**
$K(v) = K_3(v+1)$
end if
end if
else
Update the solution using equation (34)
if fitness $(K_4(v+1)) < fitness(K(v))$ **then**
Using equation (35), evaluate quality function
end if
return

Accuracy is an essential metric which supports to detect the performance of classification in classifying various classes of crimes on twitter data. It is defined as the ratio of correctly categorized data to the total amount of data. The formulation of accuracy is given as,

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{38}$$

Where, *tp* indicates true positives, *tn* represents true negatives, *fp* mentions false positives and *fn* represents false negatives.

- Specificity

The specificity measure computes the ability of the classification model to categorize true negatives of every samples. The term specificity is also said as true negatives and is evaluated as,

$$Specificity = \frac{tn}{tn + fp} \tag{39}$$

Where, *tn* represents true negatives and *fp* signifies false positives.

- Recall

The recall metrics computes the number of accurately classified data from the overall positive predictions done by the proposed DAC-BiNet model. Recall is also called as sensitivity and also this metric represents missed number of perfectly classified data. It is measured as,

$$\text{Re } call = \frac{tp}{tp + fn} \tag{40}$$

Where, *tp* indicates true positives and *fn* represents false negatives.

- Precision

The precision metric identifies the amount of accurate positive predictions made by the proposed classifier. It is defined as the ratio of perfectly classified positive samples to the overall amount of positive samples that the classifier were predicted. The formulation of prediction is expressed as,

$$\text{Pr } ecision = \frac{tp}{tp + fp} \tag{41}$$

Where, *tp* is mentioned as true positives and *fp* is signified as false positives.

- F1-score

The metric F1-score is the harmonic mean of both recall and precision which illustrates the accuracy measure of the test. The F1-score with value one is represented as better classification and is mentioned as,

$$F_1 - score = 2 \times \frac{P \times r}{P + r} \tag{42}$$

Where, *P* represents precision and *r* is mentioned as recall.

4.2 Performance evaluation

The performance evaluation is performed with the twitter data gathered from the dataset using deep learning model. The twitter data presented in the dataset involves varied classes such as fraud, crime, harassment, violence and police. To perform twitter crime detection, the proposed study used different powerful techniques. The important features extracted from the feature extraction stage reduces the computational complexity and the clustering

process groups the features in to low dimensionality of feature set. This helps the classifier to easily detect the crimes on twitter data. This section presents a comparative analysis using various deep learning approaches. For performance comparison, the proposed work uses the conventional techniques like Bi-LSTM, CNN, DBN and DNN. The confusion matrix of proposed DAC-BiNet classification is shown in Fig. 3.

The confusion matrix exhibit the prediction ability of the proposed model. In each of the testing phase, varied classes of crimes are detected. From the 3011 of crime data, 2891 are correctly classified as crime and the remaining number of data like 29, 28, 30 and 33 are wrongly classified as fraud, violence, police and harassment. Out of the 1745 fraud data, 1684 are accurately classified as fraud and only few of them are misclassified such as 31 data are misclassified as crime, 11 are misclassified as violence, 7 are misclassified as police and 12 are misclassified as harassment. Form the total of 673 violence data, 599 data are correctly classified as violence and the 16 are misclassified as crime, 17 are wrongly classified as fraud, 18 are misclassified as police and 23 are wrongly predicted as harassment. From the 101 of police data, 73 data are accurately categorized as police and the remaining are misclassified as other classes. Similarly from the total of 4516 harassment data, 4356 are correctly classified as harassment data and only the reduced number of data are misclassified. This shows that the proposed classifier is effective in terms of crime detection using different classes. Figure 4 depicts the accuracy performance comparison of proposed and existing techniques.

The accuracy value of proposed model is compared with various existing techniques like CNN, Bi-LSTM, DNN and DBN. The comparative analysis clearly shows that the proposed classifier is obtained better accuracy performance than others. In the proposed work, noisy data in the dataset are eliminated in the pre-processing stage through ten different steps. It improves the data quality and helps the classifier to perform well. Also, the optimization technique directly improves the accuracy of classification stage by updating the weight parameters. The hybrid deep learning technique of CNN and

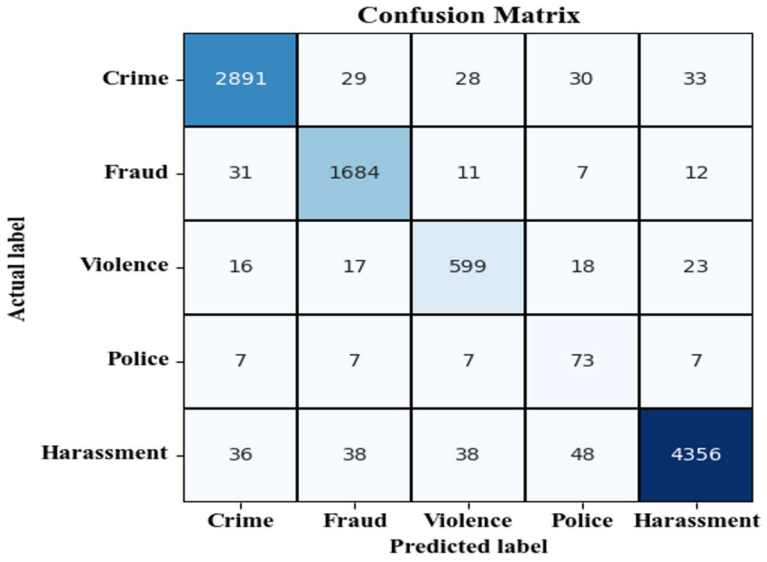


Fig. 3 Confusion matrix of proposed DAC-BiNet model

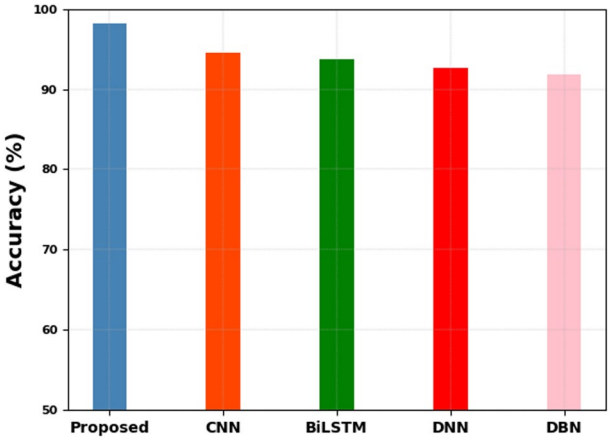


Fig. 4 Comparison analysis using accuracy performance

Bi-LSTM extracts the significant features by reducing the feature size and detect the crimes with higher accuracy. But the existing techniques fails to bring higher accuracy in crime detection because of increased computational complexity. Also, the existing methods extracts the features in vector form and mainly such features are in higher dimensionality. This limitation affects the performance of the classification and generates lower accuracy. This limitations are overcome in the proposed work by using robust techniques. The accuracy of proposed DAC-BiNet is 98.23%, CNN is 94.508%, Bi-LSTM is 93.724%, DNN is 92.698% and DBN is 91.838%. Figure 5 illustrates the specificity performance comparison of proposed and existing techniques.

The specificity analysis of the above figure apparently shows that the proposed model is superior to the other deep learning techniques. The attained specificity value of proposed work is 98.86%, CNN is 96.53%, BiLSTM is 95.65%, DNN is 95.03% and DBN is 94.46%. The specificity of conventional approaches are reduced because of several issues. The high dimensionality of features in the existing techniques generates several problems during the detection of crimes. But, the reduced size of extracted features

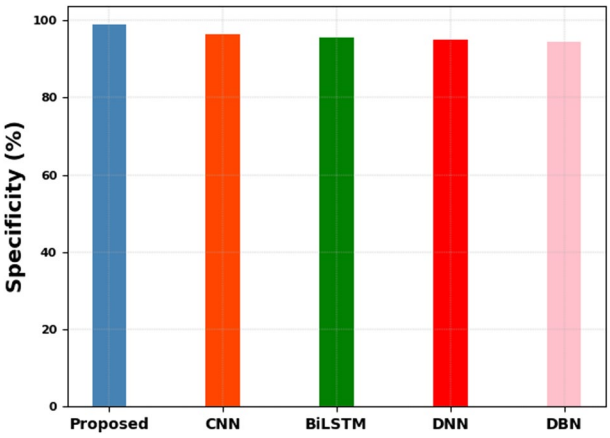


Fig. 5 Comparison analysis using specificity performance

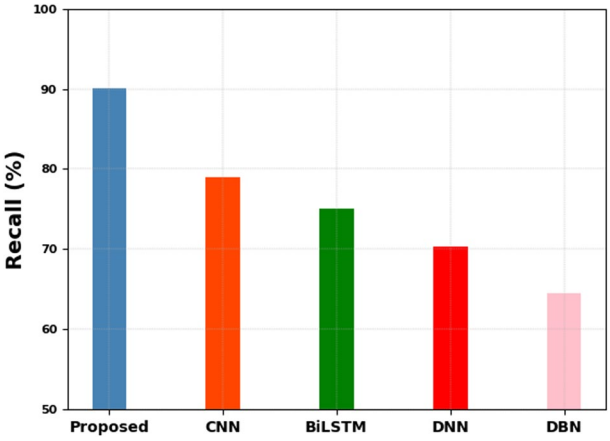


Fig. 6 Comparison analysis in terms of recall performance

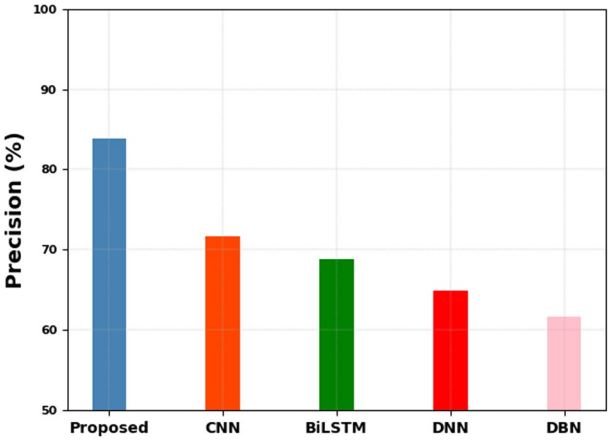


Fig. 7 Comparison analysis in terms of precision

from the CNN method support to obtain good specificity performance. Figure 6 represents the recall performance comparison of proposed and various existing approaches.

The recall performance of proposed classification technique is compared with several conventional deep learning techniques. The results analysis represents that the proposed DAC-BiNet model is better in terms of recall performance among the others. The attained recall value of proposed work is 90.05%, CNN is 79%, BiLSTM is 75%, DNN is 70.29% and DBN is 64.55%. The increased training capability in Bi-LSTM and efficient CNN makes the system recall performance as large. This analysis exhibit that the proposed techniques are more suitable for twitter crime detection with varied classes. Figure 7 illustrates the precision performance of proposed with different existing techniques.

By comparing the proposed precision performance against other deep learning techniques, it clearly seems that the precision is increased in the proposed technique. In the proposed crime detection model, the precision performance is increased as 83.86%. But, the precision of existing techniques are slightly different from the proposed one. The obtained precision value of CNN is 71.62%, BiLSTM is 68.82%, DNN is 64.90% and DBN

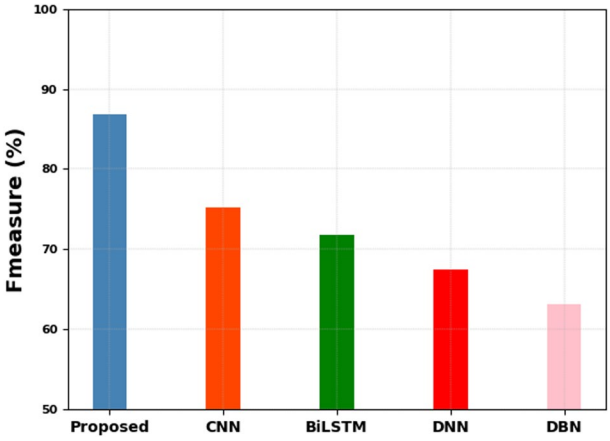


Fig. 8 Comparison analysis in terms of F1-score

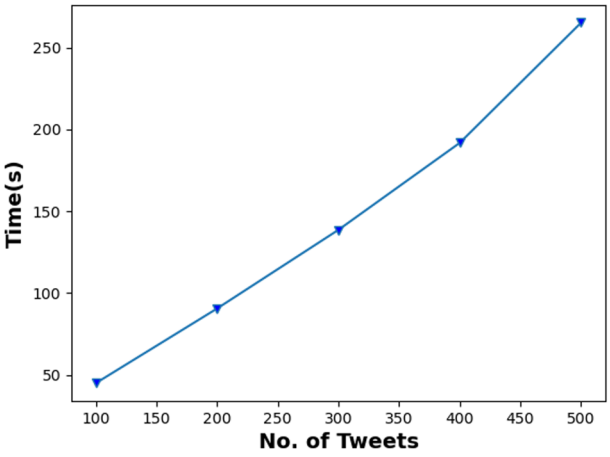


Fig. 9 Processing time of proposed crime detection model

is 61.70%. This results shows that the proposed model is highly effective than the other approaches. Figure 8 represents the performance comparison in terms of F1-score.

The attained F1-score of proposed model is compared with other conventional techniques. In this analysis, it apparently exhibit that the proposed DAC-BiLSTM classifier obtains increased F-measure value than the other techniques. The F1-score of proposed is 86.84%, CNN is 75.13%, BiLSTM is 71.78%, DNN is 67.49% and DBN is 63.09%. This represents that the proposed techniques are outperformed other existing approaches. The processing time of proposed model is depicted in Fig. 9.

The processing time of proposed work is analysed by varying the number of input tweets. The proposed model attains reduced processing time for detecting the crimes. The clustering approach in the proposed model groups the feature set with minimized size and it helps to reduce the running time of the system. The reduced processing time directly increases the system's performance and represents that the proposed model is appropriate for crime detection. The overall performance of the proposed work is shown in Table 5.

The ROC curve of proposed model with different classes of tweet data are represented in Fig. 10.

Table 5 Overall performance

Methods	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1 score (%)
Proposed	98.23	83.86	90.05	98.86	86.84
CNN	94.508	71.62	79	96.53	75.13
BiLSTM	93.724	68.82	75	95.65	71.78
DNN	92.698	64.90	70.29	95.03	67.49
DBN	91.838	61.70	64.55	94.46	63.09

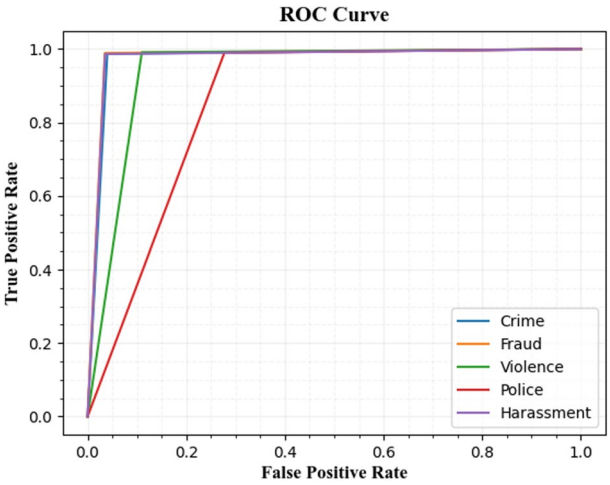


Fig. 10 ROC curve of proposed model

The ROC curve of proposed model is analysed by varying the true positive rate and false positive rate. For each of the classes, the value of ROC is obtained one and it mentions that the proposed classification is better for detecting the crimes in different classes. In the proposed work, initially the noisy data are removed and obtain the high quality of data for further process. After that, the features are extracted using three different techniques. From that, the features which are suitable for crime detection are taken to the next process. This feature extraction stage minimizes the complexity of the system. Then, the clustering stage groups the features in to minimized set which improves the prediction capability of classifier. Because of reducing the feature size, the computational is become very easy and also the crimes are detected with reduced time. But the existing CNN approach are not able to detect the crime accurately because of overfitting issue and large computational complexity. Also, the previous approaches are suffers from large feature dimensionality problem. But, the proposed work uses high efficient techniques and obtains improved results for crime detection mechanism.

5 Conclusion

This paper introduces a new DAC-BiNet model for crime detection in twitter platform. Because of higher efficiency in the proposed model, the classification accuracy of proposed model is increased as 98.23%. For crime analysis, the proposed study utilized

publicly available dataset and the raw inputs are pre-processed to eliminate the unwanted noises using diverse steps. After pre-processing, the most essential features are extracted to enhance the classification performance in the feature extraction stage. Then, the features are clustered to reduce the high feature dimensionality issue using Possibilistic Fuzzy LDA scheme in which the features are turned in to reduced set of clusters. Also, the proposed feature reduction stage helps to minimize the overall processing time by reducing the amount of feature sets and is clearly represented in Fig. 9. With the help of attained features, the crime data from the input tweets are detected and classified by the proposed DAC-BiNet model. Because of the presence of loss function, the performance of classifier may get affected. Thus, to solve this problem, the parameters are fine-tuned using Aquila optimization. Therefore, the classification performance of proposed model is enhanced as compared with other existing methods. The superior outcomes that achieved by the proposed technique in terms of varied matrix is accuracy as 98.23%, precision as 83.86%, recall as 90.05%, specificity as 98.86% and F1-score as 86.84%. However, the reduced detection accuracy attained by an existing CNN is 94.5%, BiLSTM is 93.725, DNN is 92.6% and DBN is 91.8%. In future study, different word embedding techniques will be introduced in the feature extraction stage which may produce additional features and bring improved results. Moreover, the proposed model is extended to process with real time data in the further research.

Funding No funding was received for conducting this study.

Data availability The manuscript does not have any associated data which was generated.

Declarations

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Conflicts of interests The authors declare that they have no conflict of interest and have no competing interests to declare that are relevant to the content of this article.

References

1. McMullan JL, Rege A (2010) Online crime and internet gambling. J Gambl Issues 24:54–85
2. Drew JM, Moir E, Newman M (2021) Financial crime investigation: an evaluation of an online training program for police. Polic Int J 44(3):525–539
3. Almadhoor L (2021) Social media and cybercrimes. Turk J Comput Math Educ (TURCOMAT) 12(10):2972–2981
4. Razak NA, Abdul Malik HK, Rentah F, Mazlan MNH (2021) A Comprehensive Study of Privacy and Security Risk Awareness Among Mobile Internet Users for Social Networks Sites in Malaysia. Int J Bus Technol Manag 3(1):1–20
5. Sandagiri C, Kumara BTGS, Kuhaneswaran B (2021) Deep Neural Network-Based Crime Prediction Using Twitter Data. Int J Syst Service-Oriented Eng (IJSSOE) 11(1):15–30
6. Metzger MJ, Flanagin AJ, Mena P, Jiang S, Wilson C (2021) From dark to light: The many shades of sharing misinformation online. Media Commun 9(1):134–143
7. Thurtell SC (2021) When Shopping Turns Scary: Twitter Conversations about Violent Crime in Johannesburg Malls. Critical Arts 35(1):85–96
8. Mahor V, Rawat R, Telang S, Garg B, Mukhopadhyay D, Palimkar P (2021) Machine Learning based Detection of Cyber Crime Hub Analysis using Twitter Data. In 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON), pp 1–5. IEEE

9. Dutta HS, Aggarwal K, Chakraborty T (2021) DECIFE: Detecting collusive users involved in black-market following services on Twitter. In Proceedings of the 32nd ACM Conference on Hypertext and Social Media, pp 91–100
10. Asaad RR, Abdulhakim RM (2021) The Concept of Data Mining and Knowledge Extraction Techniques. Qubahan Acad J 1(2):17–20
11. Saura JR, Palacios-Marqués D, Ribeiro-Soriano D (2021) Using data mining techniques to explore security issues in smart living environments in Twitter. Comput Commun 179:285–295
12. Ojha D, Singh RP, Jadon KS (2021) An Analysis on Data Mining for Sentiment Analysis Using Different Classification Algorithms. In 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT), pp 533–538. IEEE
13. Ahadi A, Singh A, Bower M, Garrett M (2022) Text Mining in Education—A Bibliometrics-Based Systematic Review. Educ Sci 12(3):210
14. He J, Nguyen DQ, Akhondi SA, Druckenbrodt C, Thorne C, Hoessel R, Afzal Z et al (2021) Chemu 2020: Natural language processing methods are effective for information extraction from chemical patents. Front Res Metrics Anal 6:12
15. Benicio DHP, Xavier-Júnior JC, de Paiva KRS, De Araújo Sant Camargo JD (n.d.) Applying Text Mining and Natural Language Processing to Electronic Medical Records for Extracting and Transforming Texts into Structured Data
16. Raja RA, Yuvaraj N, Kousik NV (2021) Analyses on Artificial Intelligence Framework to Detect Crime Pattern. Intelligent Data Analytics for Terror Threat Prediction: Architectures, Methodologies, Techniques and Applications, pp 119–132
17. Rouhollahi Z (2021) Towards Artificial Intelligence Enabled Financial Crime Detection. *arXiv preprint* arXiv:2105.10866
18. Rigano C (2019) Using artificial intelligence to address criminal justice needs. Natl Inst Justice J 280:1–10
19. Chackravarthy S, Schmitt S, Yang L (2018) Intelligent crime anomaly detection in smart cities using deep learning. In 2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC), pp 399–404. IEEE
20. Lal S, Tiwari L, Ranjan R, Verma A, Sardana N, Mourya R (2020) Analysis and classification of crime tweets. Procedia Comput Sci 167:1911–1919
21. AlGhamdi MA, Khan MA (2020) Intelligent analysis of Arabic tweets for detection of suspicious messages. Arab J Sci Eng 45(8):6021–6032
22. Santhiya K, Bhuvaneswari V, Murugesh V (2021) Automated Crime Tweets Classification and Geo-location Prediction using Big Data Framework. Turk J Comput Math Educ (TURCOMAT) 12(14):2133–2152
23. Boukabous M, Azizi M (2022) Crime prediction using a hybrid sentiment analysis approach based on the bidirectional encoder representations from transformers. Indones J Electr Eng Comput Sci 25(2):1131–1139
24. Hissah AS, Al-Dossari H (2018) Detecting and Classifying Crimes from Arabic Twitter Posts using Text Mining Techniques. Int J Adv Comput Sci Appl (IJACSA) 9(10)
25. Abualigah L, Yousri D, AbdElaziz M, Ewees AA, Al-Qaness MAA, Gandomi AH (2021) Aquila optimizer: a novel meta-heuristic optimization algorithm. Comput Ind Eng 157(3):107250

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

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

Physica Scripta

INCLUSIVE PUBLISHING
TRUSTED SCIENCE

PAPER

Design of 8 -bit low power SAR ADC in 45 nm for biomedical implants

To cite this article: Mohit Tyagi *et al* 2023 *Phys. Scr.* **98** 116101

View the article online for updates and enhancements.

You may also like

- Design voltage comparator 14-bit for successive approximation analog-to-digital converter
Nur Zazmera Mustafa Kamal, Nabihah Ahmad, Siti Hawa Ruslan et al.
- A 14-bit 100-MS/s CMOS pipelined ADC featuring 83.5-dB SFDR
Nan Zhao, , Qi Wei et al.
- Successive approximation type resistance to digital converter
Nimal Jagadeesh Kumar and Jagadeesh Kumar

This content was downloaded from IP address 14.139.251.98 on 03/11/2023 at 05:42

Physica Scripta

PAPER

Design of 8 -bit low power SAR ADC in 45 nm for biomedical implants

RECEIVED
9 June 2023

REVISED
28 September 2023

ACCEPTED FOR PUBLICATION
10 October 2023

PUBLISHED
23 October 2023

Mohit Tyagi[1,2], **Poornima Mittal**[1] and **Parvin Kumar**[2]

[1] Department of Electronics and Communication Engineering, Delhi Technological University, Delhi-110042, India
[2] Department of Electronics and Communication Engineering, KIET Group of Institutions, Delhi-NCR, Ghaziabad, 201206, India
* Author to whom any correspondence should be addressed.

E-mail: mohit.tyagi@kiet.edu, poornimamittal@dtu.ac.in and parvin.kaushik@kiet.edu

Keywords: successive approximation register (SAR) – analog to digital converters (ADC), digital to analog converter (DAC), sample & hold circuitry (S&H), integral non-linearity (INL), differential non-linearity (DNL)

Abstract

The utilisation of low power SAR (Successive Approximation Register) Analog-to-Digital Converters holds significant importance in the domain of bio-medical signal acquisition. The present study showcases the utilisation of an 8-bit CMOS SAR-ADC for integration into the analog front end of bio-signal acquisition. The focus of this technology pertains to the monitoring of implanted bio-signal devices, with a specific emphasis on ECG/EEG signals. A capacitive digital-to-analog converter (DAC) is suggested as a means to attain power consumption in the microwatt range. This approach enables comparisons to be made without any energy consumption, leading to a substantial enhancement in energy efficiency. Furthermore, a comprehensive theoretical examination of comparator offset voltages has been conducted to enhance the offset performance of the comparator operating at low supply voltage. The analysis indicates that optimization of the comparator is achieved solely through the adjustment of transistor sizes, without the implementation of any specific offset cancellation techniques. Simulations indicate that the optimization of the offset voltage to approximately 5 mV occurs when there is variation in the common-mode input voltage at a 1 V supply. The proposed Analog-to-Digital Converter (ADC) layout has been successfully executed utilizing the 45 nm Complementary Metal-Oxide-Semiconductor (CMOS) technology. The Analog-to-Digital Converter (ADC) attains a Spurious-Free Dynamic Range (SFDR) of 64.02 dB and consumes 1.9 μW of power at a sampling rate of 1.1 MHz and a supply voltage of 1 V.

1. Introduction

As CMOS technology advances, there is a greater need for integrated low power cardiac implanted medical devices such as defibrillators as well as pacemakers. It demands circuit implementation with low power consumption for these biomedical devices. The state-of-the-art for low-power bio-signal acquisition and processing is generally recognized as being a physiologically detecting analog front-end (AFE), which consists of a low-noise amplification circuit, an ADC, and prime actual bio-feature filters utilizing analog circuitry [1, 2].

Analog-to-Digital Converters represent the most power hungry and crucial component in monitoring cardiovascular signals among the different components present in implanted devices. Biomedical cardiac signals require analog to digital converter to perform with medium to high resolution at low to medium speed conditions [3]. As a result, creating an ADC in cardiac devices with minimal power consumption and satisfying the required conversion precision is a significant problem. In this view, low power Architecture design of 8-bit SAR ADC is the best choice to design among the available ADC architectures like flash, delta-sigma and two step ADCs due to the simple construction of SAR ADC along with the usage of maximum digital circuitry in its implementation [4]. Technology scaling has reduced the energy consumption of ADC to a great extent. The energy efficiency of ADC is characterised by a Walden figure of merit as:

© 2023 IOP Publishing Ltd

$$\text{Figure of Merit} = \frac{\text{Power}}{2 \ \text{Bandwidth} \ 2^{\text{ENOB}}} \quad (1)$$

Where ENOB stands for effective number of bits. Figure of merit denotes the energy required per each conversion step in Joule/conversion step as its unit. So there exists a tradeoff between power consumption in an ADC circuit and its maximum operating frequency [5]. Power consumption increases substantially with an increase in the operating speed of a device [6]. In a SAR ADC, comparator and digital to analog converter (DAC) modules are the major contributors of power consumption in the circuit implementation. Further, Leakage current and kickback noise in the comparator block is required to be focused upon to improve linearity of the device [7]. Also, DAC consumes power during switching of capacitors between Vref and ground in the switching process of preset-measure-reset in SAR logic [8]. Reduction of switching energy per conversion cycle in designed DAC is one area of importance for the researchers working in mixed signal domain of ADC's implementation. With the mentioned considerations, Jajodia *et al* [9] worked on energy efficient DAC switching technique for single ended SAR Analog to digital converters in which the reference voltage was reduced to Vref/2 to minimise the energy needed per conversion step in the switching of DAC module. Proposed single ended SAR ADC achieved the resolution of 10 bits at sampling rate of 40 MS/s with SNDR of 55.97 dB.

Song *et al* [10] worked on 12-bit 1MS/s low power SAR ADC in 180 nm technology node. The design considered the low fan in SAR logic to reduce the capacitive load on the comparator to reduce the power consumption. Hsieh *et al* [11] implemented the semi-resting (SR) digital to analog converter which reduces the differential nonlinearity and integral nonlinearity to be $\pm$ 0.5 LSB. With the semi-resting approach, the same signal swing was incorporated with reduction in the reference voltage. Further, duty cycled reference generation in 65 nm CMOS technology is implemented by Liu *et al* [12]. The stable reference voltage generator has only $\pm$ 1.5 percent variation over process corners. The implemented circuit comes out with figure of merit of 2.4 fJ/ conversion step.

X. Tang [13] reviewed design techniques for a low power SAR ADC. In the work, noise analysis of comparator in terms of kickback noise analysis and jitter analysis in DAC module of SAR Further In order to enhance the speed and reduction of the power consumption of the SAR ADC, an asynchronous method is implemented [14–16]. Additionally, efforts have been made to reduce the complexity and area of the control logic for SAR. Dithering is an alternate approach that has been suggested to eliminate DAC faults and increase linearity [17, 18]. Liu *et al* [19] demonstrated a reconfigurable CDAC to achieve many conversion modalities without reducing capacitance to attenuate the switching energy per conversion cycle. Liu *et al* [20] worked on monotonic capacitor switching in which input common mode voltage converged to ground gradually to reduce the power consumption and the circuit consumed 0.826 mW of power at sampling rate of 50 MS/s.

Analog front end sensing requirements are analyzed for an analog to digital converter keeping in view of biomedical signals by Jo *et al* [21] and a correction strategy for the sensor's baseline was described by Rezaeiyan *et al* [22] which has the potential to reduce circuit overloading and improve sensor sensitivity. Roh *et al* [23] proposed a SAR-assisted SAR ADC using a dual clock-rate coarse decision approach to minimise the conversion time. The ADC obtains a SNDR of 58.1 dB and a SFDR of 72.1 dB using an 80-MHz input. Seong *et al* [24] introduced the on-chip background offset. A comparator offset is used to compensate clocked edge mismatch for sub-ADC sampling harmonic distortion and the circuit achieved SNDR of 40 dB in Nyquist input while consuming 7.57 mW, resulting in a FoM of 37.2 fJ/conversion-step. In [25, 26] two-step and binary weighted switching schemes are implemented in order to identify capacitor mismatches in DAC module.

This paper proposes an 8 bit Low-power SAR ADC for biomedical implants. The design includes a novel low power dynamic latched comparator that reduces kickback noise for improvement of linearity at 1 V using 45 nm CMOS technology node. Also, this paper presents energy efficient switching scheme based capacitive digital to analog converter (CDAC). The bottom plate sampling approach along with variable reference voltage between Vrefn—Vrefp through switch buffering is considered in the DAC implementation. The DAC design employs the reduction in the amount of total capacitance, resulting in a reduction in power usage and DAC area. The array capacitance value within DAC gets adjusted down to the closest integer value, resulting in uniform geometries. With the mentioned inclusions in the design, proposed Analog-to-Digital Converter (ADC) attains a Spurious-Free Dynamic Range (SFDR) of 64.02 dB and consumes 1.9 μW of power at a sampling rate of 1.1 MHz and a supply voltage of 1V.

Parameters considered for the performance measurement of an overall designed SAR ADC are Signal to noise and distortion ratio (SNDR), SFDR, Effective number of bits (ENOB), power consumption, maximum sampling rate, Differential nonlinearity (DNL) and Integral nonlinearity (INL). Also, parametric optimization at individual block level of comparator module and DAC module in a SAR ADC are also point of importance. Optimization parameters in comparator design are offset voltage, leakage power minimization and delay reduction while in a DAC switching energy CV_{DD}^2 has been reduced during each conversion process of preset-measure-reset.

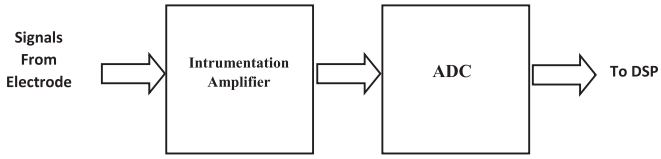


Figure 1. Diagrammatic illustration of the ECG monitoring system.

Table 1. Significant features of diverse bio signals [27].

Bio-Signal Name	Frequency Range (Hz)	Amplitude (mV)
ECG	0.01 to 250	0.5 to 4
EMG	Up to 2 K	0.1 to 5
EEG	0.5 to 150	0.0005 to 0.3

2. Proposed methodology

The human body generates bio-potential signals including EMG, EEG, and ECG. In the low frequency region, these bio-potential signals have both amplitude and voltage. The frequency spectrum and voltage amplitude of particular biopotential signals are shown in table 1 [27]. Monitoring and documenting bio potential signals has been essential for medical diagnosis, but current medical standards demand that these signals be documented. Often, patients are attached to extended recording devices to collect diagnostic body signals. This restricts their mobility and causes worry. Consequently, acquisition time is reduced, and ongoing patient monitoring is avoided, influencing the total disease diagnosis.

Figure 1 depicts the Analog Front End of an ECG/EEG recording device. This recording device consists of several parts, including an operational amplifier and an ADC. The operational transconductance amplifier (OTA) is considered a bio amplifier in bio signal detecting devices [28]. Since the bio amplifier is the first and most critical component in a bio potential signal detecting system, it may use a significant amount of energy while functioning, hence increasing the overall power consumption of the detection system.

The technique is presented completely as a step-by-step solution for the problem. According to figure 2, the design parameters must be established based on a review of the relevant literature. Second, device technology must always be chosen depending on the requirements. The proposed circuit would then be simulated using optimization techniques to ensure that it fits the specified requirements. Lastly, the proposed technique to enhance performance is offered. For low power operation, the transistors are intended to perform in weak inversion or moderate inversion regions. In the unified all-regional theory, often known as the 'EKV Model,' the question of building or operating a transistor in the inversion area has been examined. The concept of inversion coefficient has relied heavily on the unification of all-regional MOSFETs.

The model, famous for its creators, is utilised to reduce both noise and power [29–34]. The EKV configuration can be determined using the MOS transistor formula below.

$$IC = \frac{I_D}{I_S} \tag{2}$$

$$g_m = \frac{I_D}{\eta \Phi_t} \frac{1 - \exp(-\sqrt{IC})}{\sqrt{IC}} \tag{3}$$

$$I_s = 2\mu C_{ox} \frac{W}{L} \varnothing_t^2 \tag{4}$$

Where, I_S is the normalization current, drain current is represented by $I_{D,}$ η is denoted as slope factor, which is normally taken as 1, C_{ox}- oxide capacitance, $\in_0$ is permittivity of free space, $\in_{si}$ is relative permittivity of silicon and ϕt -thermal voltage at room temperature, taken as 25.6 mV.

3. Proposed design of SAR ADC for bio-medical applications

As illustrated in figure 3, the 8-bit SAR ADC comprises a sample-and-hold (S/H) stage, a digital-to-analog converter (DAC), a comparator, and a SAR logic module. The S/H stage circuitry and comparator design would be examined simultaneously.

The concentration of the concept is on achieving the criterion of 8-bit precision for a constant voltage range of 1 V while using a very small amount of power.

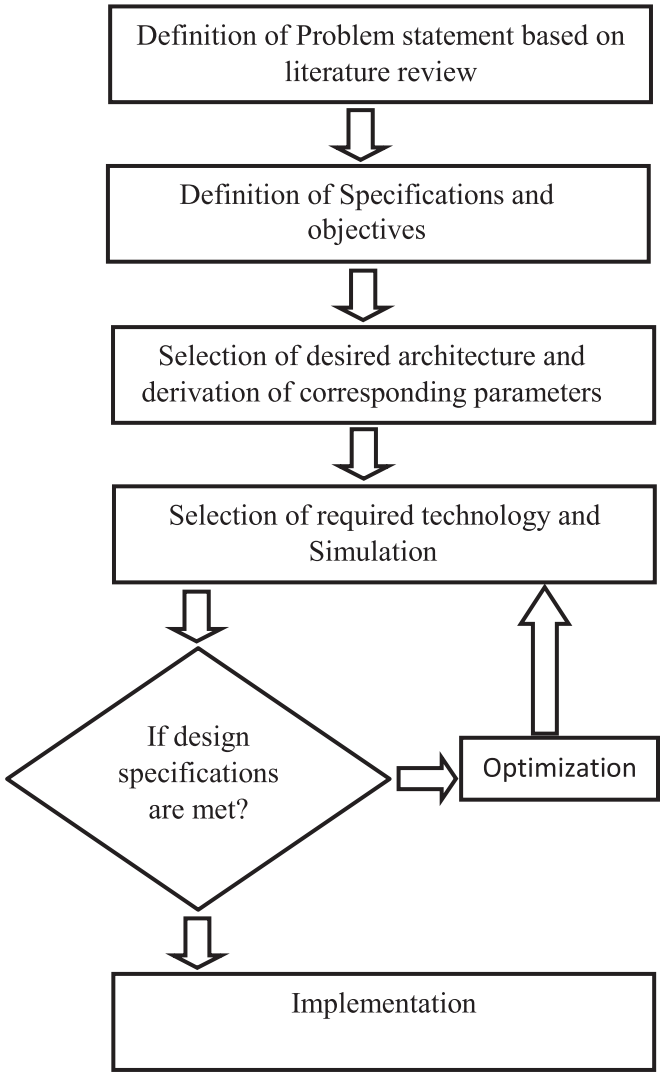


Figure 2. Methodology followed in Work.

3.1. Sample and hold circuit

Figure 4 depicts the proposed S/H circuit. The MOSFET transistors Mn1,2, Mp1,2, function as toggles and maintain the sampled voltage in the offset capacitance C_{offset} throughout the hold stage (i.e., clk becomes 0). Also, adjustment of the gate to source voltage (V_{GS}) of the primary switching device MnSW throughout the sampling or tracking stage (clk equals 1) will take place. When the source voltage of the transistor MnSW changes, the enormous value V_{GS} guarantees that it is entirely ON and remains significantly larger than the value of the threshold voltage V_{th}. MnSW gate node as well as Mp2 gate node have a common connection as Vg & Mp2's sources are kept at V_{DD} throughout sampling. To be entirely off, Mp2's gate needs to be connected to the most significant value in the entire circuit, which is Vg. In the hold stage, Mn4 couples Vg to the minimum value in the circuit to guarantee that MnSW is completely turned off. Whenever Mn4 is turned off, Mn3 is given to keep it from getting an extremely high V_{GD}. Transistor Mn3's gate voltage is maintained as V_{DD}, Mn3 will not function as an efficient pass device, therefore Mn4's drain will have to be restricted to a max value V_{DD}-V_{th} throughout the sampling procedure, but at the cost of a wider area. Mn1 and Mn2 use two opposing clocks (clk and clk_b), the clock feedthrough impact is balanced out when MpSW is introduced in parallel to MnSW.

3.2. DAC

The Digital-to-Analog Converter (DAC) is tasked with the conversion of digital code emanating from the Successive Approximation Register (SAR) logic to an analog voltage level that is subsequently contrasted by the comparator [35, 36]. As the resolution of the system grows, there is an increase in the mismatch of the capacitor array [37]. As a result, the linearity of the SAR ADC is reduced. The reduction in size of CMOS technologies facilitates the augmentation of sampling frequency. It is worth noting that the majority of the overall power

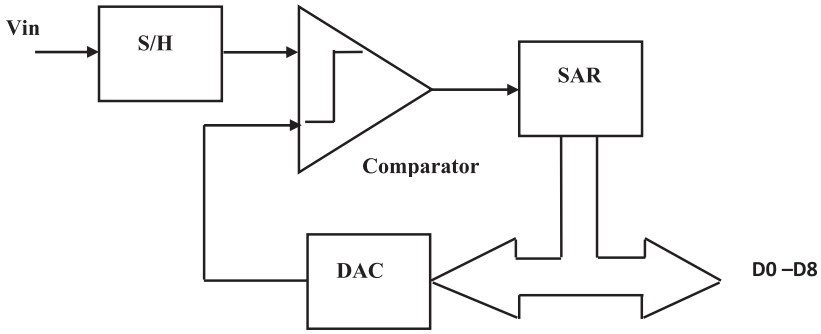


Figure 3. Architecture of SAR ADC.

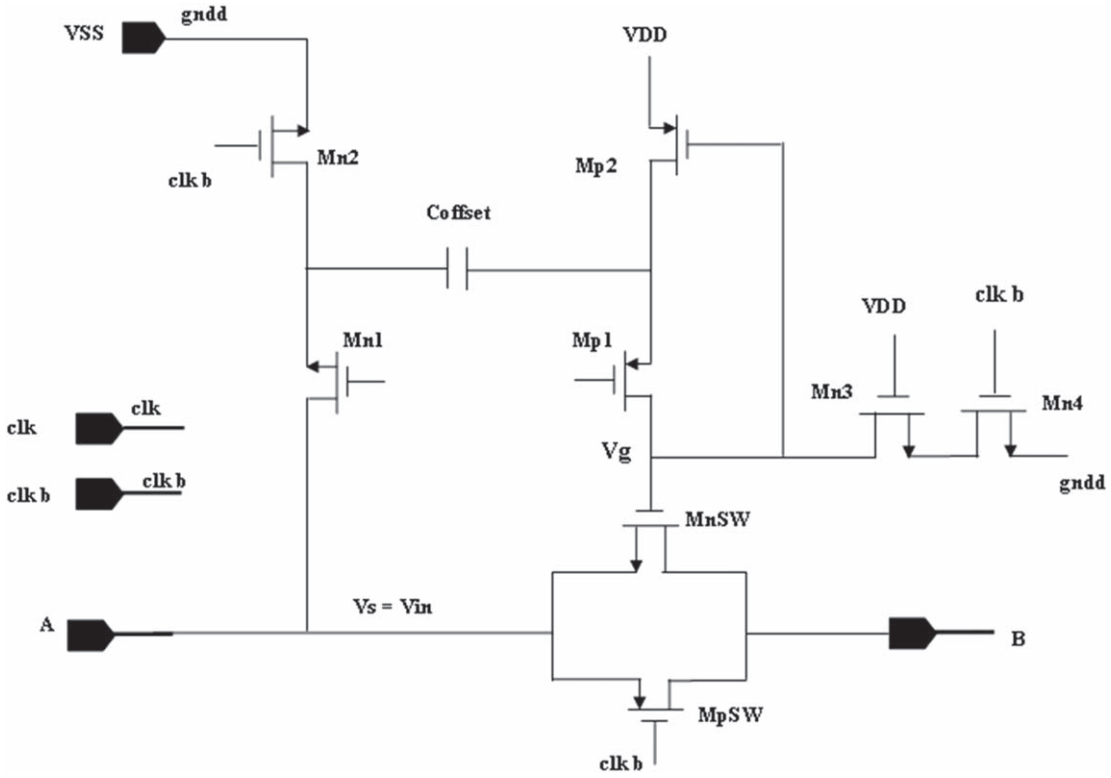


Figure 4. Circuit of proposed sample & hold circuit.

consumption is attributed to the digital-to-analog converter (DAC), which is specifically allocated for the capacitor array as well as switching mechanism [38, 39]. One of the primary concerns in mitigating energy consumption involves the shrinking of array capacitors along with the implementation of alternative approaches [6]. This paper presents energy efficient switching scheme based capacitive digital to analog converter (CDAC) in figure 5. The bottom plate sampling approach along with variable reference voltage between Vrefn—Vrefp through switch buffering is considered in the DAC implementation. The DAC design employs the reduction in the amount of total capacitance, resulting in a reduction in power usage and DAC area. The array capacitance value within DAC gets adjusted down to the closest integer value, resulting in uniform geometries. Instead of a single level voltage, the concept of level Voltages is considered. Primarily, power consumption reduction has been achieved by employing two parameters: (i) diminish the dimensions of the capacitor array and (ii) by employing varying voltage levels. The use of CDAC is deemed appropriate for the low-power SAR ADC. The diverse configurations of CDAC provide the capability to regulate the quantity of capacitors and facilitate a reduction in the area occupied by capacitors. The charge-sharing voltage switching strategy utilizing two arrays (most significant bit and least significant bit) in the CDAC structure has successfully met the desired specifications in the low-power SAR ADC designed for biomedical applications. The proposed digital-to-analog converter (DAC) incorporates an extra common-mode switch that is interposed among V_{XP} as well as V_{XN}. This

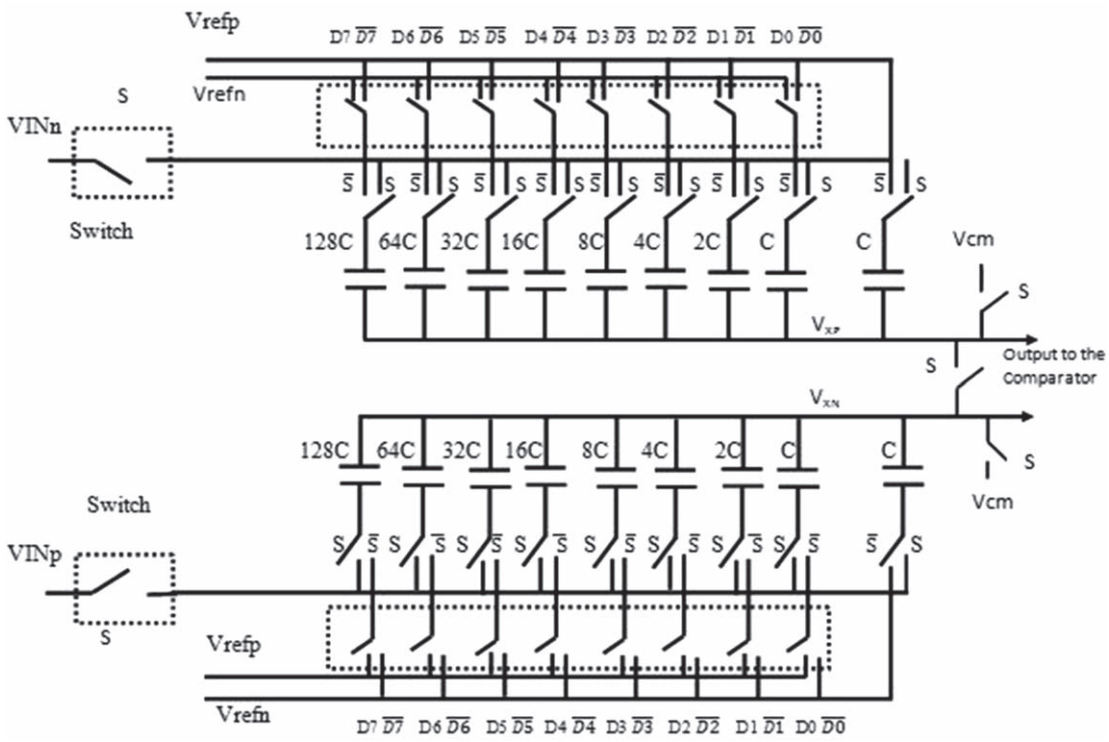


Figure 5. Architecture of proposed 8-bit capacitive DAC.

modification is intended to enhance the speed and efficacy of the connection between these components, as well as their connection to V_{cm} throughout the sampling phase with one another.

During the initial phase of a conversion process, the signal from the input V_{IN} has subjected to sampling by means of the S switches, whereby the bottom capacitive plates of each capacitor are utilised. Meanwhile, the upper plate, also known as node $\mathbf{V_X}$, has a connection to V_{cm}. The aggregate charge accumulated on the capacitors has become equivalent to the product of V_{IN} and the summation of the capacitors that are available. Equations (4), (5) depict the voltage along with charge in the node V_X, respectively. In the sampling stage,

$$V_x = V_{cm} = 0 \tag{5}$$

$$Q_x = -V_{IN}.8C \tag{6}$$

Following the sampling stage, the upper plates of the capacitors become detached from V_{CM}, whereas the lower plate is disengaged from V_{IN} as well as is modified based on the digital code emanating through the SAR logic. At first, the digital-to-analog converter (DAC) code has been set to 10000000 by setting the most significant bit (MSB) for 1 within the very first comparison.

Equation (8) illustrates the variation of V_X through the processes of distribution of charges as well as voltage division, in comparison to V_{CM} (which is assumed to be zero in this instance). This is obtained as per the comparison of V_{IN} with half of Vref. The comparator produces an output of 1 when the value of V_{IN} is greater than half of Vref and V_X is less than 0. Conversely, when this condition is not met, the comparator outcomes become 0.

$$\text{For DAC} = 10000000,\ \Delta v = \frac{4c}{4c + 2c + c + c} \tag{7}$$

$$V_x = -V_{IN} + \Delta v = -V_{IN} + \frac{V_{ref}}{2} \tag{8}$$

In the subsequent instance, the most significant bit (MSB) is going to assume the amount of the output D2 from the comparator, while the second most significant bit (MSB-1) has been assigned an integer of 1. The comparison of V_{IN} is presently conducted regardless of the top half of Vref within the event that D2 equals 1, or in the bottom half of Vref whenever D2 equals 0. Following the evaluation, the MSB-1 bit assumes the current value of a comparator D1, while the MSB-7 bit is assigned a value of 1. Upon conducting N = 8 comparisons, the successive approximation register analog-to-digital converter produces the binary code D7D6D5D4D3D2D1D0 as the ultimate output, which corresponds to the voltage level VDAC as specified . Figure 6 illustrates the

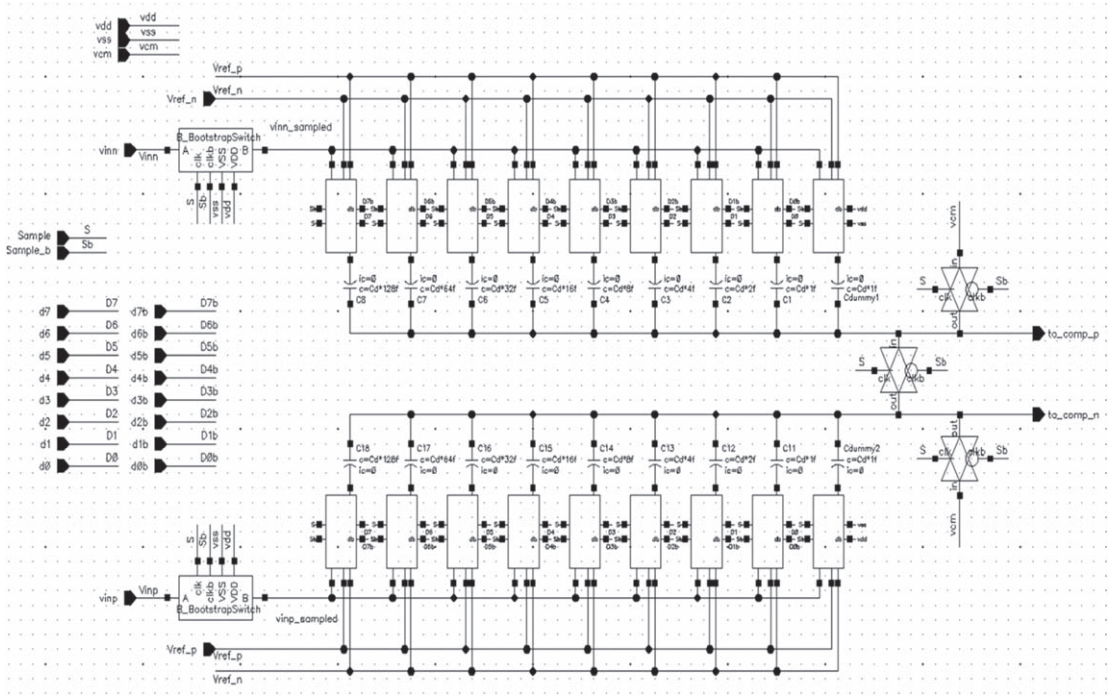


Figure 6. CMOS Implementation of 8-bit DAC.

alterations in the node voltages that affect the capacitors during each phase of the conversion procedure through the CMOS implementation of 8-bit DAC.

3.3. Comparator

Regarding the design of low power comparator that can mitigate kickback noise and offset voltage for better linearity of the device, A comparative analysis has been conducted in the literature to evaluate several architectures of dynamic latch comparators with respect to power consumption, kick-back noise, speed, resolution, and input-referred offset voltage. The single-stage dynamic latch comparator is more efficient in terms of power consumption compared to other configurations, since it utilizes the fewest number of transistors in the circuit. Nevertheless, it is affected by several drawbacks.

The performance of single-stage dynamic latch comparators is characterized by their comparatively low operating speed, as well as their significant offset voltage and least significant bit (LSB) values. Nevertheless, for biomedical applications, the required sampling rate and resolution typically fall within the low-to-medium range. This characteristic enables the utilization of single-stage dynamic latch comparators in the biomedical implants. Additionally, a commonly employed approach involves the utilization of an amplifier stage in conjunction with the dynamic comparator in order to enhance the offset performance. Various types of comparators, such as tail transistor latched comparator, two-stage dynamic latch comparator and the PMOS latched comparator, have been employed in order to enhance the offset voltage at the expense of power consumption. A comparative analysis is presented in table 2.

Figure 7 depicts the design of the suggested comparator with the kickback noise reduction approach. The suggested dynamic comparator operates in two modes: comparison/evaluation mode and reset mode. Once the clock (clk=‘0’) is zero during the reset or precharge phase, the kickback noise reduction approach reliant transistors M7-M8, M12 are turned off to prevent the circuitry from using static power. During this process, the transistors M3, M4 are turned ON, resulting in a short circuit between V_{DD} and the comparator’s output nodes. As a consequence, V_{DD} is coupled to the Comparator outputs outp and outn. When the kickback noise reduction technique transistors M7 and M8 are switched off, the comparator’s output terminals are no longer connected to the comparator’s inputs. The latch outputs have no effect on the transistors M9 and M10, which are linked to input signals. During the evaluation phase, the clock signal is set to V_{DD} (clk $=V_{DD}$), the tail transistor M11 and the M3, M6 PMOS transistors are turned off, and the Kickback noise mitigation method transistors are turned on. When the transistors M7-M8 connected to the kickback noise reduction approach are turned on, the differential voltage created at the inputs is fed into cross-coupled inverter-based latches, which then make the needed choice. If the input voltage fed to inn is greater than the input voltage fed to inp, the

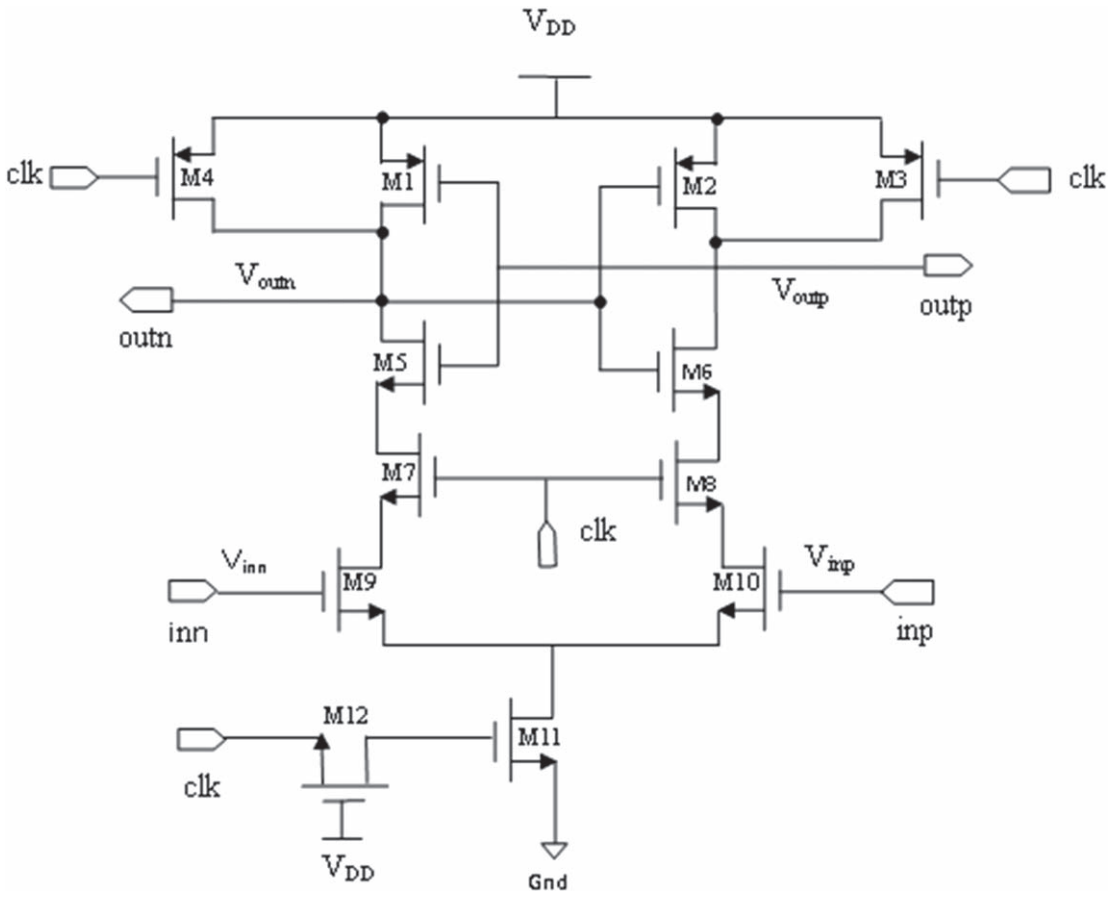


Figure 7. Design of proposed comparator.

Table 2. Comparative analysis of various configurations of comparator.

Parameter/Configuration of Dynamic Latch Comparator	Single Stage	Two Stage	Three Stage	Multiple Stage
Power Consumption	Very Low	Medium	High	Very High
Transistors Count	Low	Medium	High	Very High
Kickback Noise	High	Very Low	Medium	Very Low
Area	Less	Medium	High	Very High
Offset Voltage	High	Medium	Low	Low

corresponding voltage in the output terminal outp falls quicker than the voltage in the output node outn, and vice versa.

When compared to other terminals, the voltage of the output terminal Outp approaches $|V_{DD}\text{-}Vth|$ relatively quickly. The regeneration of inverter latches (M1-M5 & M2-M6) activates the associated PMOS transistor M1. As a consequence, the output Outn becomes V_{DD}. As a result of this PMOS transistor M2, the output outp equals 0. The primary advantage of the proposed comparator architecture is that only one regulating clock signal is needed. This eliminates the requirement for synchronization between the multiple control signals and increases the circuit's efficiency. Moreover, the proposed circuits provide good delay and power reduction. The tail transistor is small in size to give the required lower tail current for the differential input transistor stage. Since extra transistors are not required for configuring or resetting the circuitry, the circuit's complexity is reduced.

The delay specifies the discharging capacitance of load capacitance until the latch's PMOS transistor goes on. The discharge delay t_o may be expressed as

$$t_0 = \frac{C_L\ |V_{thp}|}{I_{\text{current in the INP transistor}}} \cong 2\frac{C_L\ |V_{thp}|}{I_{\text{current present in tail transistor}}} \tag{9}$$

$$\text{In(2), } I_{\text{current at the INP transistor}} = \frac{I_{\text{current present in the tail transistor}}}{2} + \Delta I_{in} \tag{10}$$

$$= \frac{I_{\text{current in tail transistor}}}{2} + g_{m(\text{input transistors})}\,\Delta I_{in} \tag{11}$$

Where ΔI_{in}—changes in the differential input voltage.

The t_{latch} delay has been generated by the latch delay of two cross coupled inverters as

$$t_{latch} = \frac{C_L}{g_{m,eff}}.\ln\left(\frac{\Delta V_{out}}{\Delta V_0}\right) \tag{12}$$

$$= \frac{C_L}{g_{m,eff}}.\ln\left(\frac{V_{dd}}{\Delta V_0}\right) \tag{13}$$

Where, $g_{m,eff}$—transconductance of the latch.

In dynamic Comparator, ΔV_0 is written as

$$\Delta V_0 = |V_{outp}(t = t_0) - V_{outn}(t = t_0)| \tag{14}$$

$$= |V_{thp}| - \frac{I_{INP} t_0}{C_L} = |V_{thp}|\left(1 - \frac{I_{INP}}{I_{INN}}\right) \tag{15}$$

where, I_{INP}- current flowing in the transistor having input voltage of INP, I_{INN} - current flowing in the transistor having input voltage of INP.

Thus, INN will be $\frac{I_{\text{current in tail transistor}}}{2}$ and delay ΔV_0 is given by:

$$\Delta V_0 = |V_{thp}|\,\frac{\Delta I_{in}}{I_{INN}} \tag{16}$$

$$\approx 2\,|V_{thp}|\,\frac{\Delta I_{in}}{I_{tail}} \tag{17}$$

$$= \frac{\sqrt{\beta_{\text{inputtransistors}} I_{tail}}}{I_{tail}} \Delta V_{in} \tag{18}$$

Where ,$\beta_{\text{inputtransistors}}$- current factor of input transistors, I_{tail} - tail transistor current. The total delay will be as follows:

$$td = 2\frac{C_L\,|V_{thp}|}{I_{tail}} + \frac{C_L}{g_{m,eff}}.\ln\left(\frac{V_{DD}}{2\,|V_{thp}|\,\Delta V_{in}}\sqrt{\frac{I_{tail}}{\beta_{1,2}}}\right) \tag{19}$$

According to this research, the recommended comparator's delay duration is proportional to the comparator's load capacitance C_L and inversely proportional to the differential voltage between the two inputs Vin.

For maximum power savings, the comparator should be constructed as simply as feasible, using as few transistors as possible. The comparator illustrated in figure 6 has been utilised in this study. Most of the transistors utilised in this comparator function within the weak inversion region, while being supplied with a voltage of 1 V. A preliminary examination of comparator offset throughout the weak-inversion zone will be given below, as no previous research has yet been conducted on this topic. The analysis is conducted as 'CLK' addresses VDD, while the circuit maintains a state of equilibrium. The balanced steady-state condition is characterised by the equivalence of currents in both branches. M1, 2, 5, 6, 7, 8, 9 are all in the weak-inversion zone at this moment. The functional correlation between the current flowing through the drain of a transistor as well as the gate-to-source voltage within the weak-inversion region [21] is established by:

$$I_D = \frac{W}{L} I_{D0}\,.\,e^{\frac{(V_{GS}-V_{th})}{n\ V_T}}\left[1 - e^{-\frac{V_{DS}}{V_T}}\right] \tag{20}$$

$$\approx \frac{W}{L} I_{D0}\,.e^{\frac{(V_{GS}-V_{th})}{n\ V_T}} (V_{DS} \gg V_T) \tag{21}$$

The disparity within M3 as well as M4 does not impact the offset of the comparator as the comparing process entails the deactivation of these two MOSFETs. The comparator offset remains unaffected by M11 and M12 as they are not apparent in the differential branches.

Random changes i.e., Gaussian distribution of process parameters (for transistors may be modelled as

$$\frac{\sigma_W}{W} = \frac{A_W}{\sqrt{WL}}\%,\ \frac{\sigma_L}{L} = \frac{A_L}{\sqrt{WL}}\%,\ \sigma_{Vth} = \frac{A_{Vth}}{\sqrt{WL}} mV \tag{22}$$

As the main factor contributing to the comparator's static offset, the disparity in threshold voltage has been given precedence over the mismatch within geometrical dimensions in the ensuing derivation process for the sake of its simplicity. Therefore, only the imbalance in threshold voltage will be taken into account.

$$V_{th9} = \Delta_{Vth9} + V_{th9,10} \tag{23}$$

$$V_{th10} = \Delta_{Vth10} + V_{th9,10} \tag{24}$$

$$\Delta_{Vth9,10} = V_{th10} - Vt_{h9} \tag{25}$$

Substitute equations (15) to (17) in equation (9), the offset current through M9 and M10 is derived as

$$I_9 = \left(\frac{W_{9,10}}{L_{9,10}}\right) I_t e^{(V_N + \Delta V_{9,10} - V_{D0} + V_{th9})/nVt} \tag{26}$$

$$I_{10} = \left(\frac{W_{9,10}}{L_{9,10}}\right) I_t e^{(V_P - V_{D0} + V_{th2})/nVt} \tag{27}$$

$\Delta V_{9,10}$ is a compensation voltage that ensures the comparator works in a balanced state and is M10's drain voltage. In a well-balanced state,

$$\Delta V_{9,10} = \Delta Vth_{9,10} \tag{28}$$

Consequently, the random offset within the comparator caused by the discrepancy between M9 and M10 can be computed as follows:

$$\sigma^2_{\Delta V_{9,10}} = \frac{A^2_{VthN}}{W9,\ 10L9,\ 10} \tag{29}$$

Assume mismatch between M3 and M4 is like that between M1 and M2, the current difference flowing across M3 and M4 can be expressed as:

$$\Delta I_{5,6} = \left(\frac{W_{5,6}}{L_{5,6}}\right) I_t e^{\frac{V_{GS56}}{nVt}[1 - e^{\frac{\Delta Vth_{5,6}}{nVt}}]} \tag{30}$$

The equation represented by (30) can be formulated utilizing the initial order approximation of the Taylor's expansion.

$$\Delta I_{5,6} = \left(\frac{W_{5,6}}{L_{5,6}}\right) I_t e^{\frac{V_{GS56}}{nVt}[\frac{\Delta Vth_{5,6}}{nVt}]} \tag{31}$$

The transconductance for input-transistors M1 as well as M2 has been combined as:

$$g_{m1,2} = \frac{I_{1,2}}{nVt} \tag{32}$$

Then, the input-referred comparator offset caused by the mismatch between M5 and M6 may be calculated.

$$\Delta V_{5,6} = K_{Vth5,6}\,\Delta Vth_{5,6} \tag{33}$$

Thus, the random offset resulting from the discrepancy between M5 as well as M6 has become:

$$\sigma^2_{\Delta V_{5,6}} = K^2_{Vth5,6}\sigma^2_{Vth_{5,6}} = K^2_{Vth5,6}\frac{A^2_{Vth\ \ N}}{W5,\ 6L5,\ 6} \tag{34}$$

Likewise, there exists a random offset resulting from the mismatch in M1 and M2 transistors.

$$\sigma^2_{\Delta V_{1,2}} = K^2_{Vth1.2}\sigma^2_{Vth_{1,2}} = K^2_{Vth1,2}\frac{A^2_{Vth\ \ N}}{W1,\ 2L1,\ 2} \tag{35}$$

The total offset voltage generated by comparator:

$$\sigma^2_{OS} = \sigma^2_{\Delta V_{1,2}} + \sigma^2_{\Delta V_{5,6}} + \sigma^2_{\Delta V_{9,10}} \tag{36}$$

Equation (36) signifies that the mean square value of total offset voltage in the proposed comparator will be depending upon the random offset resulting from the discrepancy between the transistors pair M1, M2; M5, M6 and M9, M10.

3.4. SAR digital control logic

Asynchronous control logic has traditionally been utilised to attain high speed, as referenced in sources [25] and [26]. Regrettably, the comparison circles generated by the propagation of internal logical signals exhibit a significant reliance on variations in process, voltage, and temperature (PVT). The utilisation of synchronous

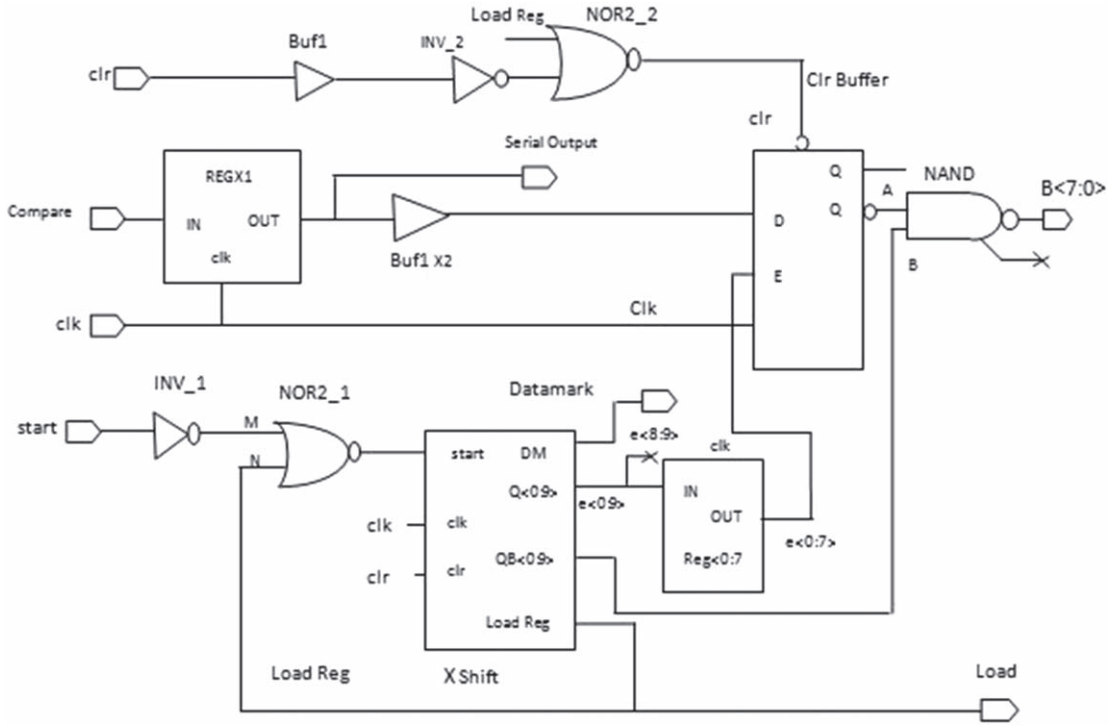


Figure 8. Block diagram of SAR digital control logic.

control system logic is illustrated in figure 8 of this study. The present control logic operates synchronously to produce output codes that encompass switching control signals for the digital-to-analog converter (DAC). The control signals are generated based on the comparison outcome, clk, and sample, which serve as the inputs to the system. The SAR logic has been developed to provide digital code to DAC depending on the output of the comparator. It is made up of digital cells that execute a binary search algorithm that produces the ADC output alongside the End of Conversion (EOC) message at the conclusion of the process. Figure 5 shows an example of SAR logic control. Shift registers, as well as a combinational switch, are the primary components. The latch serves to record the comparator's decisions, the combinational switch serves as a source of the control signals needed for the differentiating DAC [47], and the shift register is employed to generate pulses. In this study, we need n clock cycles to conduct a single conversion.

4. Simulation results

An 8-bit low power successive approximation register (SAR) analog-to-digital converter (ADC) has been developed for Bio-Medical Implants applications using 45 nm CMOS technology. The design incorporates a capacitive digital-to-analog converter (CDAC) and an optimized comparator, based on the proposed SAR logic. The current investigation was conducted utilizing CMOS 45 nm technology and utilizing Cadence tools.

Table 3 shows the measured DNL for the proposed DAC that confirms a mean differential non-linearity for the proposed design is of 0.032 V also there is average deviation of 1.87 V in the DNL value. Figure 9 denotes the Monte Carlo Analysis of DAC output for 250 samples. The precision of Virtuoso's simulation outcomes is contingent upon the calibre and precision of the utilized models.

Based on the outcomes and mathematical expressions presented in equations (9) through (24), it is possible to derive the following deductions: Enhancing the sizes can effectively decrease the comparator's static offset resulting from M9 and M10, as the variance of $\sigma^2_{\Delta V_{9,10}}$ is inversely related to their product and does not depend on the common mode voltage.

The other values of offset at M5,6 and M1,2 are dependent on common mode voltage as they have Vgs voltage in their equations. The offset voltages determined by carious transistors are depicted in figure 10. The architecture of SAR ADC is depicted in the figure 11. The SAR ADC consists of DAC, comparator, and SAR Approximate Register.

The 1024-point FFT spectrum at 20.8 KS/s corresponding to the Nyquist input frequency can also be seen in figure 12. The measured Signal-to-Noise Distortion Ratio (SNDR) and spurious-Free Dynamic Range

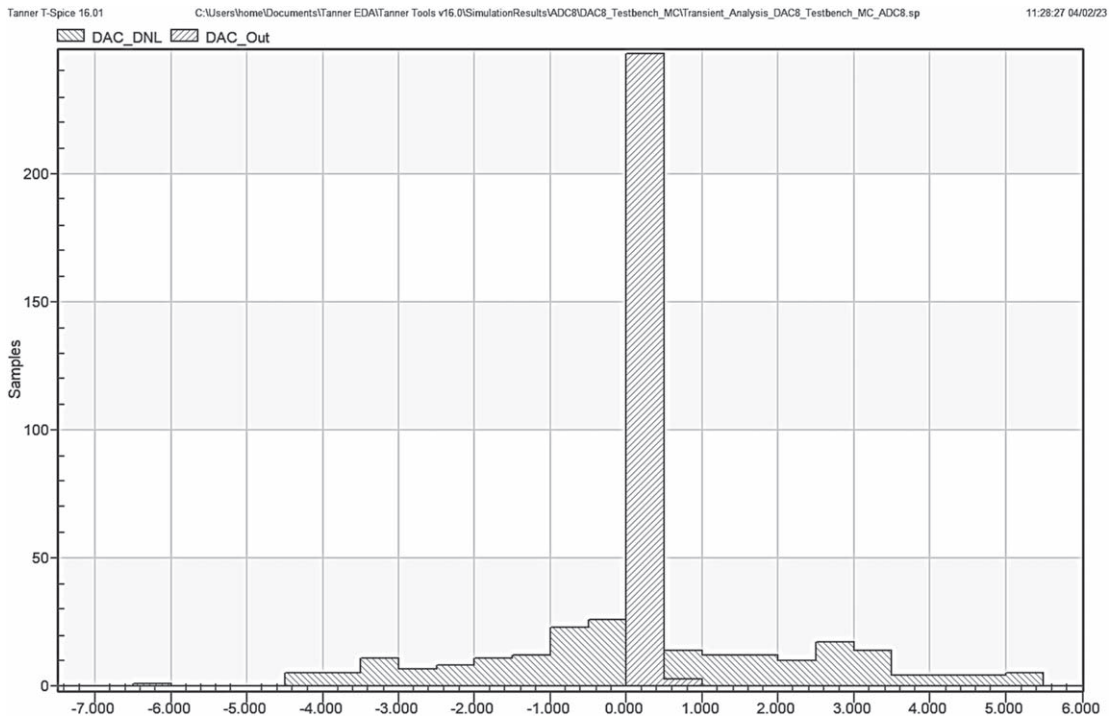


Figure 9. Monte Carlo analysis of DAC output for 250 samples.

Table 3. DNL for DAC.

Statistical parameter	DAC_Out	DAC_DNL
Minimum	273.90 m	−5.6
Maximum	413.5 m	4.3
Mean	358.30 m	32.2 m
Average deviation	17.7 m	1.87
Variance	434.90 μ	4.6
Standard deviation	33.7 m	2.2

(SFDR) are 59.16 dB as well as 64.02 dB, correspondingly. The resultant value for the effective number of bits (ENOB) is 9.53 bits. Figures 13 and 14 depicts the transient simulation results of 8-bit DAC & SAR-ADC respectively. Table 2 also tabulates and compares several performance metrics of the SAR ADC with previous studies. The suggested ADC is clearly more efficient in terms of power usage and other metrics. The proposed ADC consumes the least amount of power 1.9 μW, resulting in favourable outcomes. Our investigation demonstrates the efficacy of the suggested approach, which can boost SNDR even more. The results demonstrate that the proposed ADC technique is promising for increasing the SAR ADC's performance. Although the suggested ADC consumes less power at a medium conversation rate, it is appropriate for bio-medical applications. Implementing a more modern CMOS technology can improve performance even further. Because detecting analogue signals is required for many Bio-implantable devices, the suggested ADC has numerous application possibilities. To guarantee a long battery life, the sensor interface in these applications must consume relatively little power. The arrangement of the complete ADC is represented in figure 11. Among the most crucial phases in the full-custom design flow is layout generation. The designers specify the whole structure as well as the relative arrangement of all layers that will be employed in production. In the design Layout Editor has been used to create the layout.

Since this physical structure determines parasitic capacitances as well as resistances, as well as chip area, physical design and layout is extremely strongly connected to total device performance. Then, a thorough arrangement of CMOS logic requires a significant as well as time-consuming design process. A CMOS logic gate layout design begins with circuitry and early CMOS transistor size.

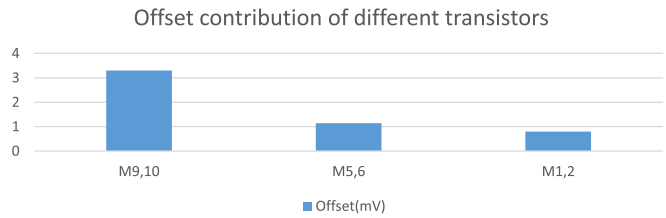


Figure 10. Offset voltages contributed by used transistors.

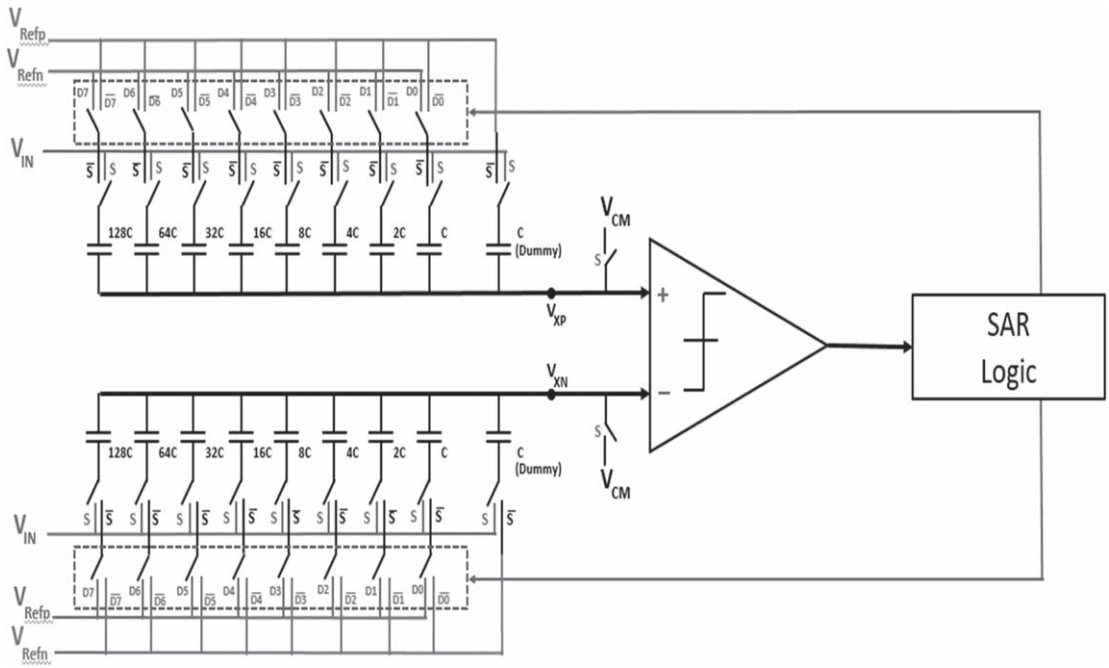


Figure 11. Total Architectural implementation of 8-bit SAR ADC.

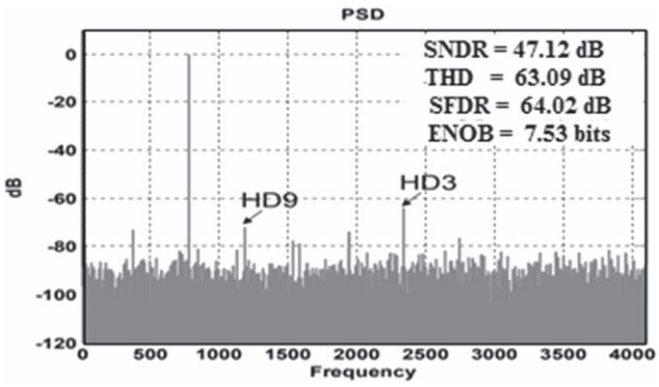


Figure 12. Power Spectral Density function of SAR ADC.

Proposed methodology utilised in this work is reduction in leakage power in comparator by keeping same potential at multiple nodes when clock is in OFF state, also clock feedthrough technique is utilised to reduce kickback noise for improvement of linearity. Further, reduction in energy consumption per conversion cycle in capacitive digital to analog converter (CDAC) is achieved with the help of variable reference voltage between vrefn and vrefp along with the bottom plate sampling for conversion of MSB bits with minimized energy consumption.

Further, two techniques have been employed in the DAC design: (i) Diminishing the dimensions of the capacitor array and (ii) Employing varying voltage levels. The use of CDAC is deemed appropriate for the

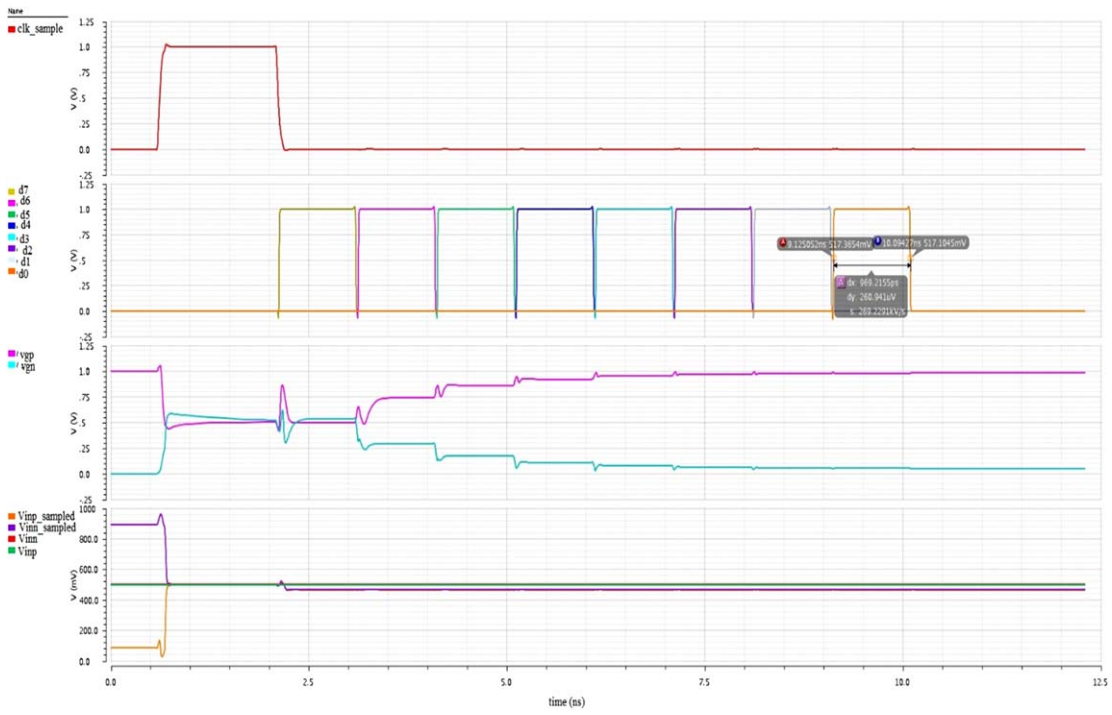


Figure 13. Transient result of DAC.

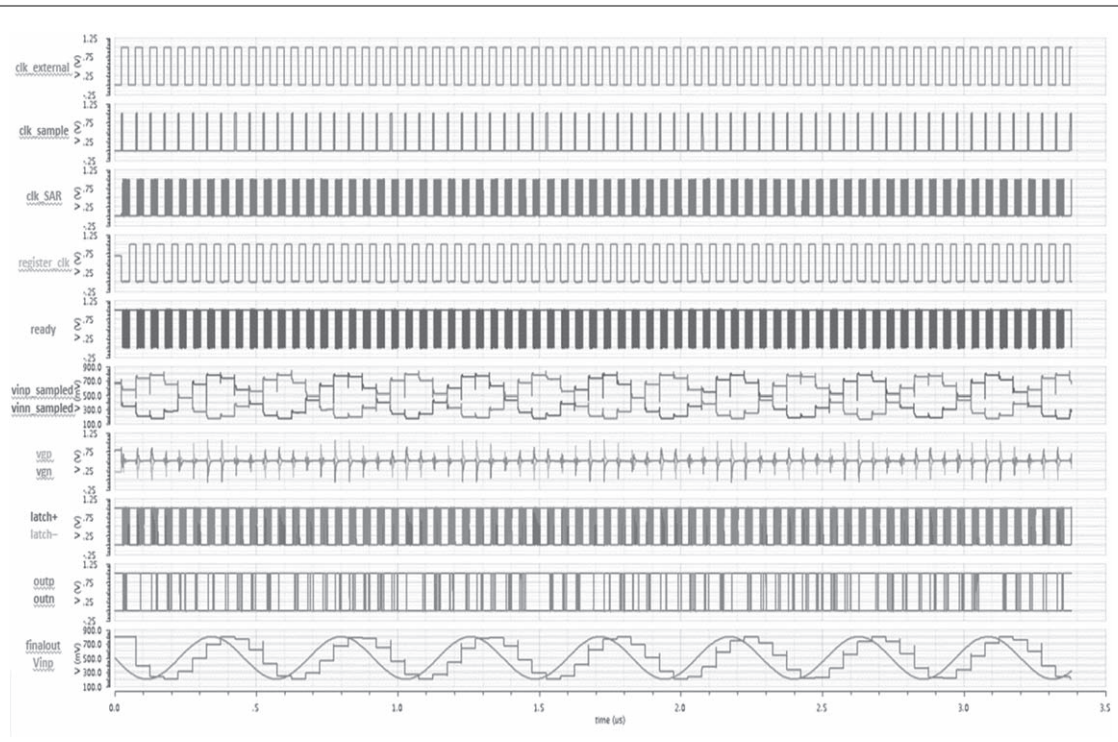


Figure 14. Transient result of SAR-ADC.

low-power SAR ADC. Nevertheless, several voltage switching techniques have the potential to enhance the efficiency of power usage during switching operations. As a result, there is an improvement in energy conservation. By employing the mentioned considerations in the design, design has achieved a good SNDR value of Spurious-Free Dynamic Range (SFDR) of 64.02 dB with power consumption of 1.987 μW at a sampling rate of 1.1 MHz and a supply voltage of 1 V as presented in table 4. The layout of proposed architecture is given in figure 15.

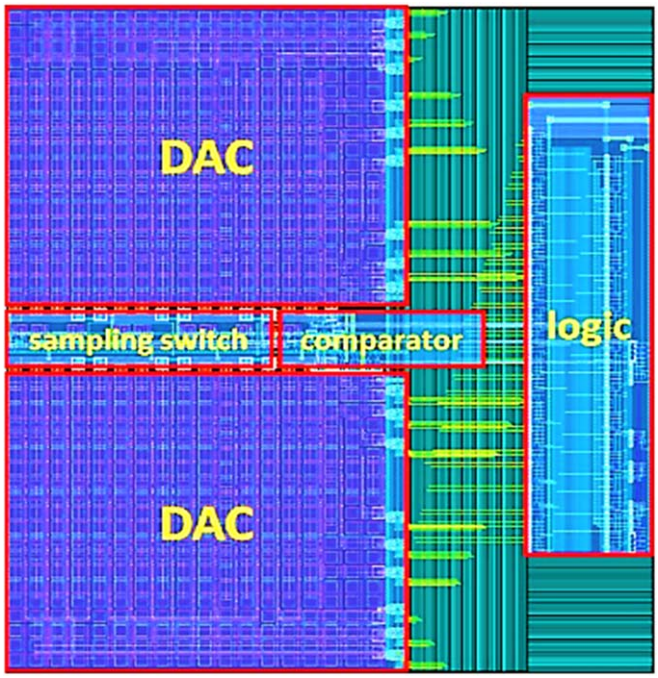


Figure 15. Layout of entire ADC architecture.

Table 4. Comparison performance of ADC.

PARAMETERS	[40]	[41]	[42]	[43]	[44]	[45]	[46]	[17]	This Work
Technology (μm)	0.18	0.18	0.18	0.13	0.090	0.065	0.055	0.040	0.045
Supply voltage (V)	3.3	1.2	0.5	1.2	1.2	1.2	0.9	1.1	1
Sampling rate (S/s)	150 K	20 M	1 M	1 M	50 M	1 M	10 M	100 M	1.1 M
Resolution (dB)	14	5	10	8	10	8	12	10	8
ENOB (bits)	11.3	4.74	7.69	7.51	9.26	7.9	10.03	9.3	7.53
SNDR (dB)	70.28	30.3	61.96	47.12	57.6	49.3	62.13	58.03	59.16
SFDR (dB)	82.91	40	68.54	57.36	65.8	61	73.52	72.6	64.02
Power (μW)	477.2 μ	76.88 μ	13.99 μ	11.56 μ	0.664 m	21 μ	40.2 μ	1.32 m	1.987

5. Conclusion

The adoption of C-2C DAC results in the suggested switching system in this work being energy-efficient and reducing switching energy. The capability exists for the SAR ADC to provide outstanding linearity. The present analysis concerns the offset characteristic of a comparator that functions within the weak-inversion regime. The Monte-Carlo simulation has been utilised to validate the optimization of the comparator. The comparator utilised in this study was designed without the implementation of an offset cancellation technique, as indicated by the analysis. The present investigation introduces a novel switching scheme as well as comparator for the development of an 8-bit SAR ADC with ultra-low power consumption, specifically designed for application in medical implantable devices. The Analog-to-Digital Converter (ADC) was realized through the utilisation of the 45 nm Complementary Metal-Oxide-Semiconductor (CMOS) fabrication process. At the frequency of input as stipulated by the Nyquist criterion, the Analog-to-Digital Converter (ADC) achieves an Effective Number of Bits (ENOB) measuring 7.53 and consumes 1.9 μW of power while operating at a sampling rate of 1.1 MS/s. This performance level indicates that the ADC has favourable energy efficiency.

Data availability statement

The data cannot be made publicly available upon publication because no suitable repository exists for hosting data in this field of study. The data that support the findings of this study are available upon reasonable request from the authors.

ORCID iDs

Poornima Mittal https://orcid.org/0000-0002-9479-8628

References

[1] Jeon D *et al* 2014 An implantable 64nW ECG-monitoring mixed-signal SoC for arrhythmia diagnosis *IEEE Int. Solid State Circuits Conf.* pp 416–7

[2] Yazicioglu R F, Kim S, Torfs T, Kim H and van Hoof C 2011 A 30 μW analog signal processor ASIC for portable biopotential signal monitoring *IEEE J. Solid-State Circuits* **46** 209–23

[3] Yang T, Lu J, Jahan M S, Griffin K, Langford J and Holleman J 2015 A configurable 5.9 μW analog front-end for bio signal acquisition *Proc. of the 2015 IEEE Custom Integrated Circuits Conf. (CICC), San Jose, CA, USA* pp 1–4

[4] Sant L, Fant A, Stojanović S, Fabbro S and Ceballos J L 2016 A 13.2b optical proximity sensor system with 130 klx ambient light rejection capable of heart rate and blood oximetry monitoring *IEEE J. Solid-State Circuits* **51** 1674–83

[5] Lin C H and Wen K A 2021 An innovative successive approximation register analog-to-digital converter for a nine-axis sensing system *Journal of Low Power Electronics and Applications* **11** 1–24

[6] Huang Y, Meng Q, Li F and Wu J 2021 A high-speed low power SAR ADC in 40 nm CMOS with combined energy efficient techniques *IEICE Electron. Express* **18** 20210156

[7] Naguib A 2020 High speed and low power comparator in 65 nm CMOS for energy efficient biomedical SAR ADCs *Proc. of the 2020 12th Int. Conf. on Electrical Engineering (ICEENG)* pp 339–42 Cairo

[8] Wu Y, Shen J, Liang J and Yao M 2020 A design method of capacitor arrays for high-resolution SAR ADCs *Circuit World* **46** 249–55

[9] Jajodia B, Mahanta A and Ahamed S R 2020 Energy-efficient DAC switching technique for single-ended SAR ADCs *AEU-International Journal of Electronics and Communications* **124** 153334

[10] Song Z, Ma Z H, Yao Y, Zou R F and Hu J G 2023 *A ENOB 10.3 bit 103 μW ON Chip Calibration SAR ADC 2nd Int. Conf. on Electronics & Integrated Circuits Technology (EICT)* Journal of Physics: Conf. series, serial no. 2524012019

[11] Hsieh S-E and Hsieh C C 2018 A 0.44-fJ/Conversion-step 11-Bit 600-kS/s SAR ADC with semi-resting DAC *IEEE J. Solid-State Circuits* **99** 1–9

[12] Liu M, Pelzer's K, Dommele R V, Roermund A V and Harpe P 2016 A 106nW 10 b 80 KS/s SAR ADC with duty-cycled reference generation in 65 nm CMOS *IEEE J. Solid-State Circuits* **51** pp 2435–45

[13] Tang X *et al* 2022 Low-power SAR ADC design: Overview and survey of state-of-the-art techniques *IEEE Transactions on Circuits and Systems - I :Regular Papers* **69** 2249–62

[14] Kelly B M, DiLello A T and Graham D W 2019 Reconfigurable analog preprocessing for efficient asynchronous analog-to-digital conversion *Journal of Low Power Electronics and Applications* **9** 25–18

[15] Cao S, Wang C, Zhang L, Luo M, Yan W and Gong Y 2020 A fully integrated 10-bit 100 MS/s SAR ADC with metastability elimination for the high energy physics experiments *Nuclear Instruments and Methods in Physics Research section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **978** 164415

[16] Chen Y, Zhuang Y and Tang H 2020 An ultra-low power consumption high-linearity switching scheme for SAR ADC *J. Circuits Syst. Comput.* **29** 2050086

[17] Konijnenburg M *et al* 2016 28.4 A battery-powered efficient multi-sensor acquisition system with simultaneous ECG, BIO-Z,GSR, and PPG *IEEE International Solid State Circuits Conference (ISSCC)* San Francisco USA pp 480–481

[18] Kim J and Ko H 2016 Reconfigurable multiparameter biosignal acquisition SoC for low power wearable platform, *Sensors* **16** 27898004

[19] Liu J, Liu S, Ding R and Zhu Z 2021 A conversion mode reconfigurable SAR ADC for multistandard systems *IEEE Trans. Very Large Scale Integr. VLSI Syst.* **29** 895–903

[20] Liu C C, Chang S J, Huang G Y and Lin Y Z 2010 A 10-bit 50 MS/s SAR ADC With a Monotonic Capacitor Switching Procedure *IEEE J. Solid-State Circuits* **45** pp 731–40

[21] Jo S H, Cho H W and Yoo H J 2019 A fully reconfigurable universal sensor analog front-end IC for the internet of things era *IEEE Sens. J.* **19** 2890211

[22] Rezaeiyan Y, Zamani M, Shoaei O and Serdijn W A 2018 Mixed-signal IC with pulse width modulation wireless telemetry for implantable cardiac pacemakers in 0.18-μm CMOS *IEEE Trans. Biomed. Circuits Syst.* **12** 589–600

[23] Seo M J, Roh Y J, Chang D J, Kim W, Kim Y D and Ryu S T 2018 A reusable code-based SAR ADC design with CDAC compiler and synthesizable analog building blocks *IEEE Transactions on circuits and systems II: Express Briefs* **65** 1904–8

[24] Seong K, Han J S, Shim Y and Baek K H 2022 A 2.5 GS/s 7-Bit 5-way time-interleaved SAR ADC with on-chip background offset and timing-skew calibration *IEEE Trans. Circuits Syst. Express Briefs* **69** 4043–7

[25] Lee J, Lim Y, Lee J, Jang T, Kang K, Cho J, Oh S and Lee J 2022 A 0.56-mW 63.6-dB SNDR 250-MS/s Two-Step SAR ADC in 8-nm FinFET *IEEE Symp. on VLSI Technology and Circuits (VLSI Technology and Circuits 2022)* Honolulu, HI, USA, June 12-17

[26] Konwar S, Roy H, Chiang S H W and Sahoo B D 2022 Deterministic dithering-based 12-b 8-MS/s SAR ADC in 0.18-μm CMOS *IEEE Solid-State Circuits Letters* **5** 243–6

[27] Sansen W 2015 Minimum Power in Analog Amplifying Blocks: Presenting a Design Procedure *IEEE Solid-State Circuits Magazine* **7** 83–9

[28] Rawat B and Mittal P 2023 A switching NMOS based single ended sense amplifier for high density SRAM applications *ACM Trans. Des. Autom. Electron. Syst.* **28** 1–14

[29] Morie T, Miki T, Matsukawa K, Bando Y, Okumoto T, Obata K, Sakiyama S and Dosho S 2013 *A 71 dB-SNDR 50 MS/s 4.2 mW CMOS SAR ADC by SNR enhancement techniques utilizing noise, IEEE Int. solid-state Circuits Conf.* pp 272–3

[30] Monika P M 2021 A novel modified current comparator based on extremely low voltage high compliance current mirror, *Int. J. Inf. Technol.* **14** 323–31

[31] Harpe P, Cantatore E and Roermund A V 2014 An oversampled 12/14b SAR ADC with noise reduction and linearity enhancements achieving up to 79.1 dB SNDR *in Proc. ISSCC Dig. Tech. Papers* pp 194–5

[32] Kapusta R, Shen J, Decker S, Li H, Ibaragi E and Zhu H 2013 A 14b 80 MS/s SAR ADC with 73.6 dB SNDR in 65 nm CMOS *IEEE J. Solid-State Circuits* **48** 3059–66

[33] Kramer M, Janssen E, Doris K and Murmann B 2015 A 14b 35 MS/s SAR ADC achieving 75 dB SNDR and 99 dB SFDR with loop-Embedded Input Buffer in 40 nm CMOS, *IEEE J. Solid-State Circuits* **50** 1–10

[34] Xu J, Mitra S, Van Hoof C, Yazicioglu R F and Makinwa K , A A 2017 Active electrodes for wearable EEG acquisition: review and electronics design methodology *IEEE Rev. Biomed. Eng.* **10** 187–98
[35] Lee T, Kim M K, Lee H J and Je M 2022 A multimodal neural-recording IC with reconfigurable analog front-ends for improved availability and usability for recording channels *IEEE Trans. Biomed. Circuits Syst.* **16** 185–99
[36] Wang T Y, Li H Y, Ma Z Y, Huang Y J and Peng S Y 2018 A bypass-switching SAR ADC with a dynamic proximity comparator for biomedical applications *IEEE J. Solid-State Circuits* **53** 2819164
[37] Kumar N and Mittal P 2021 Dynamic power consumption and delay analysis for ultra-low power 2 to 1 multiplexer designs *Advances in Electrical and Electronic Engineering* **19** 145–54
[38] Li Z, Dutta A, Mukherjee A, Tang X, Shen L, He L and Sun N 2020 A SAR ADC with reduced kT/C noise by decoupling noise PSD and BW *IEEE Symp. on Circuits*
[39] Yang C *et al* 2021 An area-efficient SAR ADC with mismatch error shaping technique achieving 102-dB SFDR 90.2-dB SNDR Over 20-kHz bandwidth *IEEE Transactions on Very Large-Scale Integration (VLSI) Systems* **29** 1575–85
[40] Zhang X, Fan W, Xi J and He L 2023 14-Bit fully differential SAR ADC with PGA used in readout circuit of CMOS image sensor *Journal of Sensors* **2021** 6651642
[41] Aminzadeh H 2019 A low-cost tiny-size successive approximation ADC for applications requiring low-resolution conversion with moderate sampling rate *Circuits Syst. Signal Process.* **38** 242–58
[42] Mandrumaka K K and Noorbasha F 2019 A low power 10-bit SAR ADC with variable threshold technique for biomedical applications *SN Applied Sciences* **1** 918
[43] Sotoudeh M and Rezaei F 2020 A four-level switching scheme for SAR ADCs with 87.5% area saving and 97.85% energy reduction *Circuits Syst. Signal Process.* **39** 4792–809
[44] Lu C C and Huang D K 2020 A 10-bits 50-MS/s SAR ADC based on area-efficient and low-energy switching scheme *IEEE Access* **8** 28257–66
[45] Fahmy G A and Zorkany M 2021 Design of a memristor-based digital to analog converter (DAC), *Electronics* **10** 1850230–43
[46] Hong H K, Kang H W and Kim C H 2020 A 62dB-SNDR 40.2 μW 10MS/s ADC for power efficient IoT and biomedical read-out systems *Proc. of the 2020 IEEE Int. Symp. on Circuits and Systems (ISCAS)* pp 1–4
[47] Mishra N, Mittal P and Kumar B 2019 Analytical modeling for static and dynamic response of organic pseudo all-p inverter circuits *J. Comput. Electron.* **18** 1490–500

2023 3rd Asian Conference on Innovation in Technology (ASIANCON)
Pune, India. Aug 25-27, 2023

Detection of Lung Disorders From X-Ray Images Using Hybrid Deep Learning

Aneesha Sharma[1], Anshika Yadav[2], Manoj Kumar[3]
[1,2,3] Department of Computer Science and Engineering
[1,2,3] Delhi Technological University, New Delhi, Delhi
aneeshasharma142@gmail.com[1], anshikay.581@gmail.com[2], mkumarg@dce.ac.in[3]

***Abstract*—Worldwide, lung illness is a ubiquitous event. With the use of emerging technology, machines, and computer power, we can help in the timely and more precise diagnosis of diseases, particularly lung disease, which can save a great deal of lives and lessen the strain on the system. These include Tuberculosis, pneumonia, asthma, fibrosis, COPD, and others. To predict lung disease, several types of deep learning techniques exist that includes convolutional neural networks (CNN), vanilla neural networks (VNN), visual geometry group neural networks (VGG) , and capsule networks. We propose a hybrid model where CNN is combined with VGG, data augmentation, and the spatial transformer network (STN) that beats the existing techniques for both sample datasets as well as complete datasets in terms of measures like precision, recall, F0.5 score, and validation accuracy. The purpose of this research is to develop an autonomous, exact, dependable, strong, and intelligent aiding system that most effectively diagnoses lung illness from a chest X-ray. The "National Institutes of Health Chest X-Ray Dataset" was utilized by the authors of this paper to train the proposed system. The model reported an unmatched accuracy of 95.43% on sample dataset and 97.56% on full dataset comprising of lung X-rays. Moreover, a comparison with the existing models and approaches shows that the proposed system finds clearer demarcation between the highly similar chest X-rays of various diseases.**

Keywords—Capsule Network, Convolutional Neural Networks, Deep Learning, Pre-processing, Spatial Transformer Network, Visual Geometry Group.

I. INTRODUCTION

With an increase in health-altering conditions comes an increase in the responsibility of doing our best to overcome them. Around 3.3 million people passed on in 2019 from constant obstructive pneumonic illness (COPD), which is commonly welcomed by Smoking, and contamination causes 400,000 asthma passings each year [1]. Especially in arising and low-pay countries, where a huge number of individuals battle with destitution and air contamination, there is an extremely high risk of lung problems. As per WHO gauges, diseases connected to home air contamination, for example, asthma and pneumonia, bring about roughly 4 million preventable passings each year, let apart COVID-19. Hence, it has become even more important to give this issue priority. While preventing these by fighting air pollution and opting for a healthy lifestyle stays at the top, we also need to look into efficient diagnostic practices as well. This is where our research comes into play.

For this, Machine learning and AI can be very helpful. Computerized innovation has, as of late, acquired significance on a worldwide scale. This study can point doctors and different analysts in the correct way for utilizing profound figuring out how to recognize lung issues. The method talked about here can likewise assist with more exact disease identification, safeguarding numerous vulnerable individuals and bringing down the fatality rate. Considering the tremendous rise in computer science for health and medical science, this approach can add to cutting clinical uses. The NIH chest dataset of X-ray pictures is acquired for the implementation from the Kaggle which is an open-source stage [2]. This exploration presents a hybrid method that effectively groups lung diseases utilizing the NIH dataset. The most significant outcome of this study is the creation of a revolutionary hybrid system using deep learning that can correctly identify lung disease from X-ray images.

There were numerous research papers proposing the detection of some or the other kind of lung disease [3-5]. But some diseases which have similar symptoms were difficult to be distinguished with great accuracy. Older methods required powerful computers with a lot of data to process. The suggested approach is driven by the requirement for accurate and reliable lung illness detection and classification, including fibrosis and viral or bacterial pneumonia.

A. Problem Statement

A sizable collection of X-ray lung data was just made available on Kaggle, along with data on lung diseases and patients. We have used this dataset to forecast that if the patient has a lung ailment or not. We will go through some preprocessing followed by machine learning and deep learning algorithms to make this prediction. The input for this project is information on the patient (age, gender, X-ray pictures, and View Position), and the output is the presence or absence of illnesses. Our first problem would be this new dataset, and we'll be one of the first to learn it. As indicated, the dataset is immense however has never been completely handled, there is a lot of noise, and an X-ray of the lung is probably not going to offer an adequate number of indicators to decide if a patient might be unwell without other external symptoms like fever, bleeding, etc. However, we will try to overcome these and provide as accurate a diagnosis as possible.

To examine the information and construct models for patient diagnosis, we will be using Machine learning and Deep Learning. Our primary concerns will follow: making the data ready, coordinating the handling of patient information with their X-ray information, using CNN with a few pre-trained models and training some for various diseases, and utilizing the CapsNet network.

The rest of the paper is organized as follows: Section 2 discuss Literature review; Section 3 talks about the Dataset

2023 3rd Asian Conference on Innovation in Technology (ASIANCON) | 979-8-3503-0228-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ASIANCON58793.2023.10270314

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:45:48 UTC from IEEE Xplore. Restrictions apply.

Description; Section 4 explain the Proposed Methodology; Section 5 illustrates the Results Discussion and Future Work; and Section 6 concluding the research work done.

II. Literature Review

Earlier with the wave of digitalization, Computer Aided Detection was used for detecting tumors and various other diseases. However, with time, it is not sufficient to accurately detect in time. Now, these diseases can be efficiently detected using various Convolutional Neural Network models, each one of them having its own advantages and disadvantages. Therefore, to understand and analyze them before choosing models for our project, we went through a number of papers [1-8]. To diagnose lung cancer, COPD, Pneumonia, Asthma and other lung disorders, numerous deep learning models are used.

Ali Hussain et. al. [3] identified the various COPD stages using the feature selection technique Recursive Feature Elimination and Cross-Validated (RFECV). The RFECV method's properties were discovered to be able to achieve accuracy of 96%.

C.-H.Liang et al. [4] made it quite evident that on chest radiographs, CAD is crucial for the early detection and diagnosis of pulmonary nodules and masses. Chest radiography is amongst the most commonly used diagnostic test in routine medical practise because it is easily accessible, reasonably priced, and widely available in outpatient facilities.

Taufik Rahmat et. al. [5] set out to examine the methods for classifying CXR images used in the medical image analysis. This knowledge and insight will be crucial for advancing future research, particularly when using machine learning to enhance the precision of CXR picture interpretation.

K.Kallianos et. al. [6] proposed to make greater use of the existing annotated examples, a network and training technique which significantly relies on data augmentation. On the ISBI challenge, for the segmenting neural structures in the electron microscopic stacks, it was demonstrated that such a network outperforms the previous top method (a sliding-window convolutional network).

Olaf Ronneberger et. al. [7] discussed that had various classification models that not just classified pneumonia, tuberculosis and covid, but also went ahead to classify covid on a severity basis such as low, medium and high. The models used for this paper were VGG-16, DenseNet-161 and ResNet-18. What really caught our eye was the accuracy achieved by them which is as high as 98.9%.

Madhura Inamdar et. al. [8] showcased that used models VGG-19, Resnet 50V2 and Densenet201. They used the Confusion matrix and ROC curve and achieved accuracies of 98%, 92% and 97% respectively. The use of adam optimizer here resulted in achieving highly accurate and consistent results.

Latheesh Mangeri et. al. [9] focused on fVDSNet and Vanilla NN. It also implemented Capsule Network. Here, VDS essentially consists of CNN, a spatiotemporal transducer system, and VGG information enrichment. And although VDSNet outperformed all the other models, it took longer to develop while requiring a lesser learning curve and vice-versa.

III. Dataset Description

A. Data Exploration

In this section, we study and explore the dataset in depth. In data exploration, we see what the dataset is comprised of and try to summarize it fully. In data visualization, we visualize the information to understand the dataset in a better way to be able to utilize this dataset further for our model.

The summary of the entire dataset is as mentioned:

- The dataset consists of 12 folders united by 112,120 X-ray photos, with a 1024 x 1024 resolution from 30,805 unique patients. The biggest freely open dataset in Kaggle before this information was made accessible was Openi, which had 4143 chest or lung X-beam filters and therefore this is a very big improvement to make our model more accurate than the already existing ones.
- We have created the testing dataset by taking the 10% of the whole dataset i.e., 11,212 images.

The following describes how the class is portrayed. There are 15 classifications (14 illnesses and one called "No findings"). Pneumothorax- 542 images, Consolidation- 452 images, Infiltration- 1934 images, Emphysema- 254 images, Atelectasis- 1016 images, Effusion- 1328 images, Fibrosis- 168 images, Pneumonia- 124 images, Pleural thickening- 352 images, Hernia- 26 images, Cardiomegaly- 282 images, Nodule- 626 images, Mass- 568 images, and Edema- 236 images are all diseases that can be classified. For supervised learning, these classifications are said to be approximately 90% accurate.

The patient information and class names from the entire dataset are listed below: • Victim ID • Picture index • X-ray orientation (View position) • Victim gender • Victim age • Actual Picture Height • Actual Picture Width • Actual Picture Pixel Spacing_x • Follow-up • Actual Picture Pixel Spacing_y.

Next, This information encases significant features for the arrangement of information developed as: patient gender, patient age, view position and even lung X-ray pictures. This vital data will be used to prepare the CNN model.

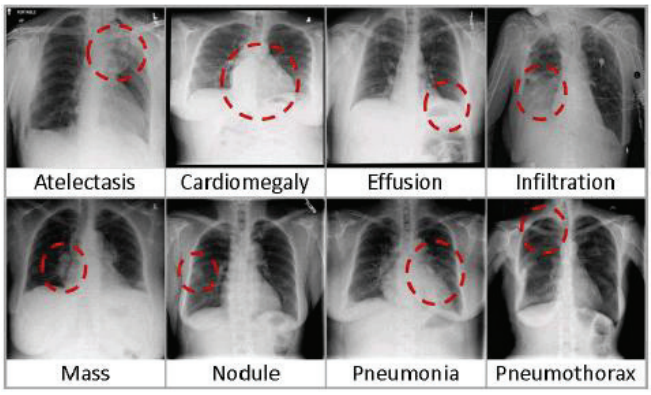


Fig. 1. Eight typical thoracic disorders were detected in chest X-rays, supporting a challenging task with a fully automated result.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:45:48 UTC from IEEE Xplore. Restrictions apply.

B. Data Visualization

Visualization of the dataset is very beneficial for knowing various aspects of the data like the distribution of data that helps us determine the split for training and testing. We also learn about some patterns like how a factor affects other factors, outliers, etc. Therefore, we performed visualization of the Chest X-ray Dataset by National Institute of Health (NIH).

We first visualized the dataset on the basis of gender where we saw that male entries are more than female entries yet the distribution is not highly skewed.

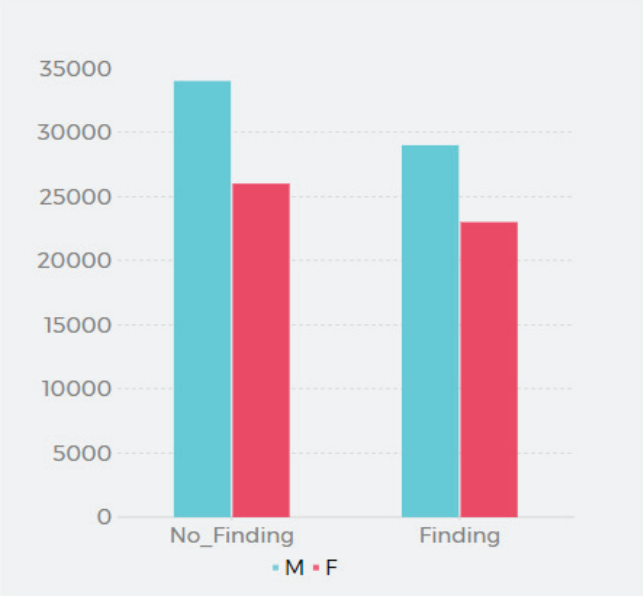


Fig. 2. Representation of the total dataset's patient population by gender and disease about binary categorization

As seen in Fig 2, In this dataset, the total number of men is higher than the total number of women, and the number of confirmed cases is higher than the total number of men with diagnosed lung illnesses.

Next, we visualized the distribution of the diseases. This visualization assisted us in understanding that, despite selecting the largest dataset available, the majority of data entries include no findings while other diseases such as infiltration, edema, emphysema, etc. are represented by other diseases. As a result, it can be claimed that the distribution of diseases is quite skewed. This also aided in the selection of the training dataset, which allowed for higher accuracy and inclusion of all disorders.

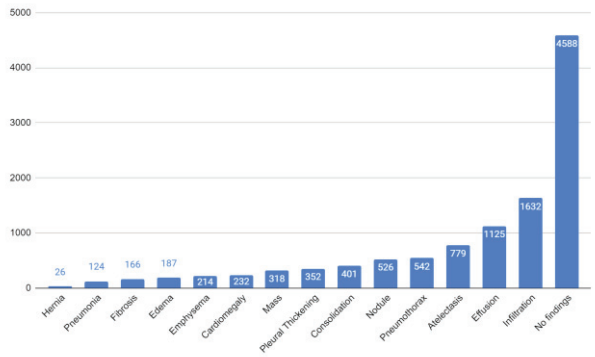


Fig. 3. Representation of the total dataset's patient population by gender and disease about multiclass categorization

Next, we observed the distribution of the view positions, including both Posterior-Anterior (PA) and Anterior-Posterior (AP) positions. When the patient is completely cooperative and in a fixed position, PA images are thought to be more accurate and taken. Since these are most accurate, having more occurrences of these is beneficial for our model. Whereas AP views are taken when the patient is unable to cooperate. Having both these views can help us build a more efficient model by taking various points of the lung in consideration.

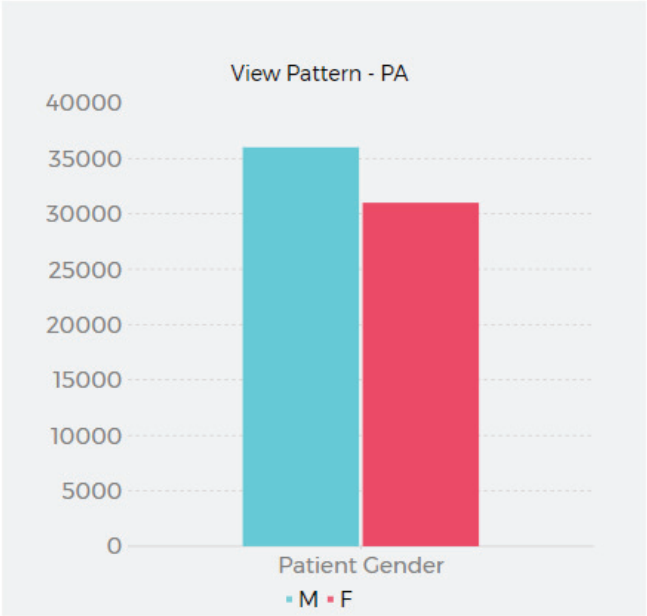


Fig. 4. Posterior-Anterior distribution of patients for sample dataset

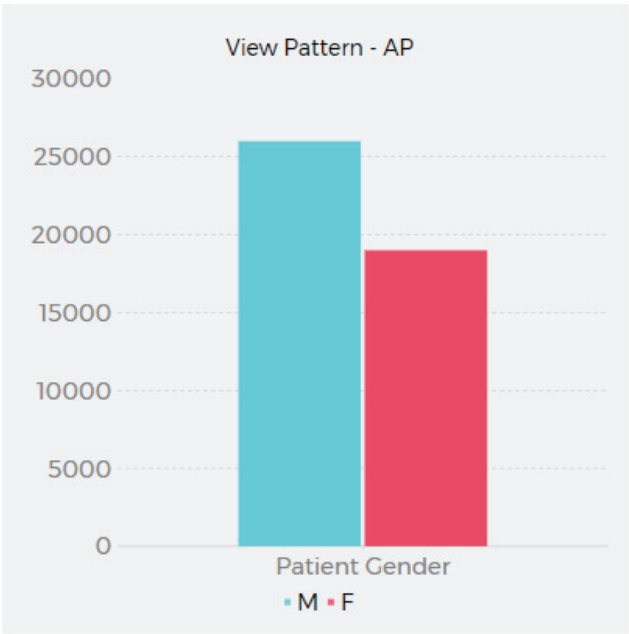


Fig. 5. Anterior-Posterior distribution of patients for sample dataset

IV. Proposed Methodology

In this section, we discuss how to prepare, handle, and split a dataset, as well as the model architectures used for preparing the ready dataset. Next, we explain the existing models' and proposed model's training details.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:45:48 UTC from IEEE Xplore. Restrictions apply.

A. Data Preprocessing

Numerous X-ray images make up the dataset. Additionally, the dataset can be used to gather other statistics like age or gender distribution.

The following lists the preprocessing steps performed in this study:

1) For images

- First, resize all of the images to make them smaller so that training would go more quickly.
- All of the images are converted to RGB and grayscale, then run in parallel on multiple models.
- The image matrix is divided by 255 to normalize the NumPy array that is being used to read the images at the time.

2) For additional information

- Redefine a few of the specific features for further information.
- After converting the age data to a numeric system, normalize the year field as well.
- Remove the outliers in the age attribute.
- This paper will conduct two crucial attributes in both datasets without distinction: "view position" and "patient gender."

After processing, all image data is stored for further use. Resized picture's form are one of the pre-processing's adjustable parameters. Following is one of the ways in which we have done this.

```
No. of rows before removing rows having age >100 :  112120
No. of rows after removing rows having age >100 :  112104
No. of datapoints having age > 100 :  16
```

Fig. 6. Snippet of ways in which we have done the preprocessing

B. Model Description for Existing Methods

The current algorithms, CNN, and the capsule network (CapsNet) are explained in this section. These two algorithms can incorporate the key traits from age, gender, and other demographic data as well as picture data. Amongst the most potent deep learning networks which can have numerous layers, mostly hidden, is CNN. To extract any level features from the data input, these hidden layers execute convolution and subsampling extremely well. Therefore, for this dataset, the effectiveness of CNN is assessed foremost.

- CapsNet

Sabour et al. suggested CapsNet in 2017 [12]. Equivariance, that preserves the spatial connection of features inside an image without changing the size or direction of the image, is amongst one of the most important characteristics of the network. Since it has been in use for many models like for classifying brain tumors from brain MRI images.

CapsNet with modified parameters produces a decreased feature map (feature size reduction) and reliable prediction accuracy. Many comparisons reveal that CapsNet outperforms all various networks in the scenario of a constrained as well as unbalanced dataset. Its effectiveness is contrasted with the other models in the instance of the huge dataset. Accuracy and training time calculations are used to assess the performance capacity of basic and modified CapsNet. Hence, a hybrid model will be implemented in this paper.

- CNN

CNN offers a lot of benefits, such as the ability to extract significant characteristics from pictures with little computing effort. Numerous CNN-related issues are taken into account in this work. These are training settings, data augmentation, and preprocessing parameters that can be sufficiently tuned. A basic CNN model will serve as the benchmark model.

A simple CNN model would be ideal to compare to what's optimal and apply the Advanced techniques as no team has ever developed a comprehensive model for this dataset as we hope to achieve in this project. Therefore, using these architectures, we will propose a hybrid model that will use the best features of both the models described above and compare them with these models as well. We have used TensorFlow and Keras for this and implemented this in a Jupyter notebook.

C. Model Description for Proposed Method

For the proposed model, the calculations are carried out utilizing Jupyter Notebook and various open source libraries such as Tensorflow, Keras etc. The execution methods are portrayed underneath. This method is the vital plan of this paper and is available on the Jupiter Notebook named as "VDSNet for training dataset" and "VDSNet for full dataset".

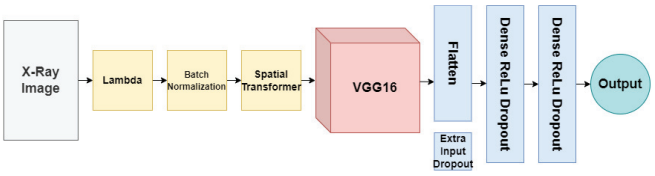


Fig. 7. Architecture of the proposed VDSNet

The following request is built up of three essential layers:

1) Layers of spatial transformers

- Three layers are present.
- The first segment, lambda, moves the default routing [-0.5: 0.5], demonstrating that X-ray images highlights have an average of 0.
- Batch normalization is done in the next section.
- The spatial transformer, the third layer, is used to remove the most crucial features for characterization of lung infection.

2) Extraction of layer elements

- We have pre-trained a VGG16 model.
- The VGG16 architecture contains three dense layers, five maximum pooling layers, and thirteen convolutional layers. Hence, while it contains 21 total complete layers, there are only 16 weight layers.
- As shown in table 1 below, five models are applied to VGG16. After the convolutional layers, for instance, model 3 has eight layers. These layers are:

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:45:48 UTC from IEEE Xplore. Restrictions apply.

the GAP layer, the FC layer with 512 neurons, the dropout layer, the 2nd FC layer with 256 neurons, the 2nd dropout layer, the 3rd FC layer with 128 neurons, the 3rd dropout layer, and a classification layer with the capacity to enact SoftMax. The dropout layer's drop rate is half in each of the scenarios.

TABLE I. FIVE LAYERS OF VGG16

Layers	Configuration
1	GAPFC(4096) → FC(4096)→ Softmax
2	GAP→Softmax
3	GAP→FC(512) →Dropout(0.5) →FC(256)→Dropout(0.5)→FC(128)→Dropout(0.5)→Softmax
4	GAP→FC(512) →Dropout(0.5) →Softmax
5	GAP→FC(512) →Dropout(0.5) →FC(512)→Dropout(0.5)→FC(256)→Dropout(0.5)→Softmax
GAP → global average pooling FC → fully connected Drop rate: 50%	

3) Classification Layers

- As a consequence of the layers of VGG16 in this case, with an added five elements, such as "Age", "Gender Male", "Gender Female", "PA View position" and "AP View position" the primary layer is described in this case as the flattened layer. Five additional features are gathered to the next layer since they will also have an influence on the arrangement. This layer is appropriately referred to as the dropout layer.
- Dense dropout layers are the final 2 layers that decrease the depth.

This method's stages are depicted in the order where first, the dataset's storing and management into RAM as it is where the pictures are kept in RGB lung X-ray picture design and will be processed as before. Then, we implement the network structure. After that we perform the measuring capability with a precision score, a score for binary accuracy via a limit, and Fβ. Then we execute the data model generator, checkpoint, and loss of model. The next step is model training using boundary parameters, validation loss with precision training/logging training/approval. And at last, we test the dataset.

Keras uses CNN and deep learning, with Tensorflow-gpu serving as the backend. Testing and adjusting with several picture sizes revealed that 64 by 64 picture size is excellent as well as quite bright to identify the picture's condition by the classifier. The front layer is maintained as λ layer while the spatial transformer is in use. This STN layer uses a "locnet" model. This makes it easier to extract important details from the images. The underlying model has been tested in several environments using non-corresponding datasets. The most acceptable and appropriate layer is often regarded as the top layer. Alteration and advancement of the review's outside borders, correctness, and Fβ score are crucial. The dropout layer's record has to be refined.

D. Performance Metrics

In this study, various performance criteria are taken into account. These are the binary classification-specific recall, precision, and F-score (where β is 0.5). Because binary classifying is used to discover or find diseases when other programs are unbalanced, F-score here is considered to be stronger than accuracy. Consider a minor classifier that, given an unbalanced dataset, only predicts the majority class. When the size of training data is significantly larger than the testing size, this classifier will perform well; but, the accuracy will be poor when the training and testing sizes are roughly close. This study takes into account a variety of indicators for lung illness diagnosis.

Testing accuracy, recall, precision, as well as F1-score have been taken into account. These metrics are also known as TP-True positive, TN-True negative, FP-False positive and FN-False negative. TP points to the patients with probable lung illness who are accurately identified as having lung disease in the context of this work. The word TN points to the count of samples having normal lung function. The phrase FN describes suspected people who have lung disease but are still going undiagnosed by the system. In addition, FP refers to many people who do not have lung disorders but were misdiagnosed. The formulas for recall and precision are given below:

$$Recall = \frac{TP}{TP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Combining recall and precision can be a crucial statistic. The following sentence can be used to summarize the recall and precision combination known as the Fβ score:

$$F_\beta = (1+\beta^2)\frac{Recall\ X\ Precision}{\beta^2.Precision+\ recall} \quad (3)$$

When choosing the importance of recall and precision, there are two key concepts to keep in mind:

- In a real-world setting, the model will be helpful for detecting lung disorders if it has good performance outcomes. It is very important since it can be viewed as a method to assist doctors in applying additional diagnostic procedures. High precision and Low recall corresponding with a small β are therefore required. Therefore, the F-β score has been considered to be β = 0.5.
- In order to prevent infection, the suggested models should avoid pronouncing sick persons incorrectly. Models should not omit vulnerable patients. Low accuracy and High recall is connected with a big β will be preferred for this circumstance. In this instance, the F-β score has been considered β to equal 2.

The suggested effort will assist doctors in promptly diagnosing ailments as a patient must undergo numerous

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:45:48 UTC from IEEE Xplore. Restrictions apply.

tests to determine the disease. Prior to receiving the results of further tests, the affected patient will be anxious. As a result, this essay recommends an F-0.5 grade, which β is 0.5.

V. Results Discussion and Future Work

The performance of our model is now evaluated considering the loss values against the epochs. In the graphs below, 'loss' demonstrates the training loss whereas the 'val_loss' shows validation loss for both training and full dataset.

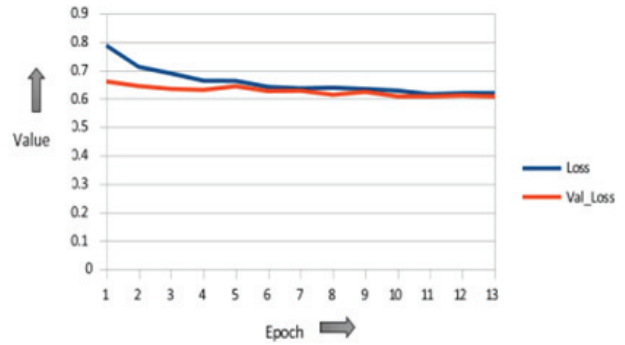


Fig. 8. Model evaluation against epochs for training dataset.

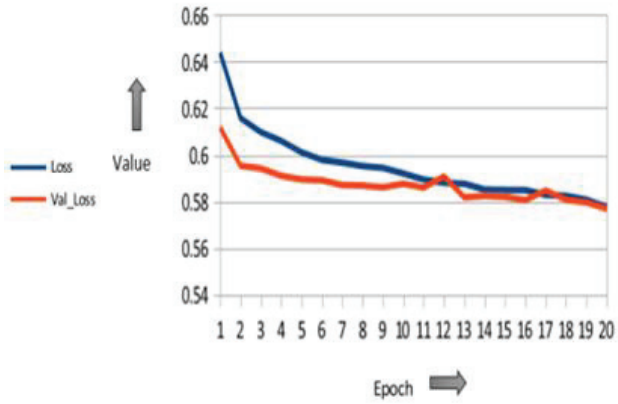


Fig. 9. Model evaluation against epochs for full dataset.

Due to the lack of information on the features of the large photos, VDSNet has a great performance but converges gradually. Extra information in the full dataset may further help in decreasing the convergence time.

Several models may be compared as indicated in Table 2 based on the techniques' accuracy on both the entire dataset and sample dataset.

TABLE II. Comparison of Several Models' Recall, Precision, F-Score, Validation Accuracy, and Training Times

Data set	Structural design	Validation Acuuracy	No. Parameters	Training Time (Sec)	Recall	Precision	Fβ (0.5) score
Sample Dataset	Vanilla Grey	50.70%	321225	2	0.5	0.58	0.56
	Vanilla RGB	51.80%	322793	2	0.59	0.62	0.61
	Hybrid CNN VGG	68%	15252133	16	0.56	0.65	0.63
	VDSNet	**95.43%**	**16482451**	**23**	**0.64**	**0.62**	**0.64**
	Modified CapsNet	59%	12167424	37	0.42	0.71	0.45
	Basic CapsNet	57%	14788864	75	0.6	0.62	0.62
Full Dataset	Vanilla Grey	67.80%	321225	51	0.58	0.68	0.66
	Vanilla RGB	69%	322793	53	0.61	0.68	0.66
	Hybrid CNN VGG	69.50%	15252133	384	0.62	0.68	0.67
	VDSNet	**97.56%**	**16482451**	**452**	**0.63**	**0.69**	**0.68**
	Modified CapsNet	63.80%	12167424	856	0.48	0.61	0.58
	Basic CapsNet	60.50%	14788864	1815	0.51	0.64	0.61

Table 2 shows that VDSNet, which outperforms the benchmark vanilla CNN, is the best model. Compared to plain CNN, the training process takes longer. But, by continuing to train with further epochs, the VDSNet model can be enhanced. The capsule network model, on the other hand, does happen to function well; although only VDSNet's number of parameters are the same, the training period is substantially longer. The validation accuracy of VDSNet is 97.56%. The algorithm still needs to be improved in order to meet the requirements for use in hospitals as there we are dealing with lives and no room for error can be left, which will take more time and computing resources.

TABLE III. Comparison of Several Models on the Same Dataset

Dataset	Author	Methodology	Accuracy	Remarks
NIH Dataset	Om Uparkar et. al. [13]	Vision Transformers	70.24%	Yields similar levels of performance and competitive alternative to CNNs
	Mei-Ling Huang et al. [14]	LightEfficientNet V2	83.42%	Transfer learning models before and after fine-tuning are executed and results are compared.
	R. Rajagopal et. al. [15]	Deep Convolutional Spiking Neural Network optimized (DCSNN) with Arithmetic	96.25%	AOA is utilized for tuning hyper parameters of DCSNN

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:45:48 UTC from IEEE Xplore. Restrictions apply.

		Optimization Algorithm (AOA)		
	FM Javed Mehedi Shamr at et. al. [17]	Fine-tuned MobileLungNetV2	96.97%	Hyper-parameter tuning was conducted to increase the architecture durability in terms of lung lesion identification
	Our Proposed	**VDSNet**	**97.56%**	**a hybrid model where CNN is combined with VGG, data augmentation, and the spatial transformer network (STN)**

We compared the outcomes of several techniques that were used on the same dataset in Table 3 above. Yet, it is clear that our method outperformed all the others on the NIH dataset.

This document can be expanded in a variety of ways as future work. The following list includes a few of these.

- In order to distinguish between the various lung illnesses, the model needs to be tested. The very skew data issue for each condition may then be understood.
- The suggested model should be trained across a large number of epochs while tweaking a few parameters for speedy convergence.
- The chance of getting major characteristics will improve if training shots can be made larger. Yet doing so can make the training process longer.
- It is possible to experiment with a number of pre-trained models in order to construct CNN with the fusion of VGG.
- An intricate "locnet" module has been utilised to produce hybrid CNN with the addition of a spatial transformer.

VI. Conclusion

This study suggests using the unique hybrid deep learning structure VDSNet to identify lung conditions in X-ray images. The suggested model is tested using a sample of NIH chest X-ray data acquired from the Kaggle library. The best validation accuracy for the complete dataset is 97.56% for VDSNet, whereas the accuracy values for vanilla grey is 67.8%, vanilla RGB is 69%, hybrid CNN VGG is 69.5%, basic CapsNet is 60.5%, and modified CapsNet is 63.8%. VDSNet has a validation accuracy value of 97.56%, which is higher than the sample dataset's accuracy score of 95.43%. However, compared to the 23 second time needed for the sample, VDSNet takes a training period of 452 seconds for the complete dataset.

Further developments are anticipated to increase the model's accuracy and make the projected VDSNet useful in medical clinics. Essentially CNN often has a poor performance at pivoting, shifting, or another odd photo direction. As a consequence, hybrid frameworks have been used to boost accuracy without lengthening training periods. According to the research's findings, deep learning models may be used to further refine the findings, which can raise the standard of treatment provided to the patient, in contrast to conventional techniques.

In chest X-ray images, our technique can effectively identify the provocative zone. As it relates to managing the enormous dataset, this investigation endeavor is still currently experiencing these challenges. Thus, the utilization of little datasets can give great exactness yet it won't be successful in genuine applications.

References

[1] World Health Organization. (n.d.). Chronic obstructive pulmonary disease (COPD). World Health Organization. Retrieved April 12, 2023, from https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd)

[2] National Institutes of Health Chest X-Ray Dataset. (2018, February 21). NIH chest X-rays. Kaggle. Retrieved April 12, 2023, from https://www.kaggle.com/datasets/nih-chest-xrays/data

[3] Detection of different stages of COPD patients using ... - IEEE XPLORE. (n.d.). Retrieved April 13, 2023, from https://ieeexplore.ieee.org/document/9370958/

[4] Author links open overlay panelC.-H. Liang a b c, a, b, c, d, e, f, g, AIMTo test the diagnostic performance of a deep learning-based system for the detection of clinically significant pulmonary nodules/masses on chest radiographs.MATERIALS AND METHODSUsing a retrospective study of 100 patients (47 with clinically significan, Greiner, M., Cheng, J.-Z., Zhang, W., Levin, D. C., Wu, F. Z., Hsu, H. T., Groot, P. M. de, Brogdon, B. G., Aberle, D. R., Bhargavan, M., … Novikov, A. A. (2019, September 11). Identifying pulmonary nodules or masses on chest radiography using Deep Learning: External validation and strategies to improve clinical practice. Clinical Radiology. Retrieved April 13, 2023, from https://www.sciencedirect.com/science/article/abs/pii/S0009926019303782?via%3Dihub

[5] Chest X-rays image classification in medical image analysis - researchgate. (n.d.). Retrieved April 13, 2023, from https://www.researchgate.net/publication/330105218_CHEST_X-RAYS_IMAGE_CLASSIFICATION_IN_MEDICAL_IMAGE_ANALYSIS

[6] Author links open overlay panelK. Kallianos a, a, b, c, Due to recent advances in artificial intelligence, Wu, M.-H., Quekel, L. G. B. A., Kim, D. H., Deng, J., Krizhevsky, A., Chartrand, G., Kohli, M., Erickson, B. J., Shie, C.-K., Kim, H. G., Ravishankar, H., Chang, J., Bar, Y., Boone, J. M., … Hogeweg, L. (2019, January 28). How far have we come? artificial intelligence for chest radiograph interpretation. Clinical Radiology. Retrieved April 13, 2023, from https://www.sciencedirect.com/science/article/abs/pii/S0009926019300194

[7] Ronneberger, O., Fischer, P., & Brox, T. (1970, January 1). U-Net: Convolutional Networks for Biomedical Image Segmentation. SpringerLink. Retrieved April 13, 2023, from https://link.springer.com/chapter/10.1007/978-3-319-24574-4_28?utm_source=getftr&utm_medium=getftr&utm_campaign=getftr_pilot

[8] U.S. National Library of Medicine. (n.d.). Shelke a[author] - search results - pubmed. National Center for Biotechnology Information. Retrieved April 13, 2023, from https://pubmed.ncbi.nlm.nih.gov/?term=Shelke+A%5BAuthor%5D

[9] Chest diseases prediction from X-ray images using CNN models: A study. (n.d.). Retrieved April 13, 2023, from

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:45:48 UTC from IEEE Xplore. Restrictions apply.

https://thesai.org/Downloads/Volume12No10/Paper_26-Chest_Diseases_Prediction_from_X_rays_Images.pdf

[10] International Journal of Mechanical Engineering Detection of lung ... (n.d.). Retrieved April 13, 2023, from https://kalaharijournals.com/resources/IJME_Vol7.1_560.pdf

[11] Chestx-Ray8: Hospital-scale chest X-ray database and ... - IEEE xplore. (n.d.). Retrieved April 13, 2023, from https://ieeexplore.ieee.org/document/8099852

[12] Sabour, S., Frosst, N. & Hinton, G. E. Dynamic routing between capsules. Adv. Neural. Inf. Process. Syst. 30, 3856–3866 (2017).

[13] Uparkar, O., Bharti, J., Pateriya, R. K., Gupta, R. K., & Sharma, A. (2023). Vision transformer outperforms deep convolutional neural network-based model in classifying X-ray images. Procedia Computer Science, 218, 2338–2349. https://doi.org/10.1016/j.procs.2023.01.209

[14] YC;, H. M. L. L. (n.d.). Stacking ensemble and ECA-EfficientNetV2 convolutional neural networks on classification of multiple chest diseases including covid-19. Academic radiology. Retrieved April 13, 2023, from https://pubmed.ncbi.nlm.nih.gov/36526533/

[15] Author links open overlay panelR. Rajagopal a, a, b, c, d, Highlights•DCSNN optimized with AOA is proposed for Lung Disease Detection using Chest X-ray Images. Deep convolutional spiking neural network optimized with arithmetic optimization algorithm for lung disease detection using chest X-ray images. Biomedical Signal Processing and Control. Retrieved April 13, 2023, from https://www.sciencedirect.com/science/article/pii/S1746809422006516

[16] Bharati S, Podder P, Mondal MRH. Hybrid deep learning for detecting lung diseases from X-ray images. Inform Med Unlocked. 2020;20:100391. doi: 10.1016/j.imu.2020.100391. Epub 2020 Jul 4. PMID: 32835077; PMCID: PMC7341954.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:45:48 UTC from IEEE Xplore. Restrictions apply.

Devanagari Character Recognition using MLP-Mixer and CNN Extracted Features

Abhay Gupta
Dept. of Computer Science and Engineering
Delhi Technological University
Delhi, India
abhaygupta_2k19co013@dtu.ac.in

Abhishek Jha
Dept. of Computer Science and Engineering
Delhi Technological University
Delhi, India
abhishekjha_2k19co019@dtu.ac.in

Abhishek Sharma
Dept. of Computer Science and Engineering
Delhi Technological University
Delhi, India
abhisheksharma_2k19co023@dtu.ac.in

Anukriti Kaushal
Dept. of Computer Science and Engineering
Delhi Technological University
Delhi, India
anukritikaushal@dtu.ac.in

2023 IEEE World Conference on Applied Intelligence and Computing (AIC) | 979-8-3503-1006-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/AIC57670.2023.10263816

***Abstract*—Devanagari is an ancient script that is used to write Hindi, Nepali, Marathi, Maithili, Awadhi, Newari, and Bhojpuri, among other Indo-Aryan languages. Thousands of individuals in India use this script to write documents in Marathi and Hindi. Indian mythology is based on this script. Because of the script's prominence, handwritten Devanagari character identification has grown in popularity over time. In this paper, we present a comparative analysis of different machine learning and deep learning models for Devanagari character recognition. Multi-Layer Perceptron(MLP) Mixer is the technology we're using because extensive research has been done in Devanagari Text Recognition using Convolutional Neural Networks(CNN). This paper aims to provide a comparison between traditional models and a newer model(MLP Mixer). We have used CNN as a stand-alone deep learning model for our problem and used it for feature extraction. We have also developed some traditional machine learning classifiers using these learned complex features.**

***Index Terms*—convolutional neural networks, ensemble learning, feature extraction, machine learning, mlp classifier**

I. Introduction

Humans have long been curious about whether computers can be programmed to understand and observe things in a similar fashion that humans can do using their sense organs. This gave rise to the advancement of technology both in the advancement of deep learning techniques and conventional machine learning approaches. Also, the advancement of hardware and introduction of GPUs led to more processing power which in turn resulted in the reduction of training times of these models.

Character recognition is an important task, if it can be done with the help of a computer, it can ease human efforts, in reading large texts, written not only in English but also in their native languages for example Devanagari(Fig. 1) [1]. Character Recognition is a challenging task, it means that the computer has to identify the character through its image, which is not at all fixed and can vary depending on the person who wrote that character, because every person has different handwriting, When the researchers were thinking to perform this mammoth task, they were amazed that how human beings are able to do it so easily given that there are characters that happen to be a lot similar to each other (Fig. 2) [2].

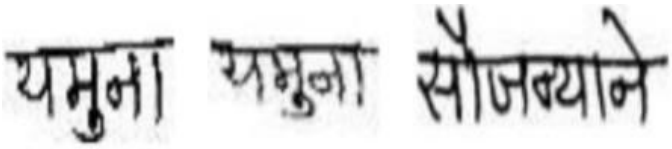


Fig. 1: Sample Devanagari Words

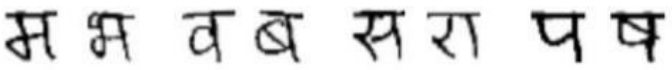


Fig. 2: Approximately Similar Devanagari Characters

People, on the other hand, have a built-in neural net, so whatever image they perceive passes through numerous layers, and this seemingly tough process is completed with ease. A fundamental task in the field of pattern recognition is OCR (Optical Character Recognition) which handles characters that are optically processed. It does not work with magnetically processed characters. OCR can be used for a variety of tasks, such as detecting a vehicle's license plate (this application is being utilized to get the car owners who are breaking the traffic rules), recognizing the address from a letter, and also recognizing the records which are kept digitally (for e.g., the Indian Government keeps the records written in Devanagari script in a digital format. Now suppose a historian wants to extract information from this huge record, but it will not be feasible as that search tool can't work on images, so in order to do that we will need character recognition for our images [3].

979-8-3503-1006-1/23/$31.00 ©2023 IEEE
DOI: 10.1109/AIC.2023.102

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 06:04:34 UTC from IEEE Xplore. Restrictions apply.

TABLE I: Related Works on Devanagari Character Recognition using Machine Learning

Reference	Dataset	Contributions	Results
Acharya et al. [4]	Devanagari Handwritten Character Dataset	Proposed the widely used DHCD dataset and the use of deep CNN for character recognition.	98.47% accuracy
Aneja et al. [5]	Devanagari Handwritten Character Dataset	Explored transfer learning starting with ImageNet weights and finetuning AlexNet, VGG, and InceptionV3 models.	95% (AlexNet), 97% (VGG), 99% (Inception)
Arora et al. [6]	Proprietary	Compared performance of SVM and ANNs on manually extracted image features: view-based, shadow, longest run features, and chain-code histogram.	94.77% (SVM) and 82.15% (ANN)
Guha et al. [7]	UCI character dataset, CVPR ISI Devanagari dataset, CMATERdb, Kaggle Devanagari dataset	Finetuned various existing models like LeNet-5, ResNet, InceptionNetV3	99.54% (UCI), 99.63% (CVPR ISI), 98.70% (CMATERdb), 97.29% Kaggle Devanagari
Jangid et al. [8]	ISIDCHAR and V2DMDCHAR	Compared performance of Deep CNN models with different adaptive gradient optimizers (RMSProp, Adagrad, Adam, and SGD)	97.30% (ISIDCHAR) and 97.65% (V2DMDCHAR)
Narang et al. [9]	Proprietary	Trained a Deep CNN focused on ancient texts.	93.73%
Singh [10]	Proprietary	Proposed an ANN trained on Histogram of Gradients (HoG) features.	97.06%

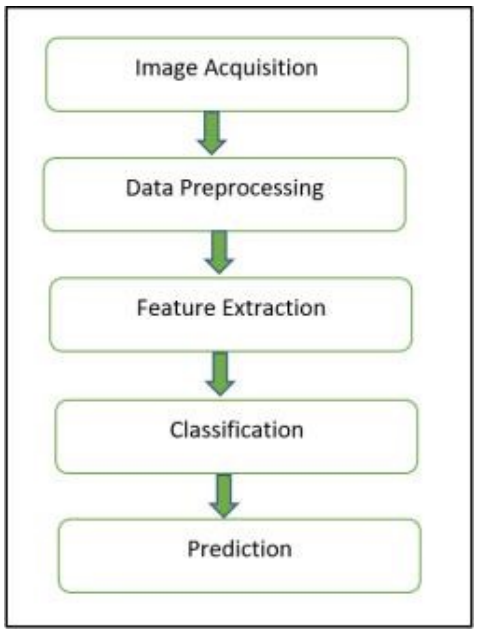


Fig. 3: Project Pipeline

II. Related Works

Sanskrit, Hindi, Marathi, and many more Indian languages are written in Devanagari, a significant script. The field of image processing and pattern recognition has shown interest in the recognition of Devanagari characters. In recent years, machine learning techniques have been successfully used for recognizing Devanagari characters [11]. This literature review discusses various machine learning-based approaches proposed for Devanagari character recognition, their contributions, and their results (Table I).

Acharya et al. (2015) proposed the widely used Devanagari Handwritten Character Dataset (DHCD) and the use of deep Convolutional Neural Networks (CNN) for character recognition. They used a 10-layer deep CNN with ReLU activation function and dropout regularization. On the DHCD dataset, the suggested technique has a 98.47% accuracy rate.

Arora et al. (2010) compared and evaluated the performance of artificial neural networks (ANN) and support vector machines (SVM) on manually derived image features such as view-based, shadow, longest-run features, and chain-code histogram. On their exclusive dataset, they attained an accuracy of 94.77% with SVM and 82.15% with ANN.

Singh (2018) proposed an Artificial Neural Network (ANN) trained on Histogram of Gradients (HoG) features for Devanagari character recognition. The proposed method acquired an accuracy of 97.06% on their proprietary dataset.

Jangid et al. (2017) compared and analyzed the performance of Deep CNN models with different adaptive gradient optimizers (RMSProp, Adagrad, Adam, and SGD) on two different datasets, ISIDCHAR, and V2DMDCHAR. With the Adam optimizer, they obtained accuracy on ISIDCHAR and V2DMDCHAR of 97.30% and 97.65%, respectively.

Aneja et al. (2019) explored transfer learning starting with ImageNet weights and fine-tuning AlexNet, VGG, and InceptionV3 models for character recognition of Devanagari characters. They achieved an accuracy of 95% with AlexNet, 97% with VGG, and 99% with InceptionV3 on the DHCD dataset.

Guha et al. (2019) finetuned various existing models like LeNet-5, ResNet, and InceptionNetV3 on multiple datasets, including the CMATERdb, CVPR ISI Devanagari dataset, Kaggle Devanagari dataset, and UCI character dataset. The accuracy obtained was 99.54% on the UCI dataset, 99.63% on the CVPR ISI dataset, 98.70% on the CMATERdb dataset, and 97.29% on the Kaggle Devanagari dataset.

Narang et al. (2021) trained a Deep CNN focused on ancient texts for character recognition in the Devanagari script. They achieved an accuracy of 93.73% on their proprietary dataset.

SVM, KNN, RF, DT, MLP, XGB, and CNN were among the machine learning models that Manocha, Shilpa, et al. [12]

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 06:04:34 UTC from IEEE Xplore. Restrictions apply.

Fig. 4: Devanagari Character Set

trained and tested using the DHCD dataset. They reported that Compared to the aforementioned models, the CNN model performed better, with an accuracy rate of 92%.

In conclusion, several machine learning methods have been put forth for the recognition of Devanagari characters, including DNN, CNN, SVM [13], ANN, LSTM [14] [15], MLP [16] and transfer learning. The results achieved on different datasets vary, with some achieving accuracy above 99%. Fine-tuning pre-trained models like AlexNet, Resnet [17] [18], VGG, MobileNet, and InceptionV3 [19] have been found to be effective. The DHCD dataset and the CVPR ISI Devanagari dataset are widely used for assessing the performance of Devanagari character recognition methods.

III. Methodology

The model uses Convolutional Neural Network (CNN) and Dense Neural Network (DNN) for feature extraction and various classifiers for classifying the images. The following approach has been made for the proposal:

1) Build and configure a CNN network to extract features.
2) Feature extraction using Convolutional and Deep Neural Networks.
3) R
    a) CNN
    b) Random Forest
    c) Multi-Layer Perceptron (MLP)-Mixer
    d) KNN
4) Training Model on 46 unique characters with 2000 images of each character.
5) Visualizing the accuracy and other metrics of each classification model.
6) Comparing the performance of the used models

A. About the Dataset

This is a publicly available handwritten Devanagari character image dataset proposed by Acharya et al. [4]. The dataset is divided into two parts: A training set makes up 70% of the dataset, and a testing set makes up 30% of it. Thousands of characters from handwritten manuscripts were extracted and manually annotated to build the dataset. The dataset size is 92000 grey-scale images with 46 classes (36 classes of characters and another 10 classes for digits). Every class contains 2000 images. The dataset is represented as a CSV file having 92000 rows and 1025 columns. The first 1024 columns stand for the grayscale pixel values (0 to 255) for each image. The last column named 'character' represents the label of each image i.e., the actual character the image represents (Fig. 4).

B. Data Preprocessing Steps

1) *Data Normalization*
   Each pixel value was divided by 255 to bring the features to a similar scale. This makes optimization algorithms like gradient descent converge faster to a solution.
2) *Encoding the Output Variable*
   First, we performed label encoding on the output variable to convert it into a numeric form. After this one hot encoding was performed to make the output suitable for deep learning models.
3) *Converting the input Vectors into Tensors*
   For CNN we need the input in form (width, height, channels) so we convert the 1024 image vector 32 x 32 x 1 tensor.
4) *Train-Test Split*
   The preprocessed dataset was split 70:30 into a training and testing dataset.

C. Feature Extraction

For feature extraction, we have used Convolution Neural Networks. After flattening, we have used two dense layers for our CNN feature extractor architecture having ReLU activation function, the output of this layer with 256 values in a vector illustrates the features extracted from the images and these are further passed to: -

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 06:04:34 UTC from IEEE Xplore. Restrictions apply.

1) Final Dense Layer of CNN Architecture having SoftMax activation function.
2) As an input to traditional Machine Learning methods such as KNN (K-Nearest Neighbors), Random Forest, and Multi-Layer-Perceptron.

In order to provide the result as a probability distribution, the SoftMax activation function is typically employed in the final layer (output layer) of a CNN. The features of the input image are represented in the output after convolution and pooling procedures. This image is classified by the Fully Connected Layer (Output Layer) based on the non-linear combination of these attributes.

Additionally, we included a dropout layer having a probability of 0.2 in our model. Dropout randomly deactivates some of the nodes during training that is during forward and backward propagation. This is done to prevent over-fitting.

D. Model Architecture

Fig. 5 gives an overview of the underlying model architecture,

1) For classification, the characteristics extracted from the Convolutional Neural Network layer are transmitted to several classifiers.
2) Based on the retrieved features, the classifiers are then used to forecast the target labels.
3) The task is a multiclass classification problem since the target class contains 46 different labels.
4) We tried out the following classifiers with the newly learned complex features as inputs.

In our CNN architecture, first, we feed the data to convolution 1 which sends it to convolution layer 1. Then this convoluted data is passed through the max pooling layer for dimensionality reduction. Then the data is fed to convolution layers 2 and 3. Then the data went to max pooling layer 2 again. Later, data passes through the dropout layer then data gets flattened. Then data passes through 3 subsequent dense layers for adding the non-linearity.

E. Classification Techniques

1) Random Forest: A well-liked machine learning algorithm is Random Forest. It is a kind of decision tree-based model that integrates different decision trees to enhance the model's functionality and robustness. With a random selection of the features and data points from the initial training data, many decision trees are produced. The outputs of the decision trees in the forest are combined to provide the final forecast after each tree has been trained on a different subset of the data.

2) K-Nearest Neighbors: A straightforward, non-parametric machine learning approach called K-Nearest Neighbours (KNN) is employed for classification and regression problems. KNN classifies a new input based on the class of the K closest points (i.e., the K "nearest neighbors") in the training data. The input data points are represented as vectors in a multidimensional feature space.

3) MLP-Mixer: MLP-Mixer is a recently proposed neural network architecture that has shown promising results in various computer vision tasks. It is a type of attention-based model that uses multilayer perceptrons (MLPs) to mix the features of different image patches. Proposed by Google Brain [20]. The MLP-Mixer architecture consists of a series of blocks, each containing two separate MLPs.The first MLP in each block is known as the "channel-mixing MLP" and is responsible for processing the features of each patch individually. It takes the patch embeddings as input and applies a series of fully connected layers with ReLU activation, which learns a set of non-linear transformations that capture the local features of each patch. The output of the channel-mixing MLP is a set of transformed embeddings that contain information about the local features within each patch.

The second MLP in each block is known as the "token-mixing MLP" and is responsible for mixing the features across different patches. It takes the transformed embeddings from the channel-mixing MLP as input and applies another set of fully connected layers with ReLU activation, which learns a set of non-linear transformations that capture the global interactions between patches. The output of the token-mixing MLP is a set of mixed embeddings that contain information about the global features across all patches.

The output of the two MLPs is then combined using a residual connecThe vanishing gradient issue is lessened by the residual connection, which also enables the model to learn more intricate features, while layer normalization helps to stabilize the training process. Compared to the most popular method for image classification, Convolution Neural Network (CNN), MLP-Mixer achieves competitive performance when trained on big datasets.

As highlighted by Fig. 6, The MLP-Mixer receives a series of picture patches (also known as tokens) as input, that are linearly projected and arranged in a table of dimensions "patches x channels". The table maintains the dimensionality of the original patches and allows for efficient processing of the data by the MLP-Mixer.

IV. Results

The standard metrics that can be interpreted from the confusion matrix have been used to evaluate our model on the test set. They are calculated using (1), (2), (3), and (4).

$$precision = \frac{TP}{FP + TP} \tag{1}$$

$$recall = \frac{TP}{FN + TP} \tag{2}$$

$$F1 = \frac{2 \times recall \times precision}{recall + precision} \tag{3}$$

$$accuracy = \frac{TN + TP}{TN + FN + TP + FP} \tag{4}$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 06:04:34 UTC from IEEE Xplore. Restrictions apply.

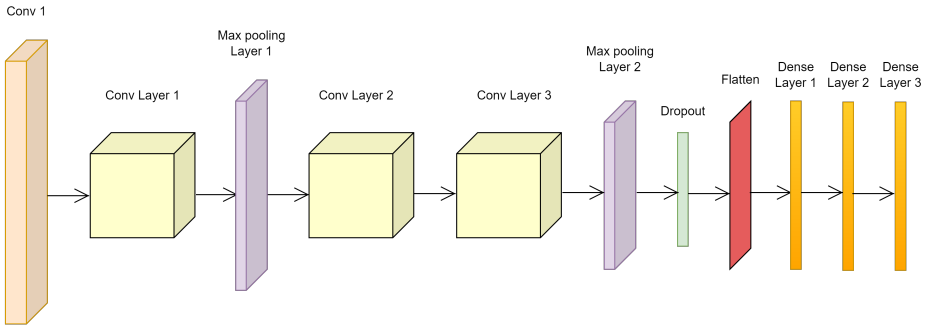


Fig. 5: CNN Model Architecture

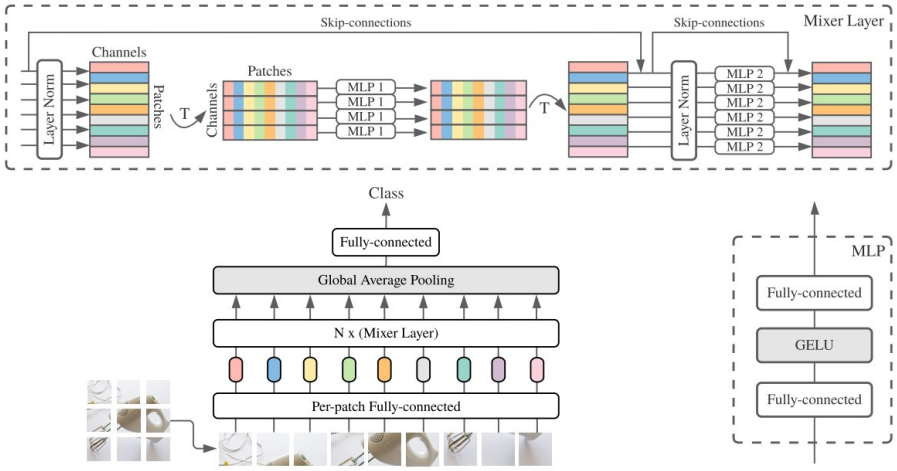


Fig. 6: MLP-Mixer: Architecture [20]

A. Model-1: CNN based neural network

After tuning the trainable parameters by the variation of kernel size, number of layers, types of layers and epochs, and other parameters, The CNN model is found to have a testing accuracy of 97.78%.

1) Training Curves: The accuracy and loss curves vs the number of epochs for both training and validation sets are shown in Fig. 7a and 7b. It can be observed from the loss vs epoch curve that choosing the number of epochs as 10 is a good choice since loss does not decrease much after 10 epochs.

2) Bar Container Plot: It can be seen that the majority of classes provide an accuracy of nearly 1, hence indicating that the model has fit the data well (Fig. 7c).

3) Classification Report: The F1-score, recall, and precision are all extremely near to 1.0. thus indicating good model performance (Fig. 7d).

B. Model-2: CNN based random forest classifier

After the features from the CNN-DNN model were extracted, the learned complex features were fed as input to the random forest classifier. This classifier gave a very good accuracy score of 98.4%.

1) Bar Container Plot: similar to model-1, the accuracy is very high and all bars are almost at the same height, hence indicating that the model has fit the data well (Fig. 8a).

2) Classification Report: F1-score, precision, and recall are just shy of 1.0, suggesting that the model is performing very well and making accurate predictions (Fig. 8b).

C. Model 3: CNN-based MLP-Mixer classifier

After extracting the features from the CNN-DNN model, the learned complex features were provided as input to the MLP-Mixer classifier. A very good accuracy score of 98.3% is obtained from this classifier.

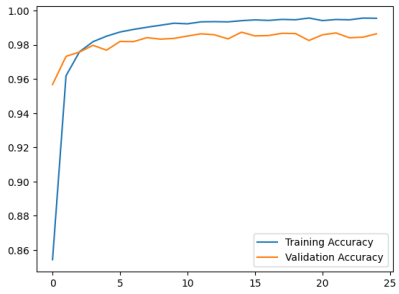
1) Bar Container Plot: From the plot, it can be observed that the model has fit the data well and hence the accuracy comes very close to 1 (Fig. 9a).

2) Classification Report: Precision, recall, and F1-score were again very close to 1.0 so this model also learned the data very well and is making accurate predictions (Fig. 9b).

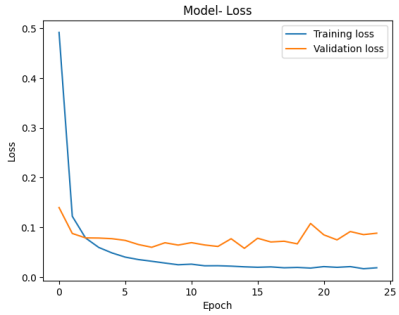
D. Model 4: CNN-based KNN classifier

After extracting the features from the CNN-DNN model, the learned complex features were fed as input to the KNN classifier. The classifier gave the best accuracy score of 98.7%.

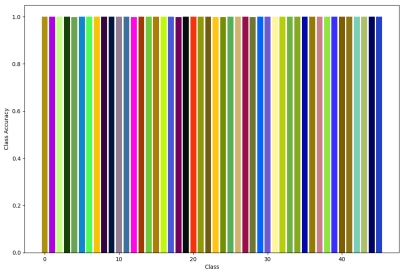
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 06:04:34 UTC from IEEE Xplore. Restrictions apply.



(a) Accuracy vs. Epochs



(b) Loss vs. Epochs

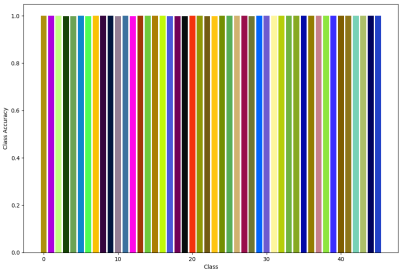


(c) Bar Plot of 46 class accuracy

	precision	recall	f1-score	support
0	0.99	0.99	0.99	590
1	1.00	0.98	0.99	605
2	0.99	0.99	0.99	572
3	0.96	0.95	0.95	601
4	0.98	0.97	0.98	597
5	0.97	1.00	0.98	560
6	0.88	0.99	0.93	597
7	0.99	0.99	0.99	631
8	1.00	0.99	0.99	618
9	0.98	0.99	0.99	609
10	0.96	0.99	0.98	600
11	0.99	0.98	0.98	601
12	0.98	0.96	0.97	598
13	0.95	0.98	0.96	589
14	1.00	0.98	0.99	558
15	0.97	0.99	0.98	604
16	0.94	0.95	0.95	643
17	0.98	0.94	0.96	588
18	0.97	0.96	0.97	592
19	0.99	0.99	0.99	626
20	0.97	0.99	0.98	627
21	0.99	0.99	0.99	589
22	0.95	0.97	0.96	606
23	0.96	0.98	0.97	596
24	0.98	0.98	0.98	615
25	0.98	0.89	0.94	586
26	0.99	0.97	0.98	584
27	0.99	0.97	0.98	590
28	0.97	0.95	0.96	631
29	0.96	0.99	0.98	590
30	0.97	0.99	0.98	641
31	1.00	0.95	0.97	595
32	0.96	0.98	0.97	581
33	0.98	0.97	0.98	624
34	0.98	0.99	0.99	593
35	0.99	0.98	0.98	622
36	1.00	0.99	0.99	624
37	1.00	0.99	0.99	611
38	0.98	0.99	0.99	597
39	1.00	0.98	0.99	574
40	0.99	1.00	0.99	613
41	0.99	0.99	0.99	574
42	0.99	0.94	0.97	596
43	0.99	0.99	0.99	577
44	0.98	0.98	0.98	586
45	0.99	1.00	0.99	599
accuracy			0.98	27600
macro avg	0.98	0.98	0.98	27600
weighted avg	0.98	0.98	0.98	27600

(d) Classification Report

Fig. 7: Model 1: Results

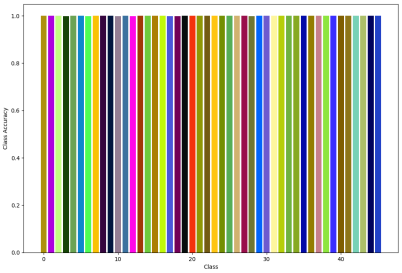


(a) Bar Plot of 46 class accuracy

	precision	recall	f1-score	support
0	0.99	0.98	0.99	590
1	0.99	0.99	0.99	605
2	0.99	0.99	0.99	572
3	0.96	0.97	0.97	601
4	1.00	0.96	0.98	597
5	0.98	1.00	0.99	560
6	0.98	0.97	0.98	597
7	0.98	0.99	0.99	631
8	1.00	0.99	0.99	618
9	0.99	0.99	0.99	609
10	0.99	0.99	0.99	600
11	0.99	0.99	0.99	601
12	0.97	0.98	0.98	598
13	0.98	0.98	0.98	589
14	0.99	0.99	0.99	558
15	0.98	0.99	0.98	604
16	0.96	0.96	0.96	643
17	0.96	0.98	0.97	588
18	0.97	0.98	0.97	592
19	0.98	0.99	0.99	626
20	0.97	0.99	0.98	627
21	0.99	1.00	0.99	589
26	0.99	0.99	0.99	584
27	0.99	0.98	0.98	590
28	0.96	0.96	0.96	631
29	0.99	0.99	0.99	590
30	0.98	0.98	0.98	641
31	0.98	0.98	0.98	595
32	0.98	0.98	0.98	581
33	0.99	0.97	0.98	624
34	0.99	0.99	0.99	593
35	0.99	0.98	0.99	622
36	1.00	1.00	1.00	624
37	0.99	0.99	0.99	611
38	0.98	0.99	0.98	597
39	0.99	0.99	0.99	574
40	1.00	1.00	1.00	613
41	0.99	0.99	0.99	574
42	0.98	0.99	0.99	596
43	0.99	0.99	0.99	577
44	0.99	0.99	0.99	586
45	1.00	0.99	1.00	599
accuracy			0.98	27600
macro avg	0.98	0.98	0.98	27600
weighted avg	0.98	0.98	0.98	27600

(b) Classification Report

Fig. 8: Model 2: Results



(a) Bar Plot of 46 class accuracy

	precision	recall	f1-score	support
0	0.98	0.99	0.99	590
1	0.99	0.99	0.99	605
2	0.99	0.99	0.99	572
3	0.97	0.97	0.97	601
4	0.99	0.98	0.98	597
5	0.97	0.99	0.98	560
6	0.98	0.97	0.97	597
7	0.99	0.99	0.99	631
8	0.98	0.99	0.98	618
9	0.98	0.99	0.99	609
10	0.99	0.98	0.98	600
11	0.99	1.00	0.99	601
12	0.98	0.98	0.98	598
13	0.98	0.98	0.98	589
14	0.99	0.99	0.99	558
15	0.98	0.99	0.98	604
16	0.96	0.96	0.96	643
17	0.96	0.98	0.97	588
18	0.97	0.98	0.97	592
19	0.98	0.99	0.99	626
20	0.97	0.99	0.98	627
21	0.99	0.99	0.99	589
22	0.97	0.97	0.97	606
23	0.98	0.98	0.98	596
24	0.98	0.98	0.98	615
25	0.98	0.89	0.94	586
26	0.99	0.97	0.98	584
27	0.99	0.97	0.98	590
28	0.97	0.95	0.96	631
29	0.96	0.99	0.98	590
30	0.97	0.99	0.98	641
31	0.99	0.98	0.98	595
32	0.98	0.98	0.98	581
33	0.99	0.97	0.98	624
34	0.99	0.99	0.99	593
35	0.99	0.98	0.99	622
36	1.00	1.00	1.00	624
37	0.99	0.99	0.99	611
38	0.98	0.99	0.98	597
39	0.99	0.99	0.99	574
40	1.00	1.00	1.00	613
41	0.99	0.99	0.99	574
42	0.98	0.99	0.99	596
43	0.99	0.99	0.99	577
44	0.99	0.99	0.99	586
45	1.00	0.99	1.00	599
accuracy			0.98	27600
macro avg	0.98	0.98	0.98	27600
weighted avg	0.98	0.98	0.98	27600

(b) Classification Report

Fig. 9: Model 3: Results

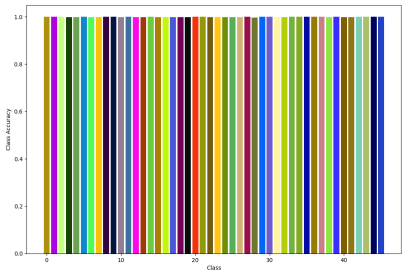
1) Bar Container Plot: Most of the bars are at similar heights, and the model is showing accuracy very close to 1, the model is performing quite well and has fit the data well (Fig. 10a).

2) Classification Report: Again, the precision, F1-score, and recall are extremely near to 1.0, which suggests that the model is making accurate predictions and learning data well (Fig. 10b).

TABLE II: Comparison of results

S. No.	Method	Accuracy Score
1.	CNN based Neural Network	97.78%
2.	. CNN based Random Forest Classifier	98.40%
3.	MLP-Mixer	98.30%
4.	CNN based KNN Classifier	98.70%

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 06:04:34 UTC from IEEE Xplore. Restrictions apply.



(a) Bar Plot of 46 class accuracy

	precision	recall	f1-score	support
0	1.00	0.99	0.99	590
1	1.00	0.99	0.99	605
2	0.99	1.00	1.00	572
3	0.98	0.98	0.98	601
4	1.00	0.97	0.98	597
5	0.98	1.00	0.99	560
6	0.99	0.99	0.99	597
7	0.99	1.00	0.99	631
8	1.00	1.00	1.00	618
9	0.99	0.99	0.99	609
10	0.99	0.99	0.99	600
11	0.99	0.99	0.99	601
12	0.98	0.99	0.98	598
13	0.98	0.97	0.98	589
14	0.99	0.99	0.99	558
15	0.99	0.99	0.99	604
16	0.96	0.97	0.97	643
17	0.96	0.98	0.97	588
18	0.98	0.98	0.98	592
19	0.98	0.99	0.99	626
20	0.98	0.99	0.99	627
21	0.99	1.00	0.99	589
22	0.97	0.97	0.97	606
23	0.98	0.98	0.98	596
24	0.98	0.99	0.99	615
25	0.97	0.97	0.97	586
26	0.99	0.99	0.99	584
27	0.99	0.98	0.99	590
28	0.96	0.97	0.96	631
29	1.00	0.99	1.00	590
30	0.98	0.99	0.98	641
31	0.99	0.99	0.99	595
32	0.98	0.98	0.98	581
33	0.99	0.98	0.99	624
34	0.99	0.99	0.99	593
35	1.00	0.98	0.99	622
36	1.00	1.00	1.00	624
37	0.99	1.00	0.99	611
38	0.99	0.99	0.99	597
39	1.00	0.99	1.00	574
40	1.00	1.00	1.00	613
41	1.00	0.99	1.00	574
42	0.99	1.00	0.99	596
43	1.00	1.00	1.00	577
44	0.99	1.00	0.99	586
45	1.00	1.00	1.00	599
accuracy			0.99	27600
macro avg	0.99	0.99	0.99	27600
weighted avg	0.99	0.99	0.99	27600

(b) Classification Report

Fig. 10: Model 4: Results

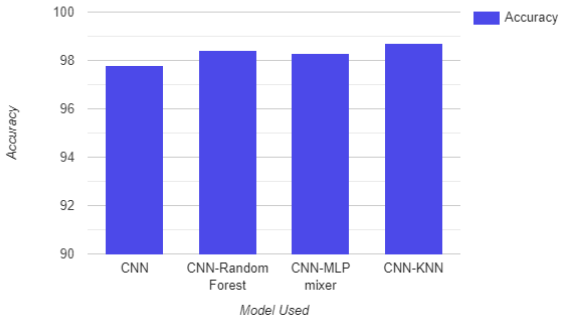


Fig. 11: Comparison plot of model performance in this study

V. Discussion

The results demonstrate that the KNN model coupled with CNN feature extraction achieved the highest accuracy of 98.7% (Table II). The CNN-based Random Forest Classifier and CNN-based MLP-Mixer classifier also performed well, with accuracies of 98.4% and 98.3% respectively (Fig. 11). These findings highlight the effectiveness of CNN-based feature extraction in achieving high accuracy in Devanagari character recognition tasks. The CNN-based deep neural network, although slightly lower in accuracy at 97.78%, still demonstrated strong performance. These results suggest that deep neural networks remain a viable approach for Devanagari character recognition, albeit with a slightly lower accuracy compared to the hybrid models. Overall, the KNN model with CNN feature extraction offers the highest accuracy. However, the Random Forest and MLP-Mixer classifiers also provide comparable accuracy and can be suitable alternatives depending on specific requirements or considerations.

VI. Future Work

We aim to develop our own dataset having images that would have some disturbances and would be different from the current dataset having clear images for all characters. By developing such a dataset we could aim to create a more foolproof model which would work well on a variety of data. Also, we are trying to include blurred, distorted images in our dataset, so that our model trains and works effectively in all circumstances. We would further develop models that would fit our own dataset and provide more satisfactory results in other Indian Languages.

References

[1] U. Pal, T. Wakabayashi, and F. Kimura, "Comparative study of devnagari handwritten character recognition using different feature and classifiers," in *2009 10th International Conference on Document Analysis and Recognition*, pp. 1111–1115, 2009.

[2] B. Dessai and A. Patil, "A deep learning approach for optical character recognition of handwritten devanagari script," in *2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, vol. 1, pp. 1160–1165, 2019.

[3] V. Agnihotri, "Offline handwritten devanagari script recognition," *International Journal of Information Technology and Computer Science*, vol. 4, 07 2012.

[4] S. Acharya, A. K. Pant, and P. K. Gyawali, "Deep learning based large scale handwritten devanagari character recognition," in *2015 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, pp. 1–6, 2015.

[5] N. Aneja and S. Aneja, "Transfer learning using cnn for handwritten devanagari character recognition," in *2019 1st International Conference on Advances in Information Technology (ICAIT)*, pp. 293–296, 2019.

[6] S. Arora, D. Bhattacharjee, M. Nasipuri, L. Malik, M. Kundu, and D. K. Basu, "Performance comparison of svm and ann for handwritten devnagari character recognition," 2010.

[7] R. Guha, N. Das, M. Kundu, M. Nasipuri, and K. C. Santosh, "Devnet: An efficient cnn architecture for handwritten devanagari character recognition," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 34, no. 12, p. 2052009, 2020.

[8] M. Jangid and S. Srivastava, "Handwritten devanagari character recognition using layer-wise training of deep convolutional neural networks and adaptive gradient methods," *Journal of Imaging*, vol. 4, no. 2, 2018.

[9] S. R. Narang, M. Kumar, and M. K. Jindal, "Deepnetdevanagari: a deep learning model for devanagari ancient character recognition," *Multimedia Tools and Applications*, vol. 80, pp. 20671–20686, 2021.

[10] N. Singh, "An efficient approach for handwritten devanagari character recognition based on artificial neural network," in *2018 5th International Conference on Signal Processing and Integrated Networks (SPIN)*, pp. 894–897, 2018.

[11] S. Ahlawat, A. Choudhary, A. Nayyar, S. Singh, and B. Yoon, "Improved handwritten digit recognition using convolutional neural networks (cnn)," *Sensors*, vol. 20, no. 12, 2020.

[12] S. K. Manocha and P. Tewari, "Devanagari handwritten character recognition using cnn as feature extractor," in *2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*, pp. 1–5, IEEE, 2021.

[13] I. Dokare, S. Gadge, K. Kharde, S. Bhere, and R. Jadhav, "Recognition of handwritten devanagari character using convolutional neural network," in *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*, pp. 353–359, IEEE, 2021.

[14] A. Moudgil, S. Singh, and B. Sareen, "Cnn-lstm based approach for recognition of devanagari manuscripts," in *2022 International Conference on Data Analytics for Business and Industry (ICDABI)*, pp. 1–5, IEEE, 2022.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 06:04:34 UTC from IEEE Xplore. Restrictions apply.

[15] M. S. V. Dharsini, N. Kamath, Y. Chaudhary, A. Sriramoju, and N. Anurag, "Devanagri character image recognition and conversion into text using long short term memory," in *2022 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI)*, vol. 1, pp. 1–5, IEEE, 2022.
[16] L. Abhishek, "Optical character recognition using ensemble of svm, mlp and extra trees classifier," in *2020 International Conference for Emerging Technology (INCET)*, pp. 1–4, IEEE, 2020.
[17] S. Patnaik, S. Kumari, and S. D. Mahapatra, "Comparison of deep cnn and resnet for handwritten devanagari character recognition," in *2020 IEEE 1st International Conference for Convergence in Engineering (ICCE)*, pp. 235–238, IEEE, 2020.
[18] M. K. Gupta, S. Vikram, S. Dhawan, and A. Kumar, "Handwritten ocr for word in indic language using deep networks," in *2023 10th International Conference on Signal Processing and Integrated Networks (SPIN)*, pp. 389–394, IEEE, 2023.
[19] P. Malik and M. K. Parida, "Devanagari characters recognition using deep neural networks," in *2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT)*, pp. 1–5, IEEE, 2022.
[20] I. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, M. Lucic, and A. Dosovitskiy, "Mlp-mixer: An all-mlp architecture for vision," 2021.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 06:04:34 UTC from IEEE Xplore. Restrictions apply.

Vol. 01, Iss. 2 (2023) 62-67, DOI: 10.61552/JME.2023.02.002

JME

Journal of Materials and Engineering

www.jme.aspur.rs

Development and Tribological Analysis of Bio-Based Lubricant Using Two Non-Edible Oils: Castor Oil and Neem Oil

Ayan Bindra[a,*], Dipali Yadav[a], Manan Anand[a], R.C. Singh[a]

[a]*Department of Mechanical Engineering, Delhi Technological University, Bawana Road, Shahbad Daulatpur Village, Rohini, Delhi – 110042, India.*

Keywords:

Tribology
Biotribology
Wear
Friction
Lubricants

* Corresponding author:

Ayan Bindra
E-mail: ayanbindra@gmail.com

Received: 21 February 2023
Revised: 23 March 2023
Accepted: 26 April 2023

A B S T R A C T

Lubricants play a very important role in multiple industries in minimizing wear and friction between two moving parts. Most lubricants are produced from petroleum oil which is poorly biodegradable and toxic in nature. Hence these lubricants are highly undesirable due to environmental hazards and the health and safety of the operator. Bio-Lubricants like Castor Oil have found a potential scope in research and are being used in various applications as a substitute for these chemical lubricants. This paper examines the Tribological Properties of developed Bio-Lubricants, prepared using Castor Oil and Neem Oil in a mixing ratio of 20%. Wear and Friction analysis has been carried out using Pin on Disc apparatus at varying loads, which has shown that the blend of Castor Oil as a base with 20% Neem Oil has promising properties, with around 70% reduction in wear rate at specific loads, hence showcasing a tremendous capacity for being used in maintenance applications. An overall reduction in wear and coefficient of friction was observed in the developed mixture, making it a potential environment-friendly replacement for chemical lubricants and an upgrade over Pure Refined Castor Oil.

© 2023 Published by ASPUR

1. INTRODUCTION

1.1 Lubricants

A lubricant is a substance used to reduce friction and wear between two surfaces that are in contact with each other. It can be a liquid, such as oil or water, or a solid, such as grease or graphite [1]. In the mechanical engineering industry, lubricants are used to prevent the failure of moving parts in machines and equipment due to excessive wear, overheating, or seizure.

The choice of lubricants and their properties, such as viscosity, chemical composition, and

additives, plays an important role in engineering applications [2]. The lubricant should have the right viscosity to form a protective film between the moving surfaces without being too thick to impede movement or too thin to evaporate or leak. Chemical compatibility of lubricant with the surface material it is used upon is important. It must hold out against corrosion, oxidation and contamination.

Moreover, modification of lubricants can be done to suit certain applications by incorporating additives, such as anti-wear agents, friction modifiers, rust inhibitors, and extreme pressure agents. These substances improve the lubricants' efficiency under high temperatures and pressure, or other extreme circumstances.

Industries like automotive, manufacturing, aerospace, and energy depend heavily on lubricants for smooth and efficient operation. Hence, lubricants play a significant role in ensuring optimal system performance.

1.2 Bio Lubricants

Escalating environmental issues have impacted many industries. Conventionally used petroleum-based lubricants harm the environment after it is released into the waterways and soil. Thus, there's a huge demand for biodegradable lubricants in the mechanical engineering industry.

Biodegradable lubricants are generally made from renewable materials like animal fats or vegetable oils. These can be easily decayed by microorganisms in the soil or water. Because of their fast and safe breakdown in the environment, they do not pose to be harmful to wildlife and surrounding, keeping it pollution-free [3].

Castor oil is a vegetable oil extracted from the seeds of the Ricinus communis, the scientific name of the castor plant. It belongs to the Euphorbiaceae family. Castor plants are non-edible, widespread, and well-known for their therapeutic applications as laxatives. In India, castor plants are widely available, and the seeds of these plants contain 40 to 60 percent oil by volume [4]. Castor oil benefits from the presence of the hydroxyl group, which makes it more oxidatively stable than its competitors and suitable for use as a base lubricant [5].

Castor Oil has been highly researched as an individual substitute to petro-based lubricants and has shown promising results in decreasing wear in mechanical applications as a substitute for present industry-grade lubricants.

Neem oil is a type of vegetable oil that is derived from the seeds and fruits of the neem tree, which is an evergreen tree that is commonly found in India. The botanical name of neem is Azadirachta indica and it belongs to Meliaceae, the mahogany family, This oil typically has a reddish-brown or greenish-brown colour [6].

The properties of Castor Oil and Neem Oil have been mentioned in Table 1.

Table 1. Properties of Oils used.

S. No.	Characteristics	Castor Oil	Neem Oil
1	Colour	Pale Yellow	Brown Yellow
2	Density (g/ml)	0.95	0.89
3	Viscosity at 30°C (mm^2/s)	273.1	36.7
4	Iodine Value	83	72
5	Flash Point (°C)	230	187

2. TESTING SETUP

2.1 Pin on Disc Apparatus

A Pin on Disc apparatus (depicted in Fig 1) has been used to perform this experimental study. The Pin on Disc machine can be used to determine the tribological properties of a wide range of materials under varying normal loads. A revolving disc is contacted by a stationary pin positioned on a pin holder moving at a predetermined speed. When lubricating oil is present, the pin slides and exerts a frictional force against the disc.

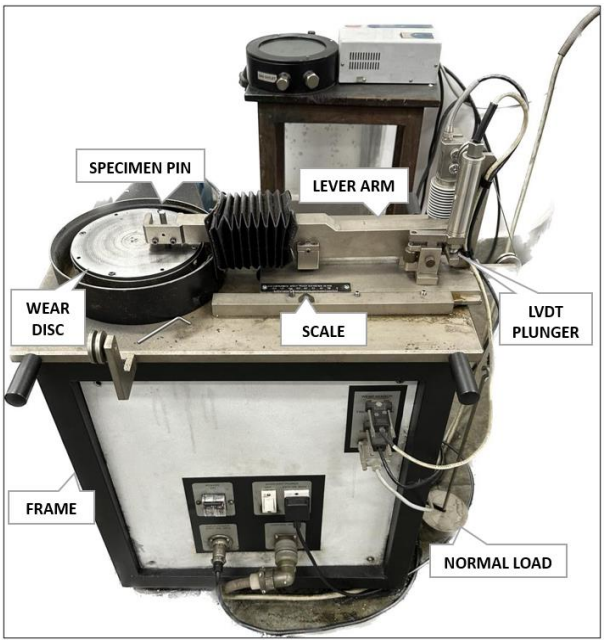


Fig. 1. Pin on Disc Apparatus.

2.2 Test Specimen

A typical cylindrical-shaped specimen has been taken for analysis, having a length of 30 mm and a pin diameter of 8 mm to perform the experiment on the Pin On Disc apparatus. Since rough surfaces make it harder to quantify wear, the pin's surface has been polished for accuracy.

Brass is the material used to make the testing pin used in this experiment.

2.3 Test Parameters

The following test parameters have been considered in this experiment:

1. Normal Load: For this experiment, a load variation of 10N to 40N has been considered, with a gap of 10N in each consecutive test.
2. Speed: The relative sliding speed between the pin and disc has been taken as 1000 RPM in each case, while the load varies.
3. Time: The running time of the experiment was 20 minutes at a selected speed of rotation, with a load increment of 10N every 5 minutes.
4. Running Diameter: The diameter of the track at which the pin rotates on the disc is taken as 80mm for this experimental study.
5. Flow rate: A constant flow of oil has been maintained throughout the experiment, with oil being pushed onto the disc at an interval of every 10 seconds.

2.4 Lubricants Used

Two lubricant blends have been used in this experiment to validate and compare their tribological performance.

Pure Neem Oil has been used to establish a baseline for Wear and Friction characteristics at different loads and operating conditions.

To further enhance the properties of the baseline, a blend of Pure Castor Oil has been created with Pure Neem Oil (Blend 1). The mixing has been done in the ratio of 80:20 since previous studies have indicated the most optimal performance to be achieved at a mixing ratio of 20%.

The various mixing ratios used in this experiment have been indicated in Table 2.

Table 2. Oil Blends Used for Analysis.

S.No.	Refined Castor Oil (mL)	Refined Castor Oil (mL)	Mixing Ratio (%)
1	1000	0	0
2	800	200	20

2.4.1 Creation of Blend

For this experiment, a base oil volume of 800mL was taken. Blend 1 was created in a ratio of 80:20 by mixing 200mL of Neem Oil with 800mL of Castor (base) oil in a magnetic stirrer. A total blend of 1000mL was achieved as a result of mixing.

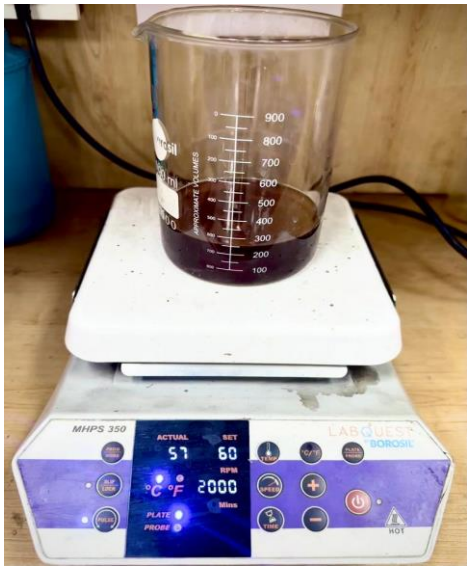


Fig. 2. Mixing blends using a magnetic stirrer.

The mixing in the magnetic stirrer was done at 2000 RPM, while the temperature was maintained at 60°C for 20 minutes.

The magnetic stirrer used for the experiment and mixing operation of the oil has been depicted in Fig 2.

3. ANALYSIS OF CHARACTERISTICS

At the point where a pin and a disc make contact, a beginning load of 10N was applied, and an oil sample flow was maintained continuously. The disc rotates at 1000 RPM (at a specific load) for a time period of 5 minutes. The coefficient of friction and wear was then measured for the specific loads and speed. Additional testing was done with increasing weights up to 40N, in a step difference of 10N each.

Testing for both the oil samples, i.e, pure castor oil and castor oil blended with neem oil, was done in the same physical conditions. Based on the experiments, the logarithmic curves showing the impact of load in N on wear rate in mm3/Nm and coefficient of friction were generated for the respective samples using the Pin on Disc apparatus. Finding an oil sample that benefits the following tribological qualities and improves lubrication is the goal of this research.

3.1 Wear Characteristics

The potential of formed oil samples was discovered for each varying load and a comparison was done for oil at high and low loading conditions to simulate mechanical applications in the industry. The graphs for both oil blends clearly show that for all oils, the wear rate increased as the applied load was increased.

The wear was tested at loads of 10N, 20N, 30N, and 40N for both oils at the constant speed of 1000 RPM. The recorded value distinctly indicated that castor oil blended with non-edible Neem Oil has better wear characteristics than refined castor oil.

A comparison graph comparing the wear was plotted for 10N and 40N loads, which has been shown in Fig 3.

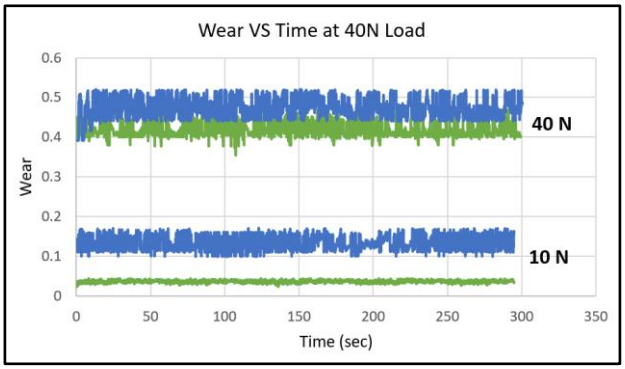


Fig. 3. Comparison graph of wear with time for oils at 10N and 40N load.

At an applied load of 10N a decrease of 70.17 % in wear was observed in the castor and neem oil blend. Whereas for the 40N load, the decrease in wear was noted to be 12.76 % in the castor and neem oil blend. From the observations, it can be stated that the neem oil when blended with castor oil in a 20% mixing ratio, proves to be a good lubricant resisting wear better than refined castor oil. Furthermore, the wear-resisting property of the Neem-Castor Blend is better at low load compared to high load but is still marginally better than castor oil alone.

3.2 Coefficient of Friction

For any equipment to perform properly and reduce energy losses, a lower coefficient of friction is preferred. As expected, a larger load led to a high coefficient of friction. More rubbing of the pin at higher loads resulted in a higher value of COF.

The figures of the coefficient of friction of pure castor oil and a blend of pure castor and neem oils (indicated in Fig 4), it is seen the latter has a lower coefficient of friction compared to pure castor oil. This is because the Neem oil in the mixture has the propensity to reduce friction.

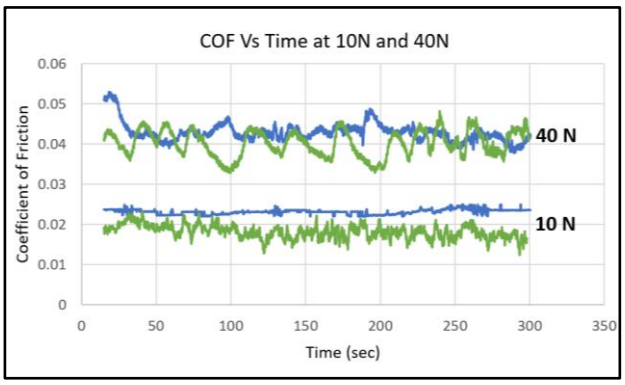


Fig. 4. Comparison graph of COF with time for oils at 10N and 40N load.

Castor-Neem blend's coefficient of friction readings range from 0.01 to 0.02 at 10N load, which is a solid indication that it may be utilized as a lubricant in mechanical applications. An average reduction of 7 to 21 % was observed by using the developed blend.

4. RESULTS AND DISCUSSION

The experimental values demonstrate that the created lubricant Blend 1 consisting of Refined Castor and Neem Oil in a 20% ratio depicts better lubrication performance and wear characteristics than Pure Castor Oil. The percentage difference in properties of Wear and Coefficient of Friction compared between Pure Castor and Blend has been described in Table 3.

Table 3. Percentage Difference in Performance.

% reduction in Wear in Blend 1 (20% Neem Oil) as compared to Pure Castor Oil	
10N (Low Load)	**40N (High Load)**
70.17	12.76
% reduction in Coefficient of Friction in Blend 1 (20% Neem Oil) as compared to Pure Castor Oil	
10N (Low Load)	**40N (High Load)**
21.64	6.29

Certain findings from the above experimental work are as follows:

- For the Castor-Neem mixture, the Coefficient of Friction was found having a value less than 0.02, indicating a standard reduction of 21.64% over the Pure Castor blend at a low load of 10N, thus resulting in lower energy losses and enhanced fuel economy when used in the industry.
- Average 70.17% wear reduction has been observed in Castor-Neem Blend over pure Castor Oil at low loads, thus indicating an increase in efficiency, an indicator of the life of the mechanical design, and reduced mechanical wear of the machinery. Enhanced anti-wear properties increase the service life of the mechanism, and result in increased longevity and longer service intervals of the machinery.

5. COST ANALYSIS

A cost analysis has been performed for the developed Bio-Lubricant to ensure that it is viable for industrial usage as per given standards.

The cost has been compared to the Servo Gear Oil 90T, which is the nearest industry standard most widely used for similar applications, and easily accessible in the market.

Servo gear oil 90T considered in the analysis has a unit cost of Rs. 220 per litre in the market. The Castor and Neem oil used for the bio-lubricant preparation is easily available in the local market at a bulk cost of Rs. 100 per liter for Pure Castor and Rs. 500 per liter for Neem oil. As per experimental observations, the blend of 80% Castor oil + 20% Neem oil has been seen to have comparable characteristics to the Servo oil.

Considering commercial production has enhanced properties for mechanical applications, including an additional cost of Rs. 40 to 80, the end user cost would still be lower than the servo-grade oil. The developed bio-lubricant has an upper hand over the Servo oil, after considering the whole preparational cost, i.e., operational, environmental harm and maintenance charge.

Castor oil-based lubricants have an enormous capacity to restore environmental balance and reduce operational costs in the mechanical industry.

6. CONCLUSIONS

The purpose of this paper is to evaluate the tribological characteristics of Castor Oil and its combination with Neem Oil. Through experimentation and analysis, several significant findings were uncovered, which have been elaborated upon.

- The research conducted made significant progress by examining how refined neem oil can be used as an additive to refined castor oil to enhance its lubrication properties, by decreasing wear and coefficient of friction. The study hence discovered that blends of refined castor oil and refined neem oil possess great potential for utilization in mechanical applications.

- The coefficient of friction has been measured to be consistently below 0.02, for Castor-Neem Blend. At a low load of 10N, the blend demonstrated a 21% reduction. This decrease in friction leads to several benefits, including improved reliability, better fuel economy, and increased performance in mechanical applications. Additionally, the reduced friction results in cost savings by lowering maintenance expenses. As the friction decreases, the wear and tear on the machine are also minimized, which is expected to lead to reduced maintenance costs over time.
- A significant reduction in wear, of about 70.1% was seen for the developed Castor-Neem Blend at a load of 10N. A wear reduction consequently leads to an increase in mechanical operating efficiency and extends the design life of the application.
- A greater reduction in wear and coefficient of friction was noted at a low load of 10N compared to a high load of 40N, indicating that the developed Castor-Neem blend might work better at lower loads.
- In all the load testing cases, the performance of the developed Castor-Neem Blend was found to be better than Refined Castor Oil. The developed blend is also cheaper than servo-grade oil. Thus, it has the ability to replace conditionally used lubricants. The proposed blend is environment-friendly and has the potential to change future trends. For specialized applications, more research needs to be done in this field.

REFERENCES

[1] J. Denis, J. Briant, J.C. Hipeaux, "Lubricant Properties, Analysis and Testing," Editions Technip, pp. 107-287, 1997.

[2] Y. Singh, A. Farooq, A. Raza, M.A. Mahmood, S. Jain, "Sustainability of a nonedible vegetable oil based bio-lubricant for automotive applications: a review," *Process Saf. Environ. Prot*, vol. 111, pp. 701–713, 2017.

[3] P.S. Chauhan, V.K. Chhibber, "Non-edible oil as a source of bio-lubricant for industrial applications: a review," *International Journal of Engineering Science and Innovative Technology*, vol. 2, pp. 299-305, 2013.

[4] E.O. Aluyor, K.O. Obahiagbon, M.O. Jesu, "Biodegradation of vegetable oils: a review," *Scientific Research and Essay*, vol. 4, pp. 543-548, 2009.

[5] A.K. Singh, "Castor oil-based lubricant reduces smoke emissions in two-stroke engines," *Ind. Crop. Prod*, vol. 33, pp. 287–295, 2011.

[6] C. Sharma, G.S. Sundar, H. Babu, S.M. Sitale, B.S. Smitha, "(Experimental Study on The Use of Neem Oil as Lubricant in IC Engine," *International Research Journal of Engineering and Technology*, vol. 4, pp. 1087-1092, 2017.

[7] "ASM Handbook: Friction, Lubrication and Wear Technology," ASM International, U.S.A., Vol.18, 1992.

[8] B.S. Chauhan, N. Kumar, H.M. Cho, H.C. Lim, "A study on the performance and emission of a diesel engine fueled with Karanja biodiesel and its blends," *Energy*, vol. 56, pp. 1–7, 2013.

[9] A.K. Jain, A. Suhane, "Research approach & prospects of non-edible vegetable oil as a potential resource for bio lubricant - a review," *Advanced Engineering and Applied Sciences: An International Journal*, vol. 1, pp. 23–32.

[10] A.K. Jain, A. Suhane, "Investigation of Tribological Characteristics of Non-Edible Castor and Mahua Oils as Bio Lubricant for Maintenance Applications," in Proc. 5th Int. & 26th All India Manufacturing Technology Con. on Design and Research (IIT Guwahati), pp. 1–6, 2012.

[11] Y. Singh, A. Sharma, N. Singh, A. Singla, P.M. Rastogi, "Prospects of inedible plant oil-driven bio-lubricants for tribological characteristics – a review," *Int. J. Ambient Energy*, pp. 1–14, 2018.

[12] R.C. Singh, R. Chaudhary, A. Kumar, "Tribological Analysis of novel Apricot oil based Biolubricant against 15W40 oil tested on High-Temperature Tribometer," *Journal of Engg. Research ICARI*, Special Issue pp. 205-214, 2021.

ECS Journal of Solid State
Science and Technology

The Electrochemical Society
Advancing solid state & electrochemical science & technology

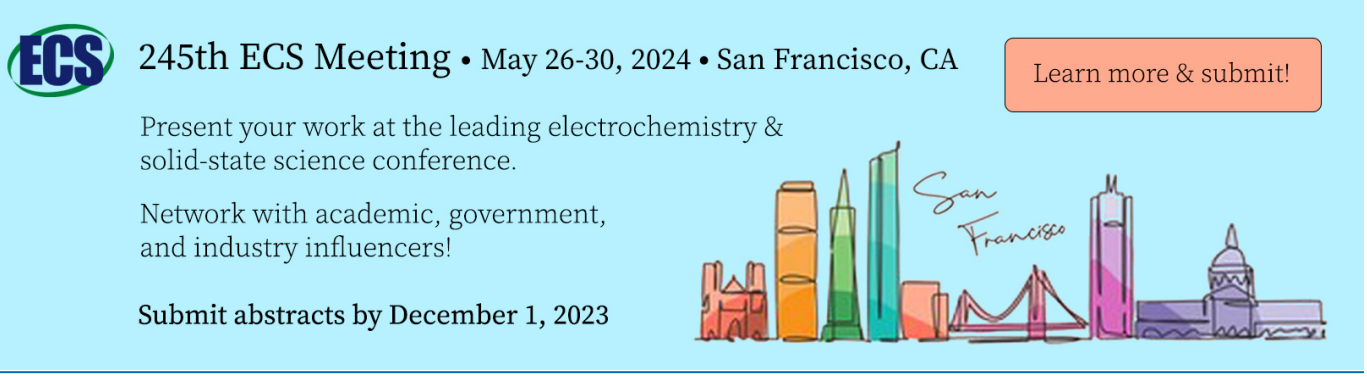
Development of Conducting Biopolymer-Based Biosensor for Heavy-Metal Ion Detection

To cite this article: Shreya Dutta *et al* 2023 *ECS J. Solid State Sci. Technol.* **12** 117001

View the article online for updates and enhancements.

You may also like

- Surface modification of silicone tubes by functional carboxyl and amine, but not peroxide groups followed by collagen immobilization improves endothelial cell stability and functionality
Nasim Salehi-Nik, Ghassem Amoabediny, Mohammad Ali Shokrgozar et al.
- Volume fraction determination of microgel composed of interpenetrating polymer networks of PNIPAM and polyacrylic acid
S Franco, E Buratti, B Ruzicka et al.
- Grafting of acrylic acid onto microwave plasma-treated polytetrafluoroethylene (PTFE) substrates
John Kenneth C. Valerio, Hideki Nakajima and Magdaleno R. Vasquez



This content was downloaded from IP address 14.139.251.98 on 03/11/2023 at 05:19

ECS Journal of Solid State Science and Technology, 2023 **12** 117001
2162-8777/2023/12(11)/117001/8/$40.00 © 2023 The Electrochemical Society ("ECS"). Published on behalf of ECS by IOP Publishing Limited

Development of Conducting Biopolymer-Based Biosensor for Heavy-Metal Ion Detection

Shreya Dutta,[1] Ravi Butola,[2] Bhupendra Pratap Singh,[3] and Prakash Chandra[1,z]

[1]*Department of Biotechnology, Delhi Technological University, Delhi 110042, India*
[2]*University School of Automation & Robotics (USAR), GGSIP University, East Delhi Campus, Delhi-110092, India*
[3]*Department of Environmental Studies, Central University of Haryana, Jant-Pali, Mahendergarh, Haryana-123031, India*

The increasing concentration of heavy metals in the water ecosystems has been a noteworthy concern nowadays. This has accounted for various diseases, due to the bioaccumulation in the human body. We propose a biocompatible, non-toxic technique for the identification of heavy metal ions from the contaminated aquatic samples. A urease biosensor, having interpenetrating network of Katira-cl-Acrylic Acid (AAc)/Polyaniline (PANI) hydrogel composite deposited onto the indium tin oxide (ITO). Enhanced hydrogel conductivity due to the incorporation of PANI has been gained by the interfacial polymerization of katira gum and aniline monomer in the ratio of 2:1 using N,N′-MBA and APS as cross-linker and initiator respectively. Structural, morphological and electrochemical characterization of PANI/ITO, Katira-cl-AAc/PANI/ITO and Urease/Katira-cl-AAc/PANI/ITO electrode conducted using SEM, FT-IR, TGA, UV spectrometry, cyclic voltammetry. A comparative study of different concentrations of cadmium, cobalt and Mercury, has been done. The inhibition rates were found to be in the order of $Hg^{2+} > Cd^{2+} > Co^{2+}$. Enzyme inhibition observed in the concentration range from 1.8 mM to 5 mM, 0.1 mM to 2 mM and 0.1 mM to 5 mM respectively. Urease/Katira-cl-AAc/PANI/ITO, exhibits greater stability, enhanced sensitivity and wide-range of detection limit (200 to 270 ng l^{-1}) with a shelf life.
© 2023 The Electrochemical Society ("ECS"). Published on behalf of ECS by IOP Publishing Limited. [DOI: 10.1149/2162-8777/ad041d]

Manuscript submitted June 7, 2023; revised manuscript received August 31, 2023. Published November 1, 2023. *This paper is part of the JSS Focus Issue on Sustainable Materials and Devices.*

Increasing concentration of heavy metals is polluting every ecosystem. The cause is due to different anthropogenic activities like agriculture and industrial wastes, leaching of soil etc. Toxicity due to the presence of these heavy metals has posed a negative effect on the aquatic flora and fauna. These enter the food chain through bio-magnification causing adverse effects on human health. Some metals are essential trace elements like copper, however in abundance cause lethality. Other metals like cadmium and Mercury are fatal even in very low concentrations. Humans, when exposed to these metals for long duration can get affected by severe physical, neurological and muscular degeneration causing Alzheimer's disease, Parkinson's disease, muscular dystrophy, and multiple sclerosis. Heavy metals such as CO^{2+}, Cd^{2+} and Hg^{2+} are very toxic and can also cause toxicity in certain organs of the human body, such as neurotoxicity, hepatotoxicity, nephrotoxicity, skin toxicity, and cardiovascular toxicity. Presence of these heavy metal ions in drinking water above the permissible range affects the physical and mental growth of infants and children and causes other chronic diseases. Since these heavy metal ions, above the permissible range is hazardous, hence its detection is major concern for researchers.[1,2] In recent years, many methods have been developed for the determination of the heavy metals in the waste water sources and in drinking water. In light of the expanding interest for the identification of toxicity caused because of heavy metal poisoning, techniques are produced which can identify them in extremely least measure of time i.e., having fast response. Some of the methods which are utilized as a part of detecting include X-ray fluorescence spectrometry,[3] atomic absorption spectrometry,[4] flame atomic absorption spectrometer,[5] anodic stripping voltammetry,[6] inductively coupled mass spectrometry,[7] neutron activation analysis, potentiometry, etc., of which biosensors are picking up frame due to its reusability, specificity, stability, in situ analysis, fast response and cheap screening methods.

Conducting polymers have drawn considerable attention of researchers and scientists for detection of various types of toxic pollutants and heavy metal ions are one of them. Conducting polymers such as polyaniline (PANI), poly(3,4-ethylene dioxythiophene) (PEDOT), polyacetylene (PA), polythiophene (PT), polypyrrole (PPy), and poly(phenylene vinylene) (PPV) and their composites have received great attention, due to their unique characteristics such as electrical conductivity, remarkable mechanical property, easy to synthesis, inexpensive, stability at room temperature good environmental stability and non-toxicity scientists. These conjugated polymer consists of numerous functional groups, these are easily modified into desired structures. Due to delocalization of π-electrons, these polymers are highly conductive. Therefore, the conductive polymers show impressive potential to be used in fabricating heavy metal ion sensors.[8–10]

Various enzymes have been used for the detection of enzyme inhibitors like heavy metals.[11,12] The interaction of the heavy metal ions with the enzyme or the enzyme-substrate complex plays a crucial role for the detection and estimation of the ions. The interaction can be competitive, non-competitive or uncompetitive. The utilization of enzyme for analyzing heavy metal toxicity has made it a sensitive and uncomplicated method for quantification. The strong inhibitory effect of the heavy metal ions on the enzyme activity has been employed for the assessment of these pollutants.[13,14] The immobilization of enzyme has been done on a variety of matrices like metal nanoparticles, conducting polymers, metal oxides, polymeric gels, hydrogels, conducting papers, etc. The non-toxicity of the immobilization matrix is a very crucial factor when experimenting with different materials is taken into consideration.

A wide variety of natural polymers has been used for the synthesis of Inter-penetrating polymer network (IPN) composite hydrogels like chitosan, alginate, starch, cellulose, guar gum, etc. We have incorporated Katira gum considering it high applicability and stability for the development of hydrogel-based sensor. Katira gum is a natural polysaccharide extracted from the plant *Cochlospermum religiosum* from the grooves of the bark. This is very commonly used in the in pharmaceuticals and food industries as an emulsifier, thickening agent, stabilizer and laxative.[15] It is also used in cigar paste and is exported from India.[16] The gum is used for treating diarrhoea, dysentery, syphilis, pharyngitis, cough, trachoma and gonorrhoea, which utilizes the antipyretic, analgesic and sedative property of the gum.[17] Katira gum is a hetero-polysaccharide having a water-swellable part, which in turn makes it a desirable material for the formulation of hydrogel.[18] The solutions of katira gum are acidic, which ranges between pH 5–6, and is stable over a wide range of pH.[19] The recent development in clinical research shows the use of this gum in formulating microspheres and

[z]E-mail: prakash0111@gmail.com

hydrogels for drug delivery.[20] Considering this, a natural biopolymer-based hydrogel used as a framework for immobilization of the enzyme in the fabrication of the biosensor is an attractive alternative.

Hydrogels have gained a large importance as an immobilization matrix in biosensors because of its different porosity and large swelling index, which is very crucial for the adsorption of the biorecognition elements onto the surface. Hydrogels have been used in the biomedical field since the early 50 s.[21] The three-dimensional (3D) structure of the hydrogel has a lot of similarity to the natural tissues and thus can mimic as an extracellular matrix for tissue growth and other biomedical applications like drug delivery, wound dressing, contact lenses, bone cement, catheters, tourniquets and biosensors.[22] This application has led to their parallel development as biorecognition layer of amperometric, potentiometric, conductometric and fibre-optic based biosensors.[23–27] Since stability, conductivity, surface modification capability and toxicity are major concern for hydrogel-based sensors, we have applied Katira gum for the purpose. Katira gum has a good film forming ability, large surface area, biocompatibility, non-toxic and fascinating electron transport ability due to the incorporation of polyaniline. Exploiting the properties of PANI, hybrid nanocomposite-based electrode having thermal, electrical and tuneable mechanical properties can be prepared with a low production cost that can be further used for the construction of a biosensor.[28–30]

We report the comparative study of urease activity inhibition by three heavy metal ions on a PANI incorporated katira gum hydrogel, coated onto the ITO electrode as a transducer for the construction of an electrochemical biosensor (Fig. 1).

Materials and Methods

Materials.—Crude Katira gum was obtained from the commercial market in New Delhi. Enzyme Urease from *Canavalia ensiformis* (Jack Bean), Ammonium persulfate (APS) and *N,N′* -Methylenebis(acrylamide) (NN-MBA) were purchased from Sigma-Aldrich. The following salts of analytical grade were used as a source for heavy metal ions: $CoCl_2.6H_2O$, $HgCl_2$ and $CdCl_2.H_2O$. All other reagents were obtained commercially and were of analytical grade.

Methods

Synthesis of conducting hydrogel composite.—6% w/v of katira gum was dissolved in Millipore water with the help of a magnetic stirrer at room temperature. 0.874 M acrylic acid was added to the mixture. Varying concentrations of NN-MBA as cross-linker and APS as initiator were then mixed into the solution until a homogeneous mixture was obtained. 0.0389 M and 0.0263 M were further chosen as standard concentration for the formulation of the hydrogel. The solution was then heated to 65 °C for 2 h in a water bath. The cross-linked hydrogel was stirred in distilled water to remove any soluble fraction. It was then dried until a constant weight was obtained. Hydrogel was immersed in an aqueous solution of aniline for 15 to 20 min under constant stirring. The hydrogel was then immersed in an oxidising solution (1.8 ml of hydrochloric acid and 0.45 gms of APS in 20 ml of distilled water) and left undisturbed for half an hour. A colour change from brown to dark green showed the incorporation of the polyaniline in the hydrogel and the formation of emeraldine salt. The PANI incorporated hydrogel was washed with methanol and kept for further use.[31]

Fabrication of biosensing platform.—The powdered conducting hydrogel Katira-cl-AAc/PANI was dispersed in DMSO solution in 1 mg ml^{-1} concentration. This solution was then sonicated for 25–30 min in the sonicator bath to achieve proper dispersion. 20 μl of the solution was pipette out and was drop cast onto the hydrolysed ITO and was allowed to dry completely. This electrode was stored in 4 °C until further used.

Immobilisation of enzyme.—Different concentration of urease (0.5 mg ml^{-1}, 1 mg ml^{-1} and 2 mg ml^{-1}) were made in PBS pH 7. The electrodes were first dipped in 2.5% gluteraldehyde solution for 90 min. The electrodes were then dried at room temperature. 10 μl of urease solution was drop cast over the functionalised electrode and was dried and stored in 4 °C for 24 h.[32]

Physicochemical characterization.—The morphology and the structure of the synthesized product was obtained by using scanning electron microscopy (SEM) [Zeiss EVO 18 Special Edition]. Dried hydrogel samples were fixed on a carbon coated copper grid. The hydrogel samples were gold sputter coated under vacuum before observation.[33,34] The functional groups and bonds present in PANI, PANI/Katira gum hydrogel and hydrogel were investigated through Fourier transform infrared spectroscopy (FTIR) [Thermo Scientific, Nicolet 380]. The Autolab Potentiostat (Netherlands) was used for the electrochemical response studies using a three-electrode system. The fabricated electrode was used as working electrode, Platinum as reference electrode and Ag/AgCl as the counter electrode. The decay of the material with respect to the temperature was carried out using Thermo Gravimetric analysis. The conductivity of the synthesised PANI pellet was measured using the four-probe system. The compound quantitative analysis was done with UV–vis Spectrophotometer [Perkin Elmer].

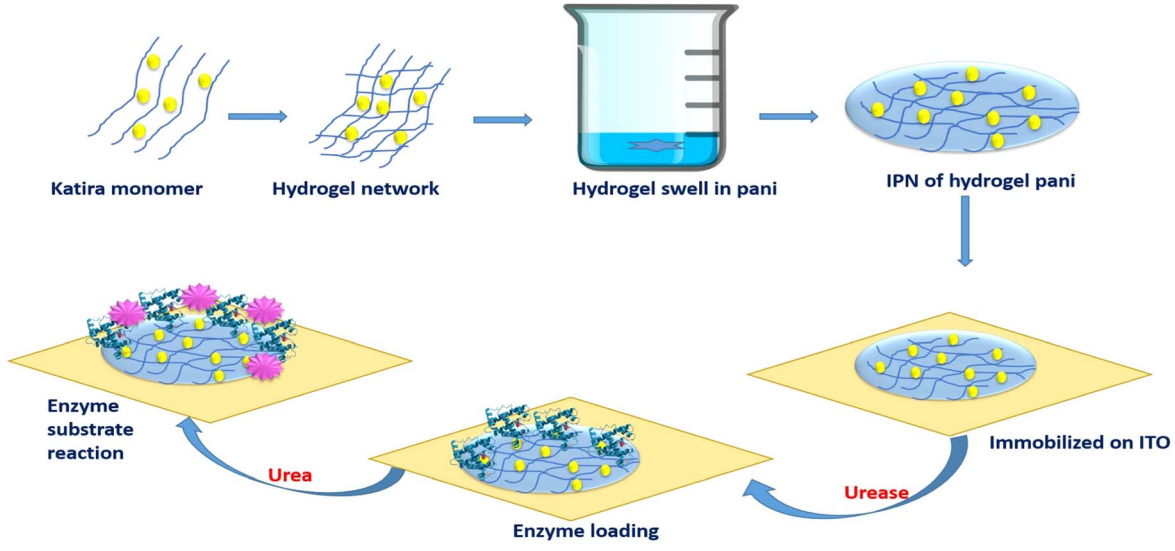


Figure 1. Fabrication step of Urease/KH/PANI-HCl/ITO electrode.

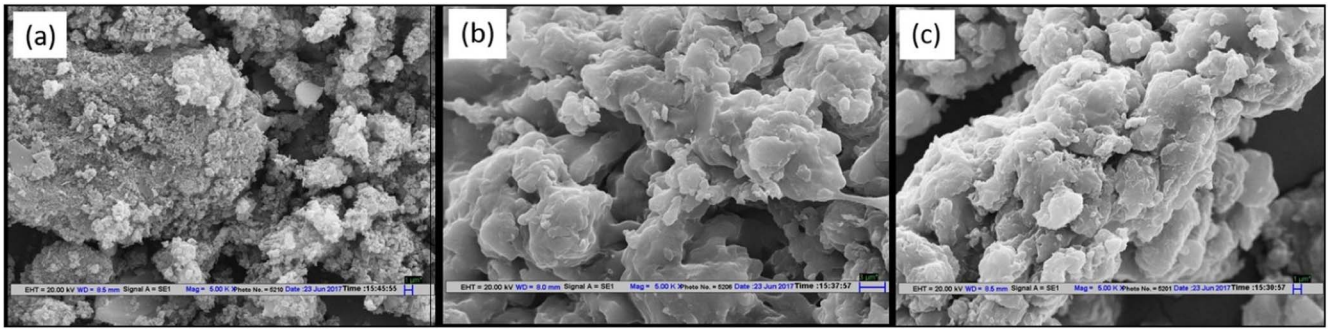


Figure 2. SEM images of (a) Katira gum hydrogel, (b) PANI, (c) Hydrogel-PANI composite, (Scale bar 1 μm).

Detection of heavy metal ions present in the sample.— Voltammetry method was used for the detection of heavy metals like cobalt, cadmium and Mercury in the solution. The electrodes (1 mg ml^{-1}) were first incubated in urea solution. 5 mM concentration of urea was standardised for the further studies. The oxidation peak current was noted in the cyclic voltammetry in the presence of the enzyme substrate reaction.[35] Various concentration of heavy metals ($CoCl_2.6H_2O$, $HgCl_2$ and $CdCl_2.H_2O$) from 0.1 mM to 5 mM in aqueous solutions were used to study the inhibitory effect of these metal ions on the urease activity of the fabricated electrode.

Results and Discussion

Scanning electron microscopy studies.—The SEM had been used for analyzing the morphology of the katira gum hydrogel, PANI and the hydrogel-PANI composite. SEM images clearly show the surface topology of the katira gum hydrogel as smooth solid sheet like structures. In the composite image, the rough clusters over the smooth sheets indicate the embodiment of PANI nanoparticles into the hydrogel network. Figures 2a and 2b shows the hydrogel and the PANI respectively. The PANI nanoparticles show a rough structure. This can be seen in the Fig. 2c as deposited over the smooth sheets of the hydrogel.

Swelling studies.—The results of the swelling studies conducted on the hydrogel were plotted as a graph of percentage swelling versus Time. The swelling studies were conducted in distilled water and PBS (pH 7.4). The phosphate buffer saline showed greater swelling percentage than water. The optimized sample of hydrogel was further used in the biosensing membrane.

PBS solvent showed a swelling percent of 534% and the water solvent showed a swelling of 218%. hydrogels exhibited a pH responsive swelling behaviour due to the association/dissociation of the hydrogen bonding between the hydroxyl groups and the carboxyl groups within the hydrogels. Figure 3 shows the swelling percentage of the hydrogel in PBS is greater than the swelling percentage in water. At a higher pH-COOH groups dissociates to form carboxylate ions that results in reduction of Hydrogen in the macromolecular network chain thus faster immobility achieved, results in increased swelling rate and swelling ratio.[36] The difference in the swelling ratio is due to the changes in the pH of the swelling medium. The swelling ratio changes with increase in the pH.[37,38] The swelling ratio is calculated by the formula Swelling ratio $= \{(W_s\text{-}W_d)/W_d\} \times 100$.

FTIR studies.—FTIR spectra of Katira gum and Katira-cl-poly (AAc) hydrogels, PANI and katira-cl-poly(AAc)/PANI are presented in Fig. 4. In case of katira gum, a broad band at 3421.0 cm^{-1} is due to O–H stretching vibrations. The peak 1072.6 cm^{-1} is of C–O and C–C stretching vibrations of pyranose rings of polysaccharide. The shoulder at 1252.5 cm^{-1} is due to the asymmetric C–O–C vibrations in glycosidic groups of polysaccharides present in the gum. The absorption bands in the region 1620.9 cm^{-1} (C=O) are due to carboxylic acid of D-galactouronic

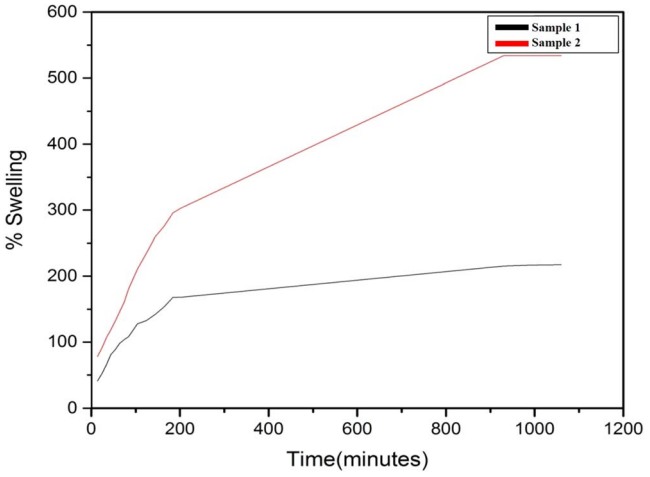


Figure 3. Slopes showing the swelling pattern of the hydrogel in (a) sample 1- PBS (pH 7.4) and (b) sample 2-water.

acid present in katira. The peak at 1423.0 cm^{-1} is assigned, to C–OH bend.[39,40]

FTIR spectrum of Katira-cl-poly(AAc) polymer shows absorption bands at 3446 cm^{-1} and 2927.0 cm^{-1} due to OH and C–H stretching vibrations respectively. The peak at 2362.5 cm^{-1} shows the overtones and combinations of O–H bending and C–O stretching vibrations. A peak at 1396.4 cm^{-1} has been assigned to CH_2 bending. The peak at 1166.1 cm^{-1} is due to C–O stretching of –COOH bond. These are the bands that are exclusively present in the hydrogel beside the bonds in the raw katira gum sample. The presence of more predominant band at 1716.7 cm^{-1} as and 1635.8 cm^{-1} which is due to COO^- asymmetric stretching, indicate the incorporation of poly(AAc) on to the polysaccharide backbone.[41]

The FTIR spectra of PANI, shows a peak at the 3437.1 cm^{-1} is due to the N–H bond stretching. The peak obtained at 2360.1 cm^{-1} is due to the NH unsaturated amine bond. The bend at 1568.9 cm^{-1} attributes to the C=N bond in the quinoidal units of the polymeric benzene ring. The 1480.5 cm^{-1} bend is assigned to the benzoid stretching vibrations. The band at the 1301.2 cm^{-1} region has appeared due to the C–N stretching vibrations in the alternate units of the quinoid-benzoid-quinoid rings. The bend at 1126.6 cm^{-1} is due to the C–H plane bending.[42]

The FTIR spectrum of the composite AAc-cl-katira/PANI shows peaks of both PANI and hydrogel confirming the incorporation of the emeraldine salt in the networks of the hydrogel. The 2359.4 cm^{-1} peak is due to the N–H unsaturated amine. The peaks at 1630.9 cm^{-1} is a broad band which is due to the clubbing of the $–COO^-$ bond stretching, the –C=N stretching of the quinoid rings and the C=C stretches of the quinoid ring. A band is observed at 1499.0 cm^{-1} which is attributed to the C–N stretching of the benzoid rings. Peaks at 2926.2 cm^{-1} and 1720.9 cm^{-1} is due to the C–H

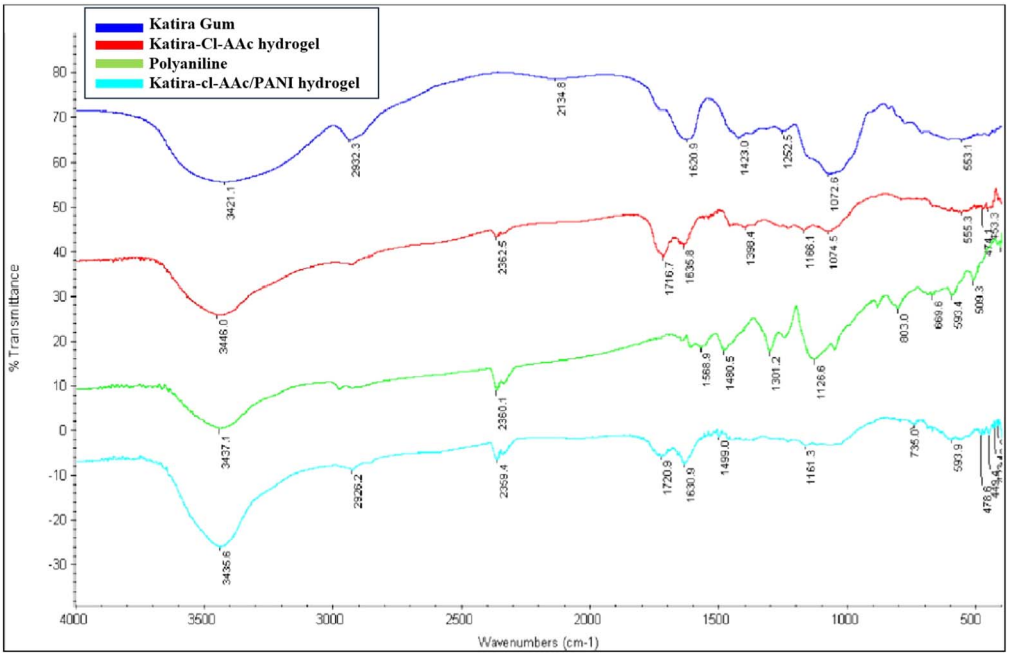


Figure 4. The FTIR spectra of katira gum, hydrogel, PANI, and PANI-Hydrogel.

alkane and –C=O bond stretching respectively. A peak at 1161.3 cm^{-1} is due to the C–O–C stretch.[43–45]

UV–vis Spectrometry results.—The Fig. 5 shows the UV-visible spectra of PANI and Katira-cl-AAc/PANI hydrogel. PANI showed two absorption peaks at 360 and 680 nm. The band at 360 nm is assigned to the benzonoid π–π^* transition. And the band at 680 nm is in the quinonoid rings due to the exciton like transition. The peak between 300–400 nm has been observed due to localized polarons. This is an essential characteristic of protonated PANI.[46] The absorption band of the katira-polyaniline hydrogel shifts due to the interaction between the quinoids and the benzoid and the katira particles. The blue shift of the absorption from 360 to 300 nm and the blue shift of the 680 to 650 nm in the absorbance may be due to the interaction of the polyaniline particles with the katira gum hydrogel.[47]

Thermo gravimetric analysis.—TGA is basically used for the finding the range of stability and for finding the decomposition temperature. In the Fig. 6, this shows the decomposition of the Katira-cl-AAc hydrogel, PANI and Katira-cl-AAc/PANI composite in an inert atmosphere of nitrogen.[48–50] The first drop in the weight have been observed from 50 °C to 175 °C, which is attributed to the loss of moisture content present in the sample. An abrupt bend in the slope has been observed from 175 °C to 380 °C, which signifies the constant decomposition of the compound. The decomposition, first starts with the volatile compounds present in the sample. The decomposition continues till 435 °C. The range of stability is attained in the temperature range of 450 °C–700 °C, before complete decomposition. At last nitrogen, is purged in the sample to ensure its complete decomposition.

Electrode studies.—The differential pulse voltammetry of the electrodes PANI/ITO, Katira-cl-AAc/ITO, katira-cl-AAc/PANI/ITO, and Urease/Katira-cl-AAc/PANI/ITO had been conducted using the three-electrode system in the electrochemical analyser. The electrodes have been immersed in PBS pH 7, 0.1 M, having 5 mM of $[Fe(CN)_6]^{3-/4-}$ with a scan rate of 50 mV s^{-1}. The potential has been changed from −0.3 to 0.5 V. In Fig. 7, It was observed that the magnitude of the peak current for PANI/ITO

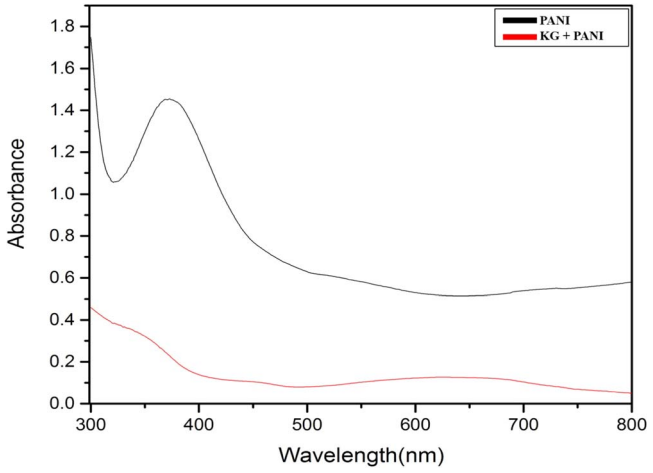


Figure 5. The UV-absorption spectrum of PANI and Katira-cl-AAc/PANI.

electrode was highest, 0.34 mA.[51] The magnitude of the peak current decreased in the katira-cl-AAc/PANI/ITO electrode to 0.22 mA, due to the decrease in the bulk concentration of polyaniline. This peak current magnitude is higher than the katira-cl-AAc/ITO electrode, as the composite is more conducting.[52] The immobilization of urease onto the electrode further increases the peak oxidation current value to 0.24 mA. The highest current is seen in the case of PANI/ITO electrode, because of pure conducting polymer nanoparticles. In all the other electrodes there is a little hindrance due to the presence of the bulkier groups.

Further, for the optimization of the concentration of urease, the cyclic voltammetry has been conducted on the Urease/Katira-cl-AAc/PANI/ITO electrode . The potential has been changed from −0.4 V to 0.5 V, with a scan rate of 50 mV s^{-1}. Different concentration of urease, varying from 0.5 mg ml^{-1} to 2 mg ml^{-1} has been prepared in PBS buffer, (pH 7, 0.1 M) shown in Fig. 8. In Fig. 9, the peak oxidation current of each electrode has been measured and compared with 1 mg ml^{-1} being the highest. Higher the concentration of the enzyme, slower becomes the mobility of the

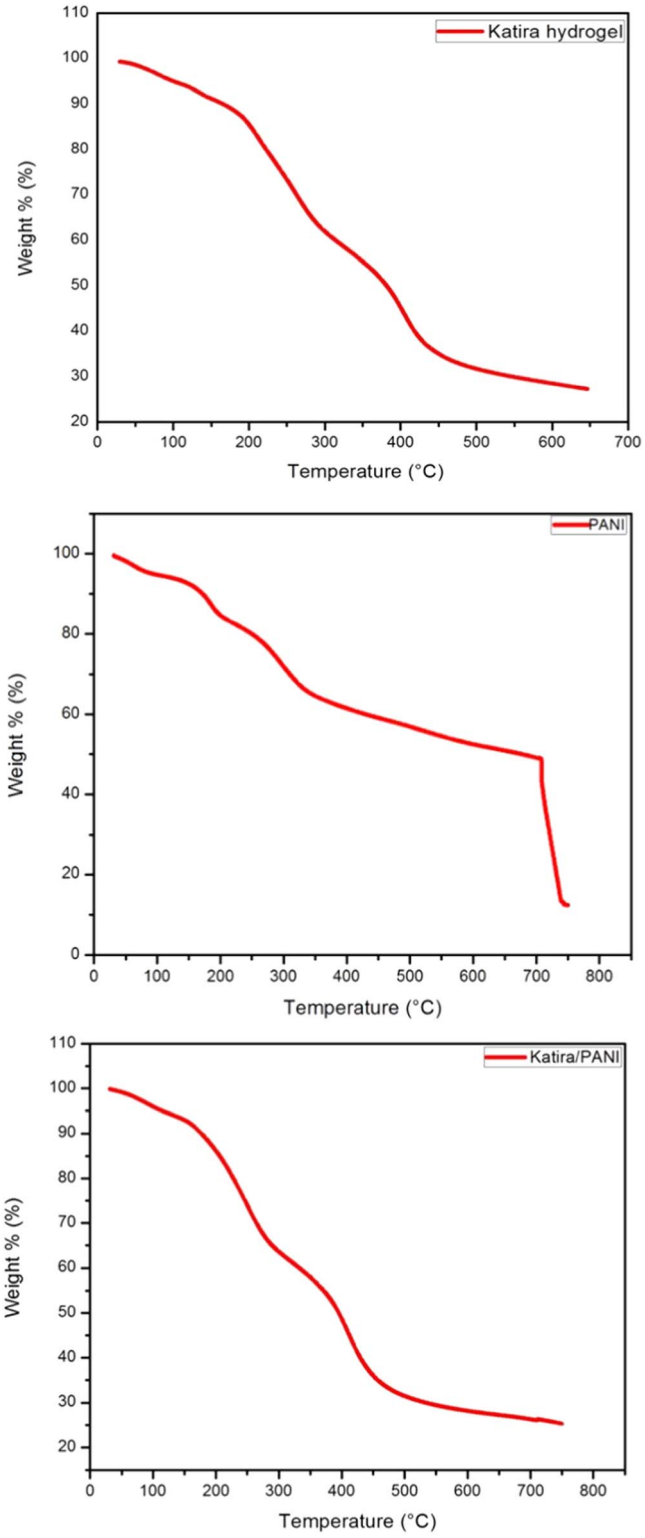


Figure 6. TGA of Katira-cl-AAc hydrogel, PANI and Katira-cl-AAc/PANI composite.

ions due to the increase in the bulk concentration. Thus 1 mg ml^{-1} has been chosen as the ideal concentration of the enzyme for the substrate detection.[53–55]

Increasing concentration of urea as a substrate, from 1 mM to 10 mM has been used for the biosensing. It has been seen that, at first, with increase in concentration of the substrate, the peak current increases. This is due to the presence of excess enzyme to substrate

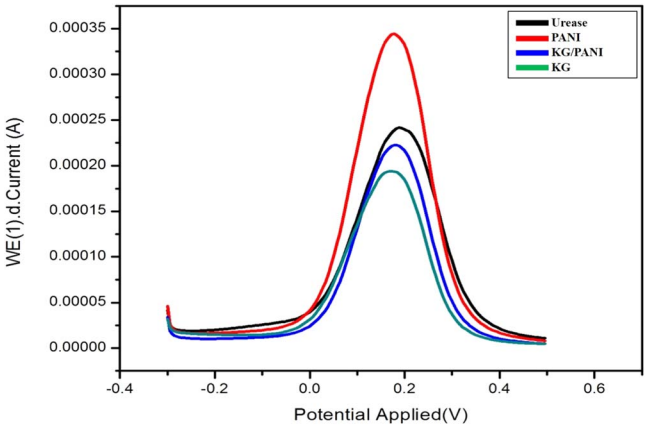


Figure 7. DPV of PANI/ITO, Katira-cl-AAc/ITO, Katira-cl-AAc/PANI/ITO and Urease/Katira-cl-AAc/PANI/ITO.

ratio. As the substrate, i.e., urea is increased, the enzyme catalyses the conversion of urea to ammonia and carbon-dioxide. A highest oxidation current has been observed in 5 mM concentration. Beyond this concentration, as the substrate is increased, there is a decrease in the peak current. This can be attributed to the cause that, all the active site on the enzyme has been occupied by the substrate molecules. Thus, decreasing the charge transfer rate, and hence reducing the oxidation current.

Incubation time studies.—The Fig. 10 shows the response of the electrode Urease/Katira-cl-AAc/PANI/ITO in 5 mM urea concentration with respect to different incubation times. It has been observed that there is an increase in the current with the increase in the incubation time from 5 min to 45 min. However, after 35 mins of incubation, the peak current of the electrode becomes constant. The electrode becomes stable at a particular oxidation current. Hence, the incubation time of 40 min has been selected for the further enzyme substrate response studies. After 40 min of incubation the enzyme substrate reaction reaches a steady state.[56]

Heavy metal salt inhibition of immobilised urease.—The enzyme electrode were first incubated with different concentrations of heavy metal salts for 5 to 10 min with the subatrate urea. Cyclic voltammetry studies were done to analyze the divergence in the peak current from the enzyme electrode in 0.1 M PBS (pH 7) containing 5 mM of $[Fe(CN)_6]^{3-/4-}$. The scan rate used was 10 mV s^{-1}. Three different heavy metal salts were used for the comparative study of the percent inhibition of the enzyme urease activity with identical measuring conditions. This is simple and portable detection method, can detect heavy metal ions selectively with less complex instrumentation.[57,58] It was observed that, there is a retarding action of the immobilized urease due to reduction in the peak oxidtion current with increase in the concentration of the heavy metal salts. For each heavy metal the concentration was varied from 0.1 mM to 5 mM. Three salts of cobalt, Mercury and cadmium were used to analyze the change in the enzyme activity. Percent inhibition was calculated using the formula,

$$\% \text{ Inhibition} = [(A_0 - A)/A_0 \times 100]$$

Where A_0 is the initial peak current of the Urease/Katira-cl-AAc/PANI/ITO electrode and A is the inhibition peak current at differnt concentrations of metal salts.[59,60]

A graph of Concentration of the heavy metal ions versus % Inhibition was plotted for each metal ion. The caused was attributed to the non-competitive inhibition, as cited in previous papers, in the case of Cadmium ion, cobalt and mercuric ion. Non-competitive inhibition is due to the binding of the metal ions to the sites other than the allesteric sites of the enzyme. The inhibitor reduces the

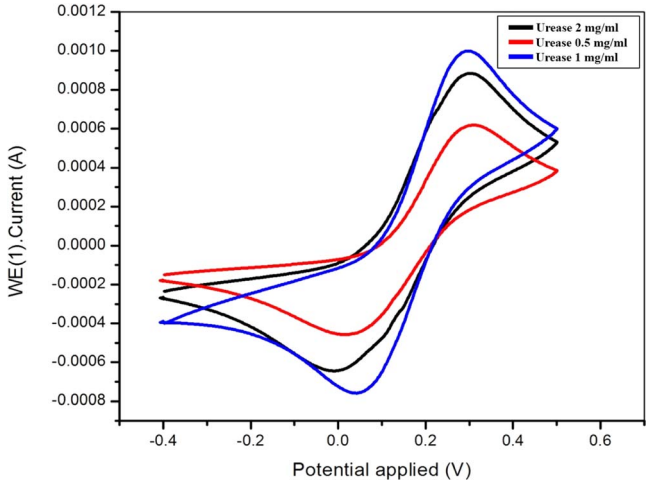


Figure 8. The cyclic voltammetry studies of different concentration of urease.

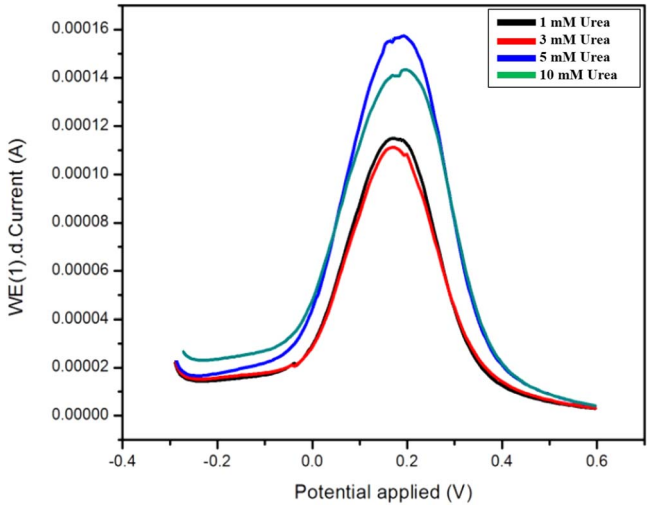


Figure 9. DPV of different concentrations of the substrate, urea for 1 mg ml^{-1} urease electrode.

catalytic activity of the enzymes by binding to the enzyme, despite of the fact that the substrate has already bound to the enzyme or not.[61] The slope of the graphs showed the following order of inhibition

$$Hg^{2+} > Cd^{2+} > Co^{2+}$$

Mercuric ion showed a sharp increased inhibition in the concentration range of 1.8 mM to 5 mM. As compared to the mercuric ion, cadmium ion showed greater inhibition when the concentration range of 0.1 mM to 2 mM. The cobalt ion showed a constant increase in the inhibition of the enzyme activity from 0.1 mM to 5 mM.

The graphs in Figs. 11a–11c shows the % Inhibition of the urease activity with respect to the increasing concentration of the metal ions.[62] In case of the mercuric ion, an increase in the peak oxidation current has been observed durin the incresing concentration of upto 1.8 mM, beyond this concentration, there is a sharp decrease in the value of the oxidation current. This is due to some discripancy in the addition of the Mercury salt which has increase the rate of charge transfer in the fabricated electrode system. There may be other cause like the Hg ion operates via other means including direct binding to the active site. A bar graph in the Fig. 11d shows the comparative

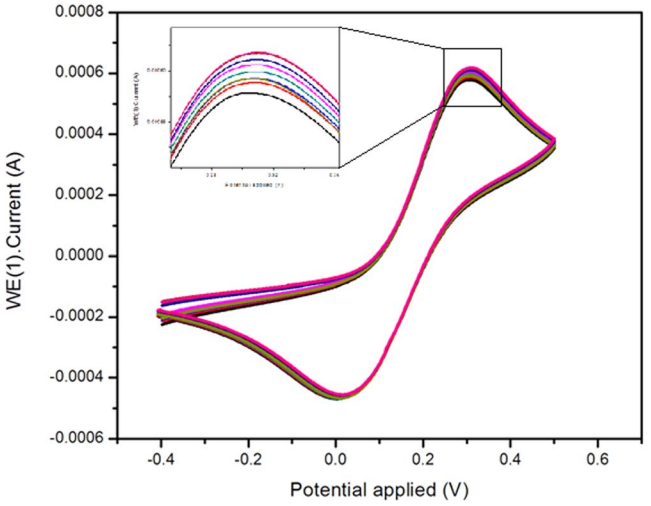


Figure 10. Incubation time study of the enzyme electrode using cyclic voltammetry.

study of the percentage inhibition of the catalytic activity of the enzyme urease with respect to the differnt concentrations of the three heavy metal ions, Hg^{2+}, Co^{2+} and Cd^{2+}. The bars have been divided in two range of concentration, first being 0.1 mM to 1.2 mM in which cobalt ion has shown minimum inhibition rate as compared to the other two metal ions. And the second is the range between 1.8 mM to 5.0 mM where the inhibition rates of Mercury and cadmium is almost same.[63] The fabricated biosensor has been demonstrated to be well-suited for rapid and fast detection of heavy metal ions. The findings reveal a remarkable sensitivity and comparable potential to other urease-based techniques for detecting heavy metal ions (Table I).

Conclusions

In this study, we have proposed a biocompatible, non-toxic hydrogel-based biosensor for the detection of the heavy metal ions present in any water sample. PANI doped with HCl, a conducting polymer has been incorporated in the hydrogel which has helped in enhancing the conducting properties of the electrode. The ratio of katira gum: aniline monomer has been polymerized in the proportion of 2:1. Firstly the katira gum hydrogel has been synthesized, and then this hydrogel has been swelled in the solution of aniline to get an interpenetrating network of conducting hydrogel composite, Katira-cl-AAc/PANI. This composite has been dried at 60 °C to remove any moisture content present. The dried product has been dispersed in an organic solvent, DMSO to form a solution, which was then drop cast over the hydrolysed ITO electrode. The area for the drop cast was kept one-third of the total surface of the electrode. Immobilization of urease enzyme was also carried out with the help of the drop casting method, wherein 20 μl of the enzyme was taken. Prior to the enzyme immobilization, the electrode was functionalized with 2.5% glutaraldehyde solution for the activation of the groups in the hydrogel composite which would ease the attachment of urease onto the electrode surface. The fabricated Urease/Katira-cl-AAc/PANI/ITO electrode is simple and has been used for the quantification of urea as well as the heavy metal ions. The composite Katira-cl-AAc/PANI has shown good conducting properties, which can be used as a bioreceptor membrane for any other electrode fabrication. The hydrogel has increased the efficiency of the electrode as its swelling property has provided a large surface area for the attachment of the enzyme. This in turn has provided a greater surface area for the interaction of the substrate urea with the enzyme. The application of the immobilized enzyme has several benefits than the free enzyme, first being the reusability of the enzyme electrode, the second being the low consumption of the enzyme.

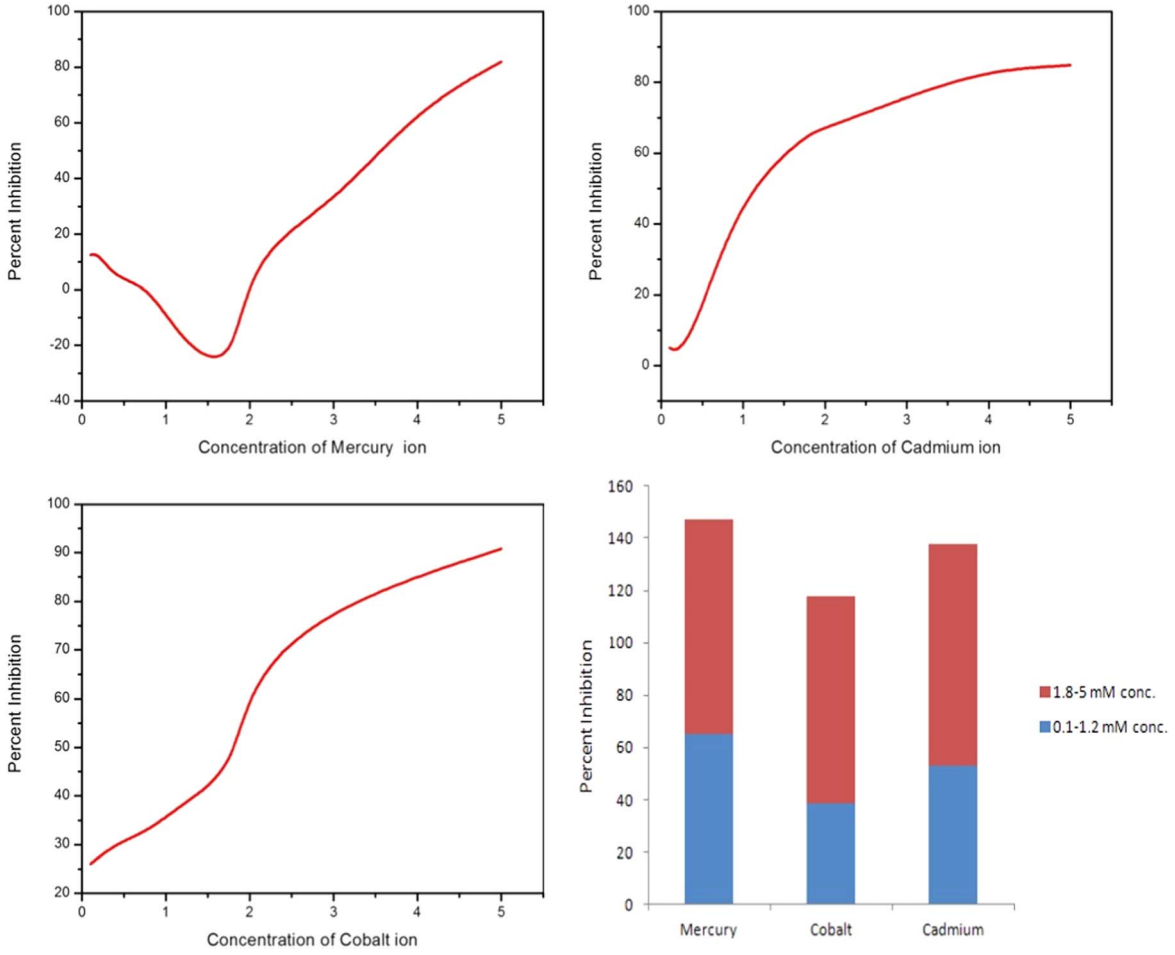


Figure 11. % Inhibition of urease with respect to (a) Hg^{2+}, (b) Cd^{2+}, (c) Co^{2+} and (d) Bar graph showing the comparative inhibition rates of heavy metals.

Table I. Characteristics of various detection techniques for heavy metals using urease.

Detection technique	Sample	Platform	Detection limit	References
Conductometric	$Hg(NO_3)_2$	Urease/Au/Al_2O_3	0–40 mM	64
Potentiometric	Heavy metals (Cu, Cd, Hg)	Urease/Carbon/Rhodium/Carboxymethyl cellulose	0–20 ppb	65
Conductometric	Heavy metals	Sol-gel screen printed electrode	0.1–10 mM	66
Conductometric	Heavy metals	Urease/TMOS sol-gel/IDA	0.1 mM	67
Electrochemical	Hg	Urease/Gold electrode	0–50 μM	68
CV	Hg, Cd, Co	Urease/Katira-cl-AAc/PANI/ITO	0.1–5 mM	Current work

The heavy metal ions that have been used for the inhibition studies are found in the aquatic system in high concentrations and has affected a wide population causing many types of diseases and increasing the level of toxicity in humans. The non-competitive inhibition of the enzyme urease by the metal salts has been used for the quantification of the metal ions.

ORCID

Ravi Butola https://orcid.org/0000-0002-6581-4131
Prakash Chandra https://orcid.org/0000-0002-8736-7993

References

1. S. Mitra et al., "Impact of heavy metals on the environment and human health: Novel therapeutic insights to counter the toxicity." *Journal of King Saud University-Science*, **34**, 101865 (2022).
2. M. Jaishankar et al., "Toxicity, mechanism and health effects of some heavy metals." *Interdisciplinary Toxicology*, **7**, 60 (2014).
3. K. Aftab, M. D. Ali, P. Aijaz, N. Beena, H. J. Gulzar, K. Sheikh, and S. Tahir Abbas, "Determination of different trace and essential element in lemon grass samples by X-ray fluorescence spectroscopy technique." *Int. Food Res. J.*, **18**, 265 (2011).
4. R. D. Beaty and J. D. Kerber, *Concepts Instrumentation and Techniques in Atomic Absorption Spectrophotometry.* (PerkinElmer Shelton CT, USA) 8 (1993).
5. V. Rai, P. Kakkar, S. Khatoon, A. Rawat, and S. Mehrotra, "Heavy metal accumulation in some herbal drugs." *Pharm. Biol.*, **39**, 384 (2001).
6. D. Stankovic, G. Roglic, J. Mutic, I. Andjelkovic, M. Markovic, and D. Manojlovic, "Determination of copper in water by anodic stripping voltammetry using Cu-DPABA– NA/GCE modified electrode." *Int. J. Electrochem. Sci.*, **6**, 5617 (2011).
7. R. J. Brown and M. J. Milton, "Analytical techniques for trace element analysis: an overview." *Trends Analyt Chem*, **24**, 266 (2005).
8. A. K. Poddar, S. S. Patel, and H. D. Patel, "Synthesis, characterization and applications of conductive polymers: A brief review." *Polym. Adv. Technol.*, **32**, 4616 (2021).
9. S. Iqbal and S. Ahmad, "Recent development in hybrid conducting polymers: synthesis, applications and future prospects." *J. Ind. Eng. Chem.*, **60**, 53 (2018).
10. M. Gerard, A. Chaubey, and B. D. Malhotra, "Application of conducting polymers to biosensors." *Biosens. Bioelectron.*, **17**, 345 (2002).
11. C. Tran-Minh, "Immobilized enzyme probes for determining inhibitors." *Ion-Selective Electrode Rev.*, **7**, 41 (1985).

12. J. Gayet, A. Haouz, A. Geloso-Meyer, and C. Burstein, "Detection of heavy metal salts with biosensors built with an oxygen electrode coupled to various immobilised oxidases and dehydrogenases,." *Biosensors Bioelectron.*, **8**, 177 (1993).
13. J. T. MacGregor and T. W. Clerkson, ed. M. Friedman *Protein Metal Interactions* (Plenum, New York) p. 463 (1974).
14. J. L. Webb, *Enzyme and Metabolic Inhibitors.* (Academic, New York) volume II, p. 729 (1966).
15. B. Bharaniraja, K. J. Kumar, C. M. Prasad, and A. K. Sen, "Modified katira gum for colon targeted drug delivery." *J. Appl. Polym. Sci.*, **119**, 2644 (2011).
16. Wealth of India, (CSIR, New Delhi) 2, p 261 (1962).
17. K. R. Kirtikar and B. D. Basu, *In Indian Medicinal Plant Cochlospermaceae*, ed. E. Blatter, J. F. Caius, K. S. Mhaskar, Singh Bishen, and Mahendra Pal Singh (Dehradun, India) 1, p. 214 (1998).
18. G. O. Aspinall and J. Baillie, "Gum tragacanth. Part I. Fractionation of the gum and the structure of tragacanthic acid." *J. Chem. Soc.*, **1**, 1702 (1963).
19. T. W. Schwarz, G. Levy, and H. H. Kawagoe, "Tragacanth solutions III. The effect of pH on the stability." *J. Am. Pharm. Assoc.*, **47**, 695e696 (1958).
20. T. K. Chatterjee et al., "Eudragit coated Gum Katira matrix microsphere formulation for sustained release of Etodolac." *Indian Journal of Pharmaceutical Education and Research*, **50** (2016).
21. R. Trigo, M. Blanco, P. Huerta, R. Olmo, and J. Teijon, "L-Ascorbic acid release from poly(2-hydroxyethyl methacrylate) hydrogels." *Polym. Bull.*, **31**, 577 (1993).
22. E. Gomez-Ordonez and P. Ruperez, "FTIR-ATR spectroscopy as a tool for polysaccharide identification in edible brown and red seaweeds." *Food Hydrocolloids*, **25**, 1514 (2011).
23. S. K. Bhardwaj et al., "Bi-enzyme functionalized electro-chemically reduced transparent graphene oxide platform for triglyceride detection." *Biomater. Sci.*, **7**, 1598 (2019).
24. J. Singh et al., "Preparation of sulfonated poly (ether–ether–ketone) functionalized ternary graphene/AuNPs/chitosan nanocomposite for efficient glucose biosensor." *Process Biochem.*, **48**, 1724 (2013).
25. R. Verma, S. K. Samdarshi, and J. Singh, "Hexagonal ceria located at the interface of anatase/rutile TiO2 superstructure optimized for high activity under combined UV and visible-light irradiation." *The Journal of Physical Chemistry C*, **119**, 23899 (2015).
26. M. Srivastava et al., "Preparation and characterization of self-assembled layer by layer NiCo2O4–reduced graphene oxide nanocomposite with improved electrocatalytic properties." *J. Alloys Compd.*, **590**, 266 (2014).
27. P. Singh et al., "Preparation, antibacterial activity, and electrocatalytic detection of hydrazine based on biogenic CuFeO 2/PANI nanocomposites synthesized using Aloe barbadensis miller." *New J. Chem.*, **46**, 8805 (2022).
28. M. Singh et al., "Recent progress on modified gum katira polysaccharides and their various potential applications." *Polymers*, **14**, 3648 (2022).
29. Z. Li and L. Gong, "Research progress on applications of polyaniline (PANI) for electrochemical energy storage and conversion." *Materials*, **13**, 548 (2020).
30. M. Beygisangchin et al., "Preparations, properties, and applications of polyaniline and polyaniline thin films—A review." *Polymers*, **13**, 2003 (2021).
31. Y. Wu et al., "Fabrication of conductive polyaniline hydrogel using porogen leaching and projection microstereolithography." *J. Mater. Chem. B*, **3**, 5352 (2015).
32. J. Y. Kim, G. Y. Sung, and M. Park, "Efficient portable urea biosensor based on urease immobilized membrane for monitoring of physiological fluids." *Biomedicines*, **8**, 596 (2020).
33. S. M. H. Bukhari, S. Khan, M. Rehanullah, and N. M. Ranjha, "Synthesis and characterization of chemically cross-linked acrylic acid/gelatin hydrogels: Effect of pH and composition on swelling and drug release." *Int. J. Polym. Sci.*, **2015**, 187961 (2015).
34. P. Chandra, J. Kim, and S. W. Rhee, "Quantitative analysis of growth of cells on physicochemically modified surfaces." *Bull. Korean Chem. Soc.*, **34**, 524 (2013).
35. N. Sanaeifar, M. Rabiee, M. Abdolrahim, M. Tahriri, D. Vashaee, and L. Tayebi, "A novel electrochemical biosensor based on Fe3O4 nanoparticles-polyvinyl alcohol composite for sensitive detection of glucose." *Anal. Biochem.*, **519**, 19 (2017).
36. Y. H. F. Al-qudah, G. A. Mahmoud, and M. A. Abdel Khalek, "Radiation crosslinked poly (vinyl alcohol)/acrylic acid copolymer for removal of heavy metal ions from aqueous solutions." *Journal of Radiation Research and Applied Sciences*, **7**, 135 (2014).
37. Y. M. Lee and S. Y. Kim, "Stimuli-responsive interpenetrating polymer network hydrogels composed of poly (vinyl alcohol) and poly (acrylic acid)." *Smart Fibres, Fabrics and Clothing* (Woodhead Publishing (Elsevier), Cambridge, UK) 93 (2001).
38. T. M. Lim, M. Ulaganathan, and Q. Yan, "Advances in membrane and stack design of redox flow batteries (RFBs) for medium-and large-scale energy storage." *Advances in Batteries for Medium and Large-Scale Energy Storage.* (Woodhead Publishing (Elsevier), Cambridge, UK)477 (2015).
39. M. Singh et al., "Recent progress on modified gum katira polysaccharides and their various potential applications." *Polymers*, **14**, 3648 (2022).
40. V. Kumar et al., "Gum katira-silver nanoparticle-based bionanocomposite for the removal of methyl red dye." *Frontiers in Chemistry*, **10**, 959104 (2023).
41. J. Nath et al., "Acrylic acid grafted gelatin/LDH based biocompatible hydrogel with pH-controllable release of vitamin B12." *Appl. Clay Sci.*, **190**, 105569 (2020).
42. B. Massoumi et al., "Poly (N-octadecylaniline) synthesis and its electrochemical parametric characterizations." *Iranian Polymer Journal*, **20**, 779 (2011).
43. B. S. Singu, P. Srinivasan, and S. Pabba, "Benzoyl peroxide oxidation route to nano form polyaniline salt containing dual dopants for pseudocapacitor." *J. Electrochem. Soc.*, **159**, A6 (2011).
44. R. Prabhakar and D. Kumar, "Influence of dopant ions on the properties of conducting polyacrylamide/polyaniline hydrogels." *Polym.-Plast. Technol. Eng.*, **55**, 46 (2016).
45. H. Xie et al., "A conductive polyacrylamide/double bond chitosan/polyaniline hydrogel for flexible sensing." *J. Mater. Sci., Mater. Electron.*, **31**, 10381 (2020).
46. B. Yan et al., "Fabrication of polyaniline hydrogel: synthesis, characterization and adsorption of methylene blue." *Appl. Surf. Sci.*, **356**, 39 (2015).
47. W. Naqash and K. Majid, "Synthesis, characterization and study of effect of irradiation on electronic properties of polyaniline composite with metal complex of Co (III)." *Mater. Res.*, **18**, 1121 (2015).
48. J. Chen et al., "A conductive bio-hydrogel with high conductivity and mechanical strength via physical filling of electrospinning polyaniline fibers." *Colloids Surf., A*, **637**, 128190 (2022).
49. M. Singh et al., "Recent progress on modified gum katira polysaccharides and their various potential applications." *Polymers*, **14**, 3648 (2022).
50. M. Moussa et al., "Free-standing composite hydrogel films for superior volumetric capacitance." *J. Mater. Chem. A*, **3**, 15668 (2015).
51. D. H. Ninh et al., "Characterization of electrochromic properties of polyaniline thin films electropolymerized in H2SO4 solution." *Open Journal of Organic Polymer Materials*, **6**, 30 (2016).
52. W. Zhang et al., "A hydrogel electrochemical electrode for simultaneous measurement of cadmium ions and lead ions." *J. Electroanal. Chem.*, **901**, 115756 (2021).
53. S. Baluta et al., "Differential pulse voltammetry and chronoamperometry as analytical tools for epinephrine detection using a tyrosinase-based electrochemical biosensor." *RSC Adv.*, **12**, 25342 (2022).
54. V. Sanko et al., "A novel comparative study for electrochemical urea biosensor design: Effect of different ferrite nanoparticles (MFe2O4, M: Cu, Co, Ni, Zn) in urease immobilized composite system." *Bioelectrochemistry*, **149**, 108324 (2023).
55. G. Sumana et al., "A novel urea biosensor based on zirconia." *Thin Solid Films*, **519**, 1187 (2010).
56. S. Upadhyay et al., "Application of bimetallic nanoparticles modified screen printed electrode for the detection of organophosphate compounds using an enzyme inhibition approach." *Anal. Methods*, **3**, 2246 (2011).
57. R. Fopase et al., "Inhibition assays of free and immobilized urease for detecting hexavalent chromium in water samples." *3 Biotech*, **9**, 1 (2019).
58. G. A. Zhylyak et al., "Application of urease conductometric biosensor for heavy-metal ion determination." *Sensors Actuators B*, **24**, 145 (1995).
59. L. Mazzei et al., "The structure of urease inactivated by Ag (I): a new paradigm for enzyme inhibition by heavy metals.'." *Dalton Trans.*, **47**, 8240 (2018).
60. H. Hou et al., "Recent advances and perspective on design and synthesis of electrode materials for electrochemical sensing of heavy metals." *Energy & Environmental Materials*, **1**, 113 (2018).
61. K. P. Delaune and K. Alsayouri, Physiology, noncompetitive inhibitor (2019).
62. T. O. Hara and B. Singh, "Electrochemical biosensors for detection of pesticides and heavy metal toxicants in water: recent trends and progress." *ACS ES&T Water*, **1**, 462 (2021).
63. A. M. Ashrafi et al., "Application of the enzymatic electrochemical biosensors for monitoring non-competitive inhibition of enzyme activity by heavy metals." *Sensors*, **19**, 2939 (2019).
64. V. N. Arkhypova et al., "Multibiosensor based on enzyme inhibition analysis for determination of different toxic substances." *Talanta*, **55**, 919 (2001).
65. P. Pal et al., "The detection of Mercury, cadium, and arsenic by the deactivation of urease on rhodinized carbon." *Environ. Eng. Sci.*, **26**, 25 (2009).
66. R. Ilangovan et al., "Enzyme based biosensor for heavy metal ions determination." *Biotechnology & Biotechnological Equipment*, **20**, 184 (2006).
67. S. M. Lee and Y. L. Won, "Determination of heavy metal ions using conductometric biosensor based on sol-gel-immobilized urease." *Bull. Korean Chem. Soc.*, **23**, 1169 (2002).
68. G. A. hylyak et al., "Application of urease conductometric biosensor for heavy-metal ion determination." *Sensors Actuators B*, **24**, 145 (1995).

Physica Scripta

INCLUSIVE PUBLISHING
TRUSTED SCIENCE

PAPER

Ditch incorporated organic thin film transistor based organic all-p inverter: a novel approach

To cite this article: Sakshi Gupta *et al* 2023 *Phys. Scr.* **98** 115038

View the article online for updates and enhancements.

You may also like

- Sub-Doppler DAVLL spectra of the D1 line of rubidium: a theoretical and experimental study
Gyeong-Won Choi and Heung-Ryoul Noh
- Search for the singlet vector-like lepton through the pair production in the W channel at the ILC
Haitao Li, Jinjin Chao and Guoqing Zhang
- Regularization parameter selection for penalized-likelihood list-mode image reconstruction in PET
Mengxi Zhang, Jian Zhou, Xiaofeng Niu et al.

This content was downloaded from IP address 14.139.251.98 on 03/11/2023 at 05:46

Physica Scripta

PAPER

Ditch incorporated organic thin film transistor based organic all-p inverter: a novel approach

Sakshi Gupta[1], **Poornima Mittal**[2,*] **and Pradeep Juneja**[1]

[1] Department of Electronics and Communication Engineering, Graphic Era (Deemed to be University) Dehradun, Uttarakhand, India
[2] Department of Electronics and Communication Engineering, Delhi Technological University, New Delhi, India
[*] Author to whom any correspondence should be addressed.

E-mail: sakshigupta.ice@geu.ac.in, poornimamittal@dtu.ac.in and pkjuneja@geu.ac.in

RECEIVED
10 May 2023

REVISED
1 October 2023

ACCEPTED FOR PUBLICATION
10 October 2023

PUBLISHED
26 October 2023

Keywords: Zero Vgs Load Logic topologies (ZVLL), Diode Load Logic (DLL), noise margin, organic all p inverter, organic thin film transistor, propagation delay

Abstract

This paper demonstrates the static and dynamic characteristics of all-p organic inverter employing a bottom gate bottom contact organic thin film transistor with a ditch incorporated into the OSC and additional p+ doping done to its S/D nearing area. Various configurations of OTFT have been experimented with and it is found that as compared to an OTFT without extra p+ doping, the devices having additional doping show better responses. Where, for SG OTFT, I_D is derived to be 18 μs, yet the one with extra doping illustrates a swooping 16.67% increased output current of 21 μs. To enhance the performance of the device further, a ditch of 30 nm, embedded 10 nm into the OSC is incorporated, further augmenting the performance of the device by 55.5% as compared to the conventional BGBC. Besides, the proposed inverter presents a considerably elevated performance in terms of robustness and low and high noise margins. This paper further compares the inverter using Diode Load Logic and Zero V_{gs} Load Logic topologies, wherein it was found that DLL shows an exceptional 211% less propagation delay τ_p of 27 μs, as compared to 84 μs delay experienced by ZVLL. But comparing the two topologies in terms of the static response, ZVLL are way better and preferred over the counterparts since ZVLL configuration displays 20.5% augmented Noise Margin, improved gain, and overall robustness. Owing to the performance parameters achieved, such organic inverters may be incorporated into integrated circuits rendering trustworthiness to digital operations in electronic circuits and numerous cascading applications. Since the organic inverters made using the proposed OTFT exhibit a decent gain hence have an apparent prospective of driving myriad-stage logic like ring oscillators and memory blocks. Moreover, these may be employed in the areas of biosensors and wearable electronics as well wherever.

1. Introduction

Of late, technology pertaining to the Organic Thin Film Transistor has gained momentum in manifold applications exhibiting elevated resilience and solution processability [1] at a curbed fabrication cost. This has facilitated a headway into using paper, glass, textile, and plastic as substrates offering pliability without staking the electronic function of the device [2]. Moreover, owing to the semiconductor attributes of the organic material employed in designing, OTFTs are being used in other diversified applications viz. conformable imagers [3], digital switches, memory circuits [4], organic memory [5], code generators [6] and e-paper [7] to name a few.

Over the last three decades, there have been massive alterations to the semiconductor manufacturing technology making it a challenge to maintain a positive and direct relation between the component density of a circuit to its cost effectiveness and an indirect relationship with dimension size of the device. In VLSI systems, digital integrated circuits may contain billions of logic gates. At the outset, constraint pertaining to the

© 2023 IOP Publishing Ltd

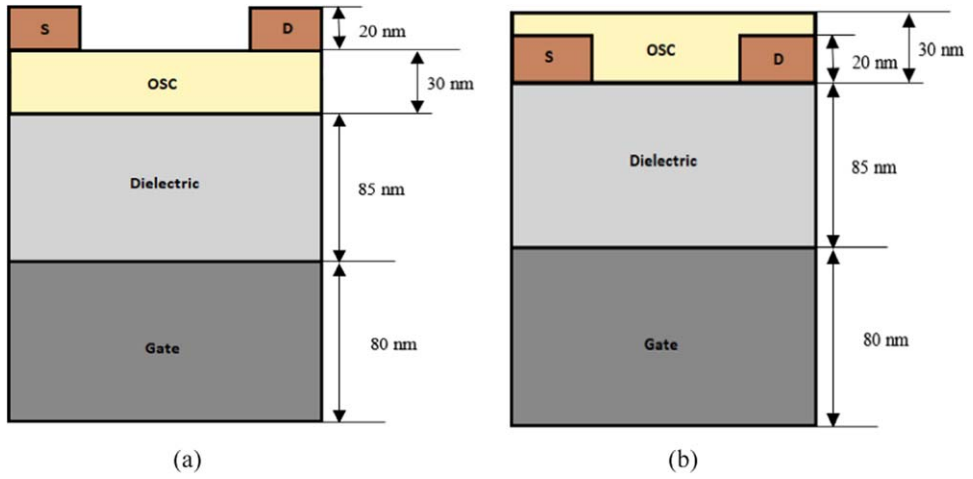


Figure 1. Representation of SG organic TFT (a) BGTC (b) BGBC.

manufacturing cost and bendability of the devices created using conventional semiconductor logic gates or inverters, was prominent. With the evolution of a-scale integration and usage of organic materials in designing the devices, this concern has been catered to, alongside achieving a legitimate performance. Although the employability of gates and inverters has seen a growth in leaps and bounds, the idea of designing logic gates based on OTFTs is yet to be scaled [8]. While forming organic inverter circuits, high noise margin, voltage swing and gain, besides maintaining a low propagation delay must be the pre-requisites [9].

In the pretext, it is worth a mention that as compared to the values of performance parameters derived in the 1980s, the parameters have been augmented to a considerable extent. Against the values of mobility as low as 10^{-6}—10^{-3} $cm^2\,Vs^{-1}$ back then, now the researchers can yield the values as high as 10^{-2}—10^{-1} $cm^2\,Vs^{-1}$. Likewise, the factor with which the current is being modulated has also increased to 10^6 from 10^3 [10].

This paper in the first section demonstrates the prologue to the organic TFT and organic inverters. The second discusses the various possible structural analysis of OTFT alongside exhibiting a single gate- bottom gate bottom contact (SG BGBC) device structure with and without the incorporation of ditch and p+ doping in the organic semiconductor layer (OSC) [11]. Thereafter, section 3 elaborates on employing the mentioned device for the construction of an all-p type OTFT organic inverter Moreover it also attracts the attention towards the inverter being evaluated under Diode-Load Logic (DLL) and Zero-V_{gs}-Load Logic (ZVLL) configurations for the calculation and derivation of the static and dynamic characteristics. It is also proposed to use additional p+ doping in the vicinity of the source and drain contact of the device for enhancing the performance by shrinking the work function gap [12], converging to proving the superiority of all-p type inverters in ZVLL, over all other. Thereafter, post discussing the fabrication steps, section 4 comprises result and discussion to summarize the entire work. Lastly in section 5, conclusion makes it clear as to how the proposed device is better than the reported devices.

2. Structural analysis of otft

The thin film transistors contain different tiers viz.: organic semiconductor layer, source [S] electrodes, drain [D] electrodes, gate dielectric layer and gate electrode. This emplacement of S/D renders varied structures to the OTFT. The varied structures are as under:

2.1. BGBC OTFT versus BGTC OTFT

When S/D are placed over the OSC, the structure that forms is known as bottom gate top contact [BGTC] whereas when they find their placement below the OSC, it forms a bottom gate bottom contact OTFT This section demonstrates a SG BGTC and BGBC. To construct this architecture of the organic TFT, the device is fabricated in layers; (a) & (b) parts of figures 1 and 2.

The first layer to be deposited is the heavily doped aluminum (p^+ Al) substrate acting as the gate electrode. The gate electrode is given a thickness of 80 nm BGBC structures are favored over the top contact counterparts with an obvious and evident reason that the active layer in top gate structures gets subjected to debauchment at excessive temperatures. $La_{0.85}Nb_{0.15}O_y$ is picked for constituting the dielectric layer of 85 nm thickness besides

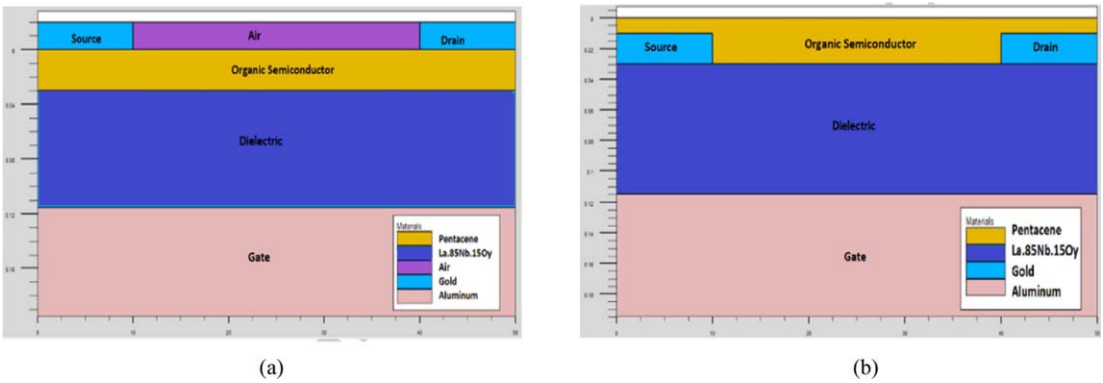


Figure 2. Simulated structures of (a) BGTC (b) BGBC.

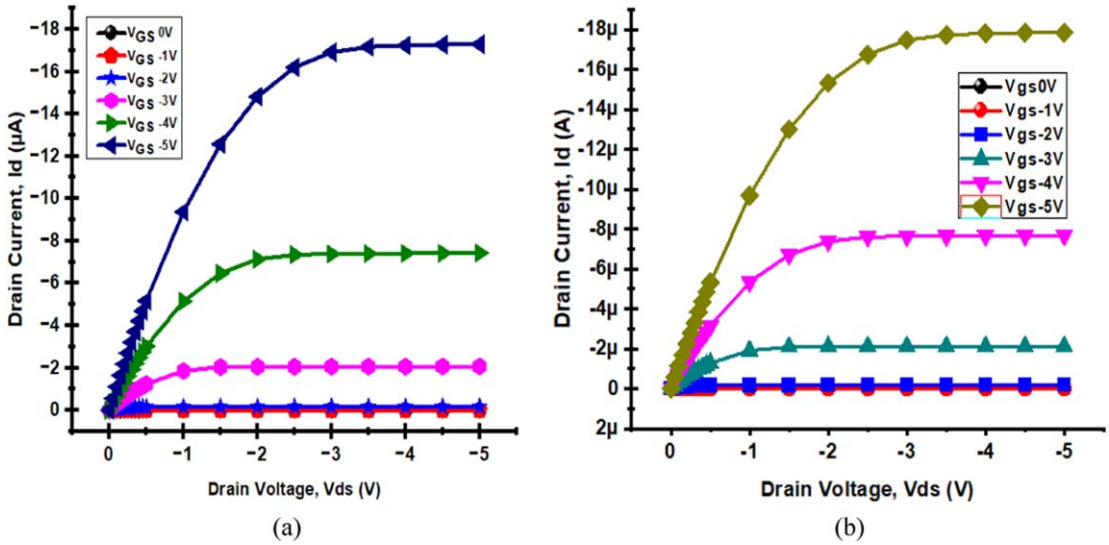


Figure 3. I_d-V_{ds} of (a) BGTC (b) BGBC structures.

Table 1. Thickness and material of different layers of OTFT.

Notation	Parameter	Material	Value
T_S, T_D	Thickness- S/D Contacts	Gold	20 nm
T_G	Thickness-Gate electrode	Aluminum	80 nm
T_{OSC}	Thickness-Semiconductor	Pentacene	30 *nm*
T_{DE}	Thickness-Dielectric	$La_{0.85}Nb_{0.15}O_y$	85 nm
L	Length-Channel		30 μm
W	Width-Channel		*195* μm

choosing Pentacene as the material for the organic semiconductor layer given the fact that it is extremely stable with respect to the environmental fluctuations and significantly high mobility.

Various tiers of a BGBC architecture are composed of diverse materials and thicknesses, as demonstrated in table 1. Although, TC and BC structures show the same I_d-V_{ds} characteristics as depicted in figures 3(a) and (b), also TC structures exhibit greater mobility and density of current as compared to the BGBC assembly, yet the favored assembly is bottom contact with a reason that it is economically more viable owing to a smoother formulation.

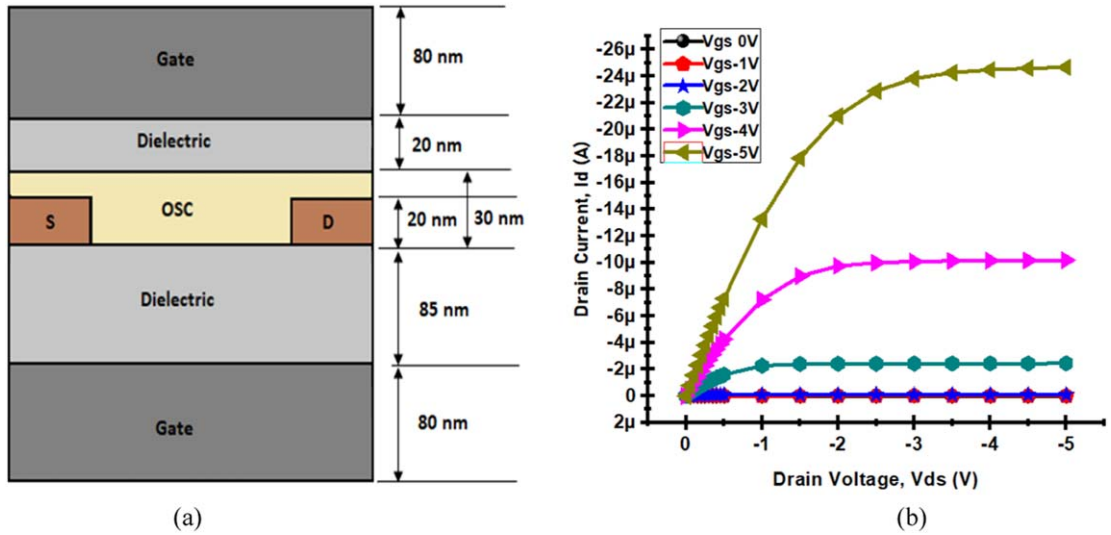


Figure 4. Dual Gate BGBC (a) Structure (b) I_d-V_{ds} characteristics.

Table 2. Evaluation factors for BGTC and BGBC.

Parameters	BGTC	BGBC
Threshold Voltage (V_T)	−1.9 V	−1.9 V
Mobility (μ)	4.82 $cm^2 Vs^{-1}$	4.63 $cm^2 Vs^{-1}$
Drain Current (ID)	18.5 μA	5.1 μA

Furthermore, channel length of BGBC (L = 30 μm) can be easily varied resulting in devices that are smaller scaled. For organic TFTs, channel thickness ($T_{Channel}$) goes from the first monolayer formed at the interface of the OSC-Dielectric to the second monolayer which is usually 1 or 2 nm from the said interface. Besides, there is a non-linear upgradation of contact resistance with upgrading thickness of the organic film which is unlikely for the structures with contacts at the top [13]. Additionally, due to assemblage of gate electrode metal, the structures encounter certain flaws, making BGBC a more endorsed choice [11].

Comparing the parameters derived for BGTC and BGBC assemblies at V_{GS} equal to −5 V, it was observed that BGTC showed superior functioning in terms of threshold voltage, mobility and drain current. The values of the extracted parameters have been summed up in table 2 below:

The table depicts that mobility of BGTC device is better than that of the BGBC device, though the threshold voltage is apparently the same for both. Moreover, the BGTC shows a drain current which is 3.64 times that of the BGBC. Besides, upon varying the doping concentration of both the configurations from 10^{12} - 10^{16} cm^{-3}, where for BGTC I_D changes by a factor of 8.6, the same for BGBC changes by a factor of 27 ranging from −0.19 μA to −5.1 μA upon reaching the doping strength of 10^{16} cm^{-3}. Though BGTC shows much fascinating performance as compared to the counterparts, yet its fabrication happens through vacuum or thermal processing which poses a lot of challenges. On the contrary, BGBC devices are fabricated using solution processing which makes it easy to be formed over the flexible substrates. Therefore, bottom gate devices prove to be a safer choice and hence are favored.

2.2. Single Gate BGBC versus Dual gate BGBC OTFTs

OTFT's may be fabricated either with single gate [SG] or dual. Dual gate [DG] devices mean that they employ an additional set of gate electrode and dielectric layer. Thus, giving a structure to the device that has one gate and dielectric at the bottom and the other at the top.

In comparison to the single gate devices, dual gate ones, as depicted in (a) and (b) parts of figure 4, prove to be promising on the front of performance, showing better control of threshold voltage, charge carrier mobility, current on–off ratio, sub-threshold slope and transconductance [12]. Table 3 precises the variation in performance parameters for both the configurations.

DG configuration yields better response since bottom gate governs the organic TFT intensely and the rear gate (also called the top gate) escalates V_T value linearly owing to its weak coupling with the channel. Also, it acts as a barrier to prevent further deterioration of the performance because of the adverse ambient conditions [13].

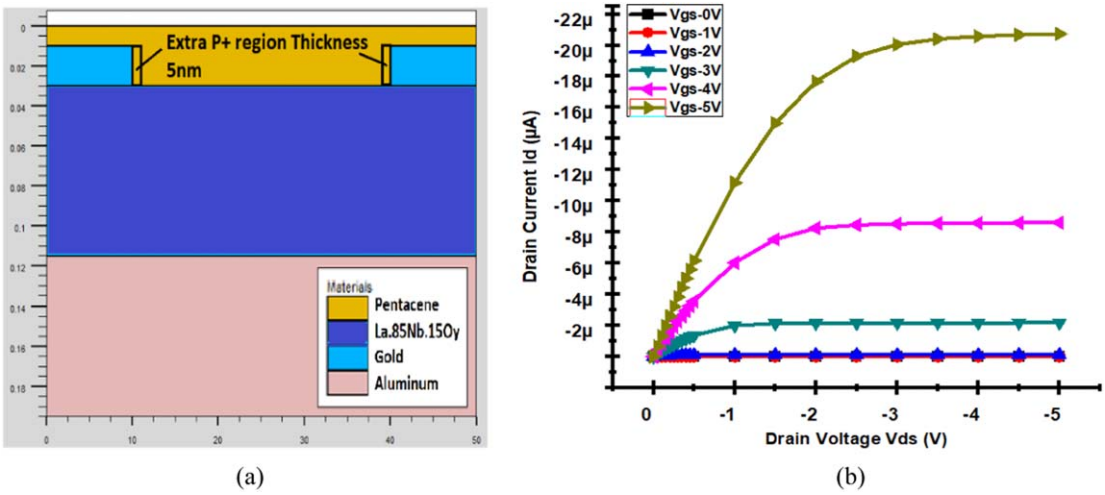


Figure 5. (a) Structure of OFFT with extra P^+ region (b) I_d-V_{ds} characteristics of the device with extra P^+ doping.

Table 3. Performance considerations for SG and DG BGBC OTFT.

Device Parameters	SG BGBC OTFT	DG BGBC OTFT
Threshold Voltage (V_T)	−1.9 V	−0.8 V
Sub-threshold Slope (SS)	0.38 mV/Decade	0.26 mV/Decade
Mobility (μ)	4.63 $cm^2 Vs^{-1}$	5.4 $cm^2 Vs^{-1}$
$I_{ON/OFF}$	6.24×10^6	4.7×10^7
Transconductance (g_m)	1.4 μ siemens	2.5 μ siemens

Nevertheless, the extra gate fabricated on organic semiconductor contaminates OSC layer. Besides, with the addition of gate, cost of the device rises further, therefore, hindering the economic viability of the device. Consequently, single gate devices are picked over the dual gate devices.

2.3. Single Gate BGBC OTFTs with extra P^+ doping in the vicinity of S/D electrodes

Section 2.2 provides ample proof that though a DG device performs better, its selection gets compromised due to higher fabrication expenses and increased contamination. Hence, SG is chosen for the construction of the device. Nevertheless, with the performance it shows, its advancement is desired.

Seemingly, BGBC assembly demonstrates poor performance with respect to drain current, transconductance and charge mobility when compared with the top contact devices [2–5]. This inequality exists due to the resistance posed by the contacts because of the barrier created at the interface of the electrodes and the dielectric material. Another evident reason for this is the semiconductor morphology surrounding the S/D [4]. An average performance is presented by the BC devices as the contact resistance existing between the interface of gold source and drain and the pentacene layer is considerably high alongside possessing petite size of the molecular grains. This results in an adverse effect leading to significant contact barrier. This structure also falls prey to lesser values of current owing to the construction wherein there is reduced area for the injection of the current thereby resulting in a considerably high barrier at the contact [4, 5]. Due to the evident morphological disorders, gaps, and surface traps creep-in into organic semiconductor layer, further hampering the charge mobility [14]. Such alterations observed in the intrinsic characteristics are principally because of the scarcity of charge carriers at the area adjoining source/drain and the channel. This scarcity leads to a significant decline in the voltage in the BGBC structures [15].

It has been demonstrated multiple times that at area surrounding the border of source and drain electrodes with channel, an augmented decline in the potential is pragmatic especially in the BGBC structures with charge concentration dropping to as less than $10^{17} cm^{-3}$ which anyhow must be more than 10^{18} cm^{-3} to retain increased performances. Hence, an insertion of extra doping (10^{20} cm^{-3}) zone is done close to S/D to resolve resistance at the contact and a meagre concentration of charge carriers [16] as shown in figure 5(a). This improves the VI characteristics of the structure as presented in (b) of figure 5.

Furthermore, this enhanced doping not only upgrades the VI characteristics but also the other important parameters as summed in the table 4 below:

Table 4. Comparison of performance parameters for SG OTFT with and without extra P^+ doping near S/D.

Parameters	SG OTFT without extra p+ doping	SG OTFT with extra P^+ doping
Threshold Voltage (V_T)	−1.9 V	−0.68 V
Sub-threshold Slope (SS)	0.38 mV/Decade	0.31 mV/Decade
Mobility (μ)	4.63 $cm^2 Vs^{-1}$	5.47 $cm^2 Vs^{-1}$
$I_{ON/OFF}$	6.24×10^6	3.9×10^7
Transconductance (g_m)	1.4 μ siemens	2.52 μ siemens

It is evident that when Au (S/D) is placed directly in contact with a semiconductor (pentacene), then Schottky barrier gets generated which leads to a remedial conduct of contacts possessing electrical properties. This occurs when a p-type semiconductor and the work function it possesses is larger as compared to the work function of the metal. The research that has been done on pentacene based OTFTs in past numerous years, proves that the combination of gold and pentacene have been often used in OTFTs as the electrode and semiconductor material respectively because of the high work function of gold and the high mobility offered by pentacene [4, 5]. Nonetheless, while forming the source and drain electrodes by deposition of gold onto pentacene (active layer), a relatively large hole-injection barrier of 0.8–1 eV from gold to pentacene is observed [9], which is much higher than expected [17]. Also, there is a shadow that gets created near the interface due to which there is a significant decrease of the charge carriers in the morphological region. Therefore, to eliminate this effect and to attain superior operation, the injection barrier must be very small. It has been depicted by the work of many researchers that upon the insertion of extra p+ charge under controlled circumstances may elevate the injection of hole by leaps and bounds thereby assisting in obtaining optimal conductivity to alleviate the decline in the potential at the interface [2]. Upon deposition of gold in the pentacene layer, because there is an evident and apparent drop in charge density near the interface, which is observed to be 10^{17} cm^{-3} against the necessary 10^{18} cm^{-3} to maintain the desired performance, therefore under controlled conditions, a small 5 nm region of extra p+ doping is created keeping the concentration of added p+ charge as 10^{20} cm^{-3}. This results in a swooping increase in the drain current of the BGBC device by 13.8% after the extra doping as indicated in the figures 3(b) and 5(b) respectively for without and with p+ doping. Therefore, upon employing any semiconductor material that shows high charge carrier availability, or by incorporating multiple p type materials to enhance the mobility of the interface area of S/D with the OSC, the mentioned p+ doping may be achieved. One such material can be fullerene (C_{60}) which causes high charge surge. The insertion of material with high mobility and high charge carrier concentration becomes the chief objective and is termed as doping to refine the contact-semiconductor interface that usually encounters the contact effect generated due to work function difference of two dissimilar materials.

2.4. Ditch incorporated Single Gate BGBC OTFTs with extra P^+ doping

To the structure of a SG BGBC with an additional P^+ region near the electrodes; figure 5(a), a ditch of 40 nm is created in the OSC, as shown in figure 6(a) to further enhance the performance of the device. Because in bottom contact structures, upon the addition of extra doping concentration near S/D, carriers tend to move in horizontal direction, horizontal component of electric field becomes dominant over the vertical one, hence reducing the force essential to extract the parameters, thereby leading to a difficult charge extraction. It is evident from figure 6(b) that by incorporating a ditch of organic semiconductor material, the charge carrier injection gets relatively augmented due to resilient upright electric field. This results in obtaining high drain current which acts in favor of proving extra P^+ region BGBC with ditch as an ideal option. Moreover, since the contact electrodes in the ditch structure are formed before the layer comprising active material, these can be fabricated with a lot of ease like the structure without any ditch.

The idea of incorporating the ditch into the OSC of a SG BGBC with spare P^+ doping yields the desired results. The performance of this device fairly competes with the DG BGBC and displays some marvelous values of V_T, SS, μ, g_m, and $I_{ON/OFF}$. The derived values have been outlined in the table 5 below:

The values of parameters derived and listed in table 5 suffice for its employability in further applications to attain augmented performance. Consequently, it may be inferred that single gate OTFT may be in devices like inverters, provided it is fabricated having excessive P^+ charge in the vicinity of electrodes and upon incorporating a ditch created into pentacene [16].

As far as the proposed TFT and comparatively high values derived pertaining to certain important parameters are concerned, the reference of the proposed transistor has been taken from the research done by C. Y. Han *et al* in their paper titled, 'High-Mobility Pentacene Organic Thin Film Transistor with $La_xNb_{(1-x)}O_y$ Gate Dielectric Fabricated on Vacuum Tape.' [18]. This work validates the device in the previous work published [5] and therefore it can be inferred that the device is highly reliable and stable. Furthermore, this work strives to

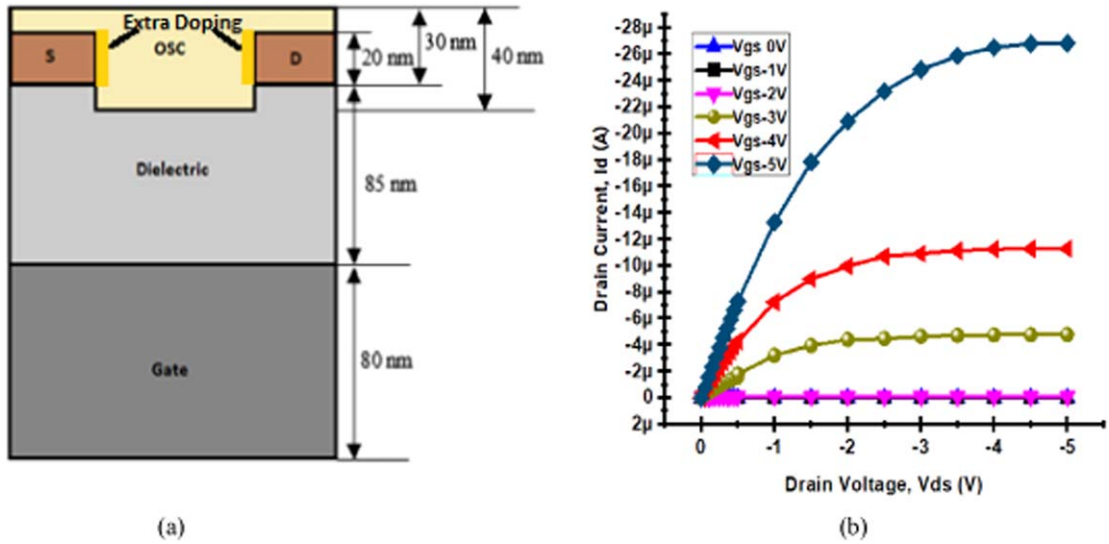


Figure 6. (a) Representation of Extra P^+ region OTFT with Ditch (b) I_d-V_{ds} Characteristics.

Table 5. Comparison of performance parameters of various OTFT device configurations.

Parameters	SG OTFT	DG OTFT	SG OTFT with extra P^+ doping	Ditch incorporated SG OTFT with extra P^+ doping
Threshold Voltage (V_T)	−1.9 V	−0.8 V	−0.68 V	−0.57 V
Sub-threshold Slope (SS)	0.38 mV/Decade	0.26 mV/Decade	0.31 mV/Decade	0.36 mV/Decade
Mobility (μ)	4.63 $cm^2 Vs^{-1}$	5.4 $cm^2 Vs^{-1}$	5.47 $cm^2 Vs^{-1}$	5.95 $cm^2 Vs^{-1}$
$I_{ON/OFF}$	6.24×10^6	4.7×10^7	3.9×10^7	3.52×10^7
Transconductance (g_m)	1.4 μ siemens	2.5 μ siemens	2.52 μ siemens	2.57 μ siemens

improve the already validated device [5] to augment the performance using numerous ways to remarkably increase the performance of the devices like to fill the traps present in the geometry of the devices, as presented in 1998 by Horowitz *et al* in their work titled, 'The concept of threshold voltage in organic field-effect transistors.' Also, for the device to show enhanced stability, better mobility, and a smooth surface, Pentacene proves to be an apt choice of OSC layer as it is by far the most opted organic semiconductor. The work aims to propose materials with high mobility. Once proposed, we may use it for design of devices(memory) using the proposed device.

3. Organic inverter

As a preface, organic thin film transistor-based circuits have been well-thought-out as one of the exceedingly favorable contenders to create a modest paradigm catering to outsized organic devices. [19]. Inverters are a rudimentary yet significant integrant of a few such circuits [20]. Ensuring apt functioning of the inverter to yield simplicity of design, robustness, and truncated power ingestion, requires keen consideration being put on the noise margin and the delay in propagation. Till date, numerous researchers have invested the effort and time in constructing inverters based on organic materials proving that complementary inverters produce better results but require technical proficiency [21]. For developing circuits that are complementary in nature either n and p-type transistors can be assimilated or two analogous (both n or both p) transistors may be employed on a common substrate [4]. Given the fact that n-type OSC exhibits inferior mobility contrasting p-type ones; there is an evident dip in the performance of entirely complementary organic inverters. Moreover, such circuits are hypersensitive to moisture, air and oxygen as compared to their p-type complements [9]. Consequently, organic semiconductors with p-type configuration endorsed for the construction of economically viable organic circuits owing to their exceptional stability and mobility [22–24]. Furthermore, challenges in finding an appropriate n-channel thin film transistor material steered the proposal of an all p-organic inverter [25, 26]. Upon directing research towards such devices, when their short comings came into limelight, hybrid complementary inverters came into existence [21].

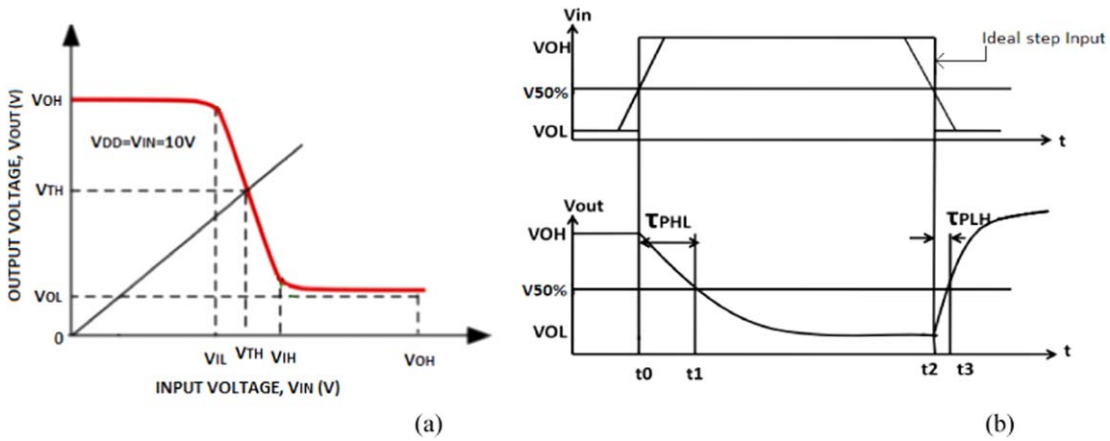


Figure 7. (a) Voltage transfer characteristics of a classical inverter (b) I/O voltage waveforms.

Figure 7(a) and (b) show the voltage transfer characteristics and I/O waveforms of a classical inverter respectively. Stability of the inverter, based logic circuits is governed through Noise Margin which expands to extreme admissible spurious signal that a device whilst being employed in a system, can entertain nevertheless yielding legitimate operation.

Low-state Noise margins (NM_L) and High-state Noise margins (NM_H) that bring about reliability and robustness in the circuits [27] must be adequately positive to obtain steady application. These values are dependent on the values of minimum input and output high voltage (V_{IH} and V_{OH} respectively) and maximum input and output low voltages (V_{OL} and V_{IL} respectively), which can be mathematically expressed as per equations (1)–(5).

$$NM_H = V_{OH} - V_{IH} \tag{1}$$

$$NM_L = V_{IL} - V_{OL} \tag{2}$$

Where,

$$V_{in} \leqslant V_{IL} \rightarrow V_{out} \geqslant V_{OH} \tag{3}$$

$$V_{in} \geqslant V_{IH} \rightarrow V_{out} \leqslant V_{OL} \tag{4}$$

$$V_{IH} > V_{IL} \tag{5}$$

Here, V_{in} and V_{out} are the input and output voltage respectively [27].

In ideal transfer characteristics, to ensure that the push - pull transistors are balanced, switching from high logic state to low logic state transpires precisely in accordance with equation (6):

$$NML = NMH = V_{DD}/2 \tag{6}$$

To attain this condition, trans conductance ratio β_R of the p-type and n-type devices needs to be worked at. Moreover, the logic threshold voltage of the inverter, V_{TH} can be devised as [24]

$$V_{TH} = V_{Tn+}\left(\frac{\sqrt{\frac{1}{\beta_R}}(V_{DD} - |V_{Tp}|)}{\left(1 + \sqrt{\frac{1}{\beta_R}}\right)}\right. \tag{7}$$

Where, threshold voltages of n and p type transistors are shown as V_{Tn} and V_{Tp} respectively. Besides, to establish the delay in the signal, while transpiring from high logic state to low, propagation delays τ_{PHL} and τ_{PLH} are made used of. Given the step input having zero rise and fall ties, the mentioned delays, may be logically obtained as:

$$\tau_{PHL} \quad (0) = \frac{C_{TL}}{\beta_n(V_{OH} - V_{T_n})}\left[\frac{2V_{T_n}}{V_{OH} - V_{T_n}} + \ln\left(\frac{4(V_{OH} - V_{T_n})}{V_{OH} + V_{OL}} - 1\right)\right] \tag{8}$$

$$\tau_{PLH}\ (0) = \frac{C_{TL}}{\beta_p(V_{OH} - V_{T_{OL}-} | V_{T_p} |)} \times \left[\frac{2 | V_{T_p} |}{(V_{OH} - V_{T_{OL}-} | V_{T_p} |} + \ln\left(\frac{4(V_{OH} - V_{T_{OL}-} | V_{T_p} |)}{V_{OH} \quad V_{OL}} - 1\right)\right] \tag{9}$$

In equations (8) and (9); τ_{PHL} (0) and τ_{PLH} (0) equate with the delays encountered during propagation upon the application of an ideal step input, C_{TL} refers to the total load capacitance, β_n and β_p are the trans conductance of p type and n type devices. Considering a finite rise and fall time (τ_r and τ_f) in the input signal, the actual delay in propagation, τ_{PHL} (1) and τ_{PLH} (1), maybe then calculated using the following equations:

Table 6. Transistors based on different materials used for fabricating the varied layers.

Transistors	Dielectric	Gate	S/D Electrode	Semiconductor
Transistor 1	Al_2O_3+SiO_2	n^+ Si	Aluminum (Al)	Pentacene (p type)
Transistor 2	SiO_2	n^+ Si	Au	CuPc (p type)
Transistor 3	SiO_2	p-Si	Au	Fe_{16}CuPc (n type)
Transistor 4	$SiO_2 + TiSiO_2 + SiO_2$	p-Si	LiF + Al	Fullerene (C_{60}) (n type)
Transistor 5	SiN_x	p-Si	Silicide	a-Si:H (n type)
Transistor 6	Al_2O_3+SiO_2	p-Si	Al	ZnO (n type)
Transistor 7 (Proposed)	$La_{(1-x)}Nb_xO_y$	Al	Au	Pentacene (p type)

Table 7. Design of Transistors based on different materials used for fabricating the varied layers.

	Thickness					
Transistors	Dielectric (T_{DE} nm)	Gate (T_G nm)	Electrode (T_S/T_D nm)	Semiconductor (T_{OSC} nm)	Length (L μm)	Width (W μm)
Transistor 1	10 + 2.5 (12.5)	10	10	50	90	500
Transistor 2	300	10	10	10	50	1000
Transistor 3	300	10	10	10	50	1000
Transistor 4	8 + 132 + 4 (144)	30	1 + 100 (101)	60	100	1000
Transistor 5	300	80	30	250	40	80
Transistor 6	10 + 2.5 (12.5)	10	10	60	90	500
Transistor 7 (BGTC with proposed design)	85	80	20	30	50	215 (nm)
Transistor 8 (Dual Gate BGBC with proposed design)	85 (D1) + 20 (D2)	80 + 80	20	30	50	297 (nm)
Transistor 7 (Proposed Single Gate BGBC)	85	80	20	40	50	195 (nm)

$$\tau_{PHL} \quad (1) = \sqrt{\tau_{PHL}^2(0) + \left(\frac{\tau_r}{2}\right)^2} \tag{10}$$

$$\tau_{PLH} \quad (1) = \sqrt{\tau_{PLH}^2(0) + \left(\frac{\tau_f}{2}\right)^2} \tag{11}$$

Inverters may be constructed using various device combinations viz.: fully organic complementary inverters with organic p-type and n-type device, hybrid complementary inverters using organic p-type and inorganic n-type device and all p-type complementary inverters employing organic p-type devices [28]. A brief comparison of these configurations has been made in table 6. Distinctively, n-type organic SC exhibits lower mobility than the p-type counterparts affecting the performance of fully complementary organic inverter circuits. Additionally, these are more sensitive to oxygen and air moisture as well, making them an even more vulnerable choice. The inorganic material used in hybrid circuitry, also has its own demerits leading to choosing an all p-type organic complementary inverter as the most preferred choice. Owing to these factors, the outcome of numerous research was not very astonishing when p-type devices were favored for the fabrication of low-cost circuits because of exemplary stability and mobility shown by the material.

Below tables (6–7) give a glimpse of some verified and validated material combinations and device dimensions of the different layers (Dielectric, gate, source and drain contact electrodes and the semiconductor layer) used for designing the transistors [16]:

Based on the transistors as shown in table 7, upon making varied material combinations, inverters may be formed and can be categorized as Fully Organic (both p and n type transistors used to make inverter are organic; in accordance with the tables 2, 7 inverter combinations are possible: Pentacene-C_{60}, CuPc-Fe_{16}CuPc) Complementary Inverters, Hybrid Complementary (material of p- type transistor is organic and that for n- type is inorganic: Pentacene-a-Si:H, Pentacene-ZnO) Inverters and Organic all-p (both transistors are organic p-type: Pentacene-Pentacene) Inverters. The critical parameters that distinctively decide the performance of the device for the three categories have been summed up in the table 8 below:

Though TFTs are an obvious and simple choice for designing inverters, there may be other ways for doing the same. Since the recent times, flexible electronics is witnessing the use of vertical organic field effect transistors (VOFET) [29], FinFET [30], [31] and GAAFET [32]. The three architectures against the OTFT are as compared below:

Table 8. Comparison of various device combinations of inverter circuits.

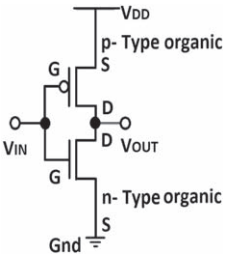
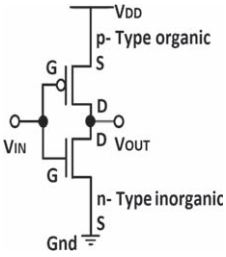
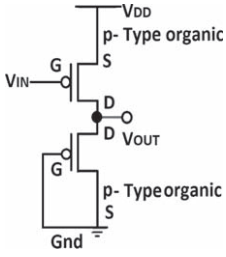
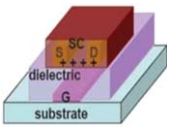
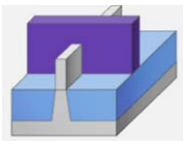
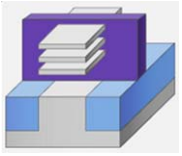
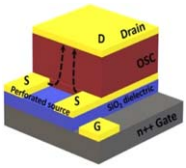
Parameter	Fully organic complementary inverter	Hybrid complementary inverter	All p organic inverter
Circuit Schematic			
NM_H	$=V_{DD}/2$	$<V_{DD}/2$	Very low
NM_L	$<V_{DD}/2$	$<V_{DD}/2$ (=NM_L of fully organic complementary inverter)	$\sim V_{DD}/2$
Propagation delay	High	High	Low
V_{TH}	$<V_{DD}/2$	$\geqslant V_{DD}/2$	$>V_{DD}/2$

Table 9. Comparison of OTFT with various upcoming architectures of transistor to be used for inverter design.

Variables	OTFT	FinFET	GAAFET	VOFET
Structure				
Structure	Gate surrounds only one side of the channel	Gate surrounds three side of the channel	Gate surrounds all the sides of the channel	Gate surrounds only one side of the channel
Placement of S/D	Horizontal	Horizontal	Horizontal	Vertical
OSC thickness to serve as channel	Thick	Thin	Nano scale	Thick
Geometry	Very Simple	Complex	Very Complex	Complex
Fabrication Cost	Very Low	High	Very High	High
Channel Length	$\geqslant 1\ \mu$m	$\geqslant 70$ nm	$\geqslant 50$ nm	$\leqslant 1\ \mu$m
Leakage current	Moderate	Low	Negligible	Very low
Power Consumption	Low	Very low	Ultra-low	Very low
Drive Strength	Dependent on channel width	Can be increased by having multiple or longer fins	High, can be increased by having multiple sheets	High due to vertical channel
Current Density	Moderate	High	Very High	High
Operating speed	Moderate	High	Very high	High
Charge Transport Direction	Parallel	Parallel	Parallel	Vertical

Though, OTFTs are not yet performing well enough for commercial usage because of lack of performance pertaining to operation speed, parasitic device capacitance, fabrication complexity on flexible substrates, device-to-device variations, and bias stability. However, since organic transistors enable truly flexible electronics, OTFTs show great potential to be used in flexible devices in the future such as active-matrix displays [2], sensor arrays [3, 4], or radio-frequency identification tags [5].

Looking at the table 9, though it is apparent that the OTFT lags in many ways from the latest and upcoming architectures, yet these are being used more as compared to the counterparts, specially while fabricating an inverter, because their geometry is very simple and are cost effective when it comes to fabrication. Moreover, not much work has been done on the upcoming architectures which could prove that their stability is better than the well-established stability of the OTFT.

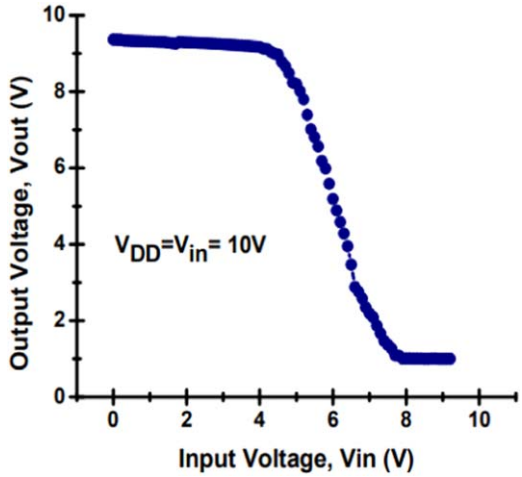


Figure 8. VT characteristics of an all p, SG BGBC Organic Inverter.

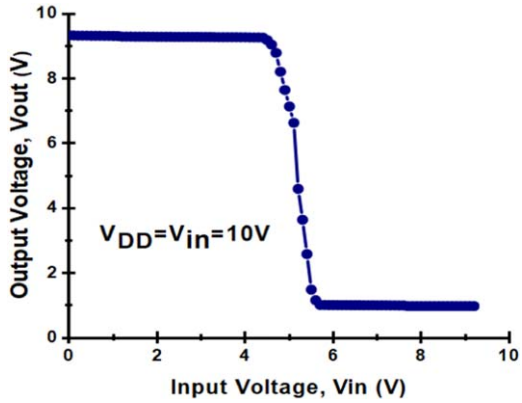


Figure 9. VTC of Inverter using SG OTFT with extra P^+ doping.

3.1. Inverter design using single gate BGBC OTFT

As depicted in table 8, an all-p organic inverter has NM_H towards the low side wherein NM_L meets the desired standard. In this section, a single gate bottom contact OTFT is used to design an inverter circuit. The VT characteristics of such an inverter are presented in figure 8.

Analyzing the characteristics and using equations (1) and (2), the respective values of NM_H and NM_L are found to be 1.25 V and 3.22 V. Ideally, the noise margin values must be near equal to $V_{DD}/2$ (equation (6)) which distinctively dies not match. Because of this, an OTFT with a 40 nm ditch has been incorporated for the design of the inverter.

3.2. Inverter design using single gate BGBC OTFT with extra P+ region in the vicinity of S/D Contacts

Designing an all-p inverter using extra doping concentration device as explained in section 2.3 aids in upgrading the performance of the device, yielding results depicted in figure 9.

On employing this device, the high and low noise margins received bettered as compared to those established from the device without extra P^+ region, witnessing an increase of 2.31 V in NM_H and an increase of 0.16 V in NM_L making them to be 3.56 V and 3.38 V respectively.

3.3. Inverter design using single gate ditch OTFT with extra P^+ region.

A 40 nm ditch incorporated in the pentacene organic semiconductor layer of the OTFT which has an additional P^+ region concentration of 10^{20} cm^{-3}, is as shown in figure 6(a). For creating a ditch, a 10 nm recession is created in the gate dielectric and then pentacene is deposited in it. Employing ditch ascertains the solution process alongside augmenting the performance of the device.

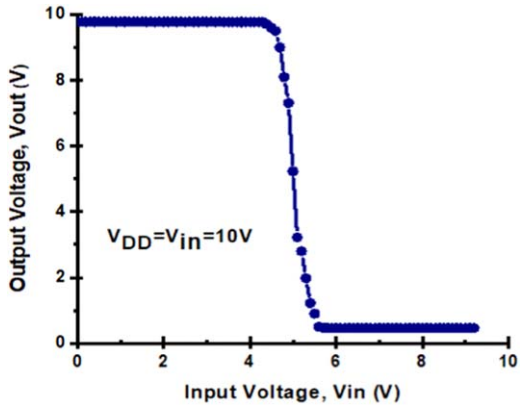


Figure 10. VTC of Inverter using Ditch in OSC in SG OTFT with extra P^+ doping near S/D.

Upon adopting such a device for the designing of the inverter, NM_H was calculated to be 4.17 whereas NM_L came out to be equal to 4.0 V. In comparison to the inverter without ditch, this one yielded improved noise margins.

The attained values verify that the inverter designed using a 40 nm ditch in the OSC of a single gate bottom contact OTFT which has a superfluous doping concentration near the electrodes, generates superior performance and therefore, may be employed in organic circuits aiming towards boosted performance as indicated in figure 10 [31]. A comprehensive comparison of the noise margins for different configurations used to design the inverter has been done in table 10.

3.4. Comparison of all p organic inverter comprising Zero-V_{gs} load logic (ZVLL) and diode load logic (DLL)

Organic device quoted in section 3.3 may be further categorized using two different conformations of the circuit viz.: Zero-V_{gs} Load Logic (ZVLL) and Diode Load Logic (DLL) [33]. These are differentiated based on the association of gate, drain and source terminals with each other as depicted in figure 11. These conformations have contrasting culminations and may be employed for varied objectives. While ZVLL configuration is selected for applications needful of advanced switching, DLL configuration is designated to devices demanding minimal extraneous noise.

Both the source and drain terminals in a diode load logic (DLL) conformation are grounded. Although the operation of driver and load OTFTs takes place in the enhancement mode, yet they may function in depletion mode as well. Inverters using DLL configuration are swift in operation as compared to the counterparts.

For making a consistent current source between V_{DD} and ground, the source and gate terminals of the transistor are shorted in the ZVLL configuration. In this conformation, the driver functions in the enhancement mode whereas the transistor connected as load, functions in the depletion mode. This arrangement gives the ZVLL configuration an upper hand over the DLL one in terms of voltage swing and noise margin. Hence making the device sturdier. The resistance of the channel of the load is way higher than that of the driver transistor at lower values of $V_{IN.}$ This results in switching the driver to 'ON' state, consequently making V_{OUT} approximately equal to V_{DD} [33]. In contrast to this, if the values of V_{IN} are set towards the higher side than, the driver and load develop a zero gate to source condition which switches both the transistors to 'ON' state. This results in the augmentation of the aspect ratio (W/L) of the load to five times as compared to that of the driver.

Aspect ratio acts a key feature while the inverter conformation is selected and designed. This yields into the maximization of noise margin, besides providing equal current to the source and sink terminals. Thereby allowing the charge and discharge of the capacitive load in identical time.

Analyzing the single gate device employing Atlas 2-D numerical device simulator [31] using the mobility genre proposed by Poole–Frenkel to essentially examine the stationary and active performance of the electronic circuits. Using the model, mobility may be articulated as:

$$\mu(E) = \mu_{\circ} \exp\left[-\frac{\Delta}{kT} + \left(\frac{\beta}{kT} - \gamma\right)\sqrt{E}\right] \tag{12}$$

Wherein, $\mu(E)$, $\mu_{\circ}$ *and* E respectively signify mobility dependent upon mobility, null field mobility and electric field, Δ which for pentacene is equal to 1.792×10^{-2} eV represents activation energy, β (for pentacene is 7.758×10^{-5} eV $(cm/V)^{0.5}$) is hole Poole–Frenkel factor, and γ represents the fitting parameter. In accordance with equation (12), mobility is directly proportional to gate voltage which is responsible for the hopping transport phenomenon in the organic transistors [33]. As shown by equations [1] and [2], V_{OH} needs to be set to

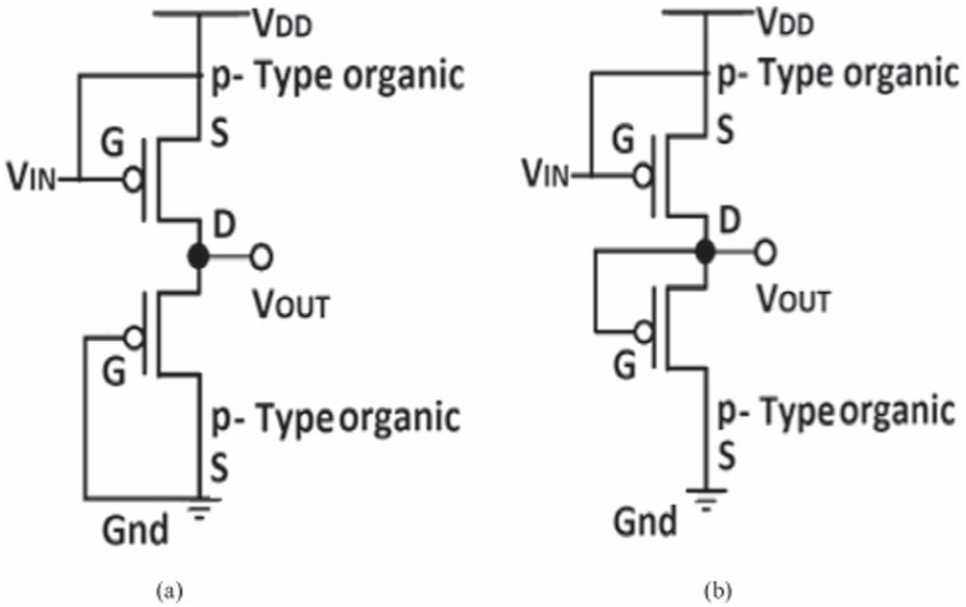


Figure 11. SG-OTFT based all p-type Inverter conformations (a) DLL (b) ZVLL.

Table 10. Comparison of Noise Margins of different Inverters.

Configuration	V_{OH} (V)	V_{IH} (V)	V_{IL}(V)	V_{OL}(V)	NM_H (V)	NM_L (V)
Inverter using OTFT without P+ region	9.05	7.8	4.3	1.08	1.25	3.22
Inverter using OTFT with P+ region	9.26	5.71	4.4	1.02	3.56	3.38
Inverter using Ditch OTFT with P+ region	9.77	5.6	4.5	0.5	4.17	4.00

Table 11. Dynamic Performance parameters of single gate all p-type inverter in DLL and ZVLL configurations.

Parameters	τ_{PHL} (μs)	τ_{PLH} (μs)	τ_p (μs)
DLL	32	21	27
ZVLL	103	64	84

high magnitude which must be as near as possible to V_{DD} for obtaining high values of noise margin. Ideally, peak attainable value of the output high voltage is restricted to the difference between V_{DD} and the threshold Voltage V_T because of the drop in V_T of the device. To attain augmented performance, V_T and the potential drop need to be compromised. As suggested by figure 12, the transient responses of DLL and ZVLL prove that in terms of the performance, ZVLL is way beyond DLL hence proving it to be a more chosen option.

Table 11 provides a synopsis of the dynamic performance parameters of single gate all p-type inverter in Diode Load Logic and Zero-V_{gs} Load Logic topologies. The table seconds the work of Myny K. [34] indicating that the ZVLL that primarily operates with devices that are normally ON, though shows a slower switching action as compared to the DLL configuration, yet compensates the same by having higher gain, enhanced noise margins and augmented robustness [35].

Upon comparing the values of the decisive parameters derived from the proposed device with the already reported inverter configurations, it can be inferred that the proposed device has better gain than all other configurations barring the hybrid inverter made from Pentacene-ZnO combination. Nonetheless, when the noise margin of the proposed and the Pentacene-ZnO hybrid inverter is compared, the proposed device excels, alongside having almost equal average propagation delays viz. 84 μs and 83 μs respectively for proposed and mentioned hybrid inverter. A comparative study of inverters designed using different transistors has been done in table 12 proving that the chosen material and configuration are an apt choice for the inverter design.

This device viz. all-p type organic inverter in ZVLL configuration using a ditch and extra p+ doping based OTFT has been simulated using ATLAS device simulating software. Though, for any further applications if this

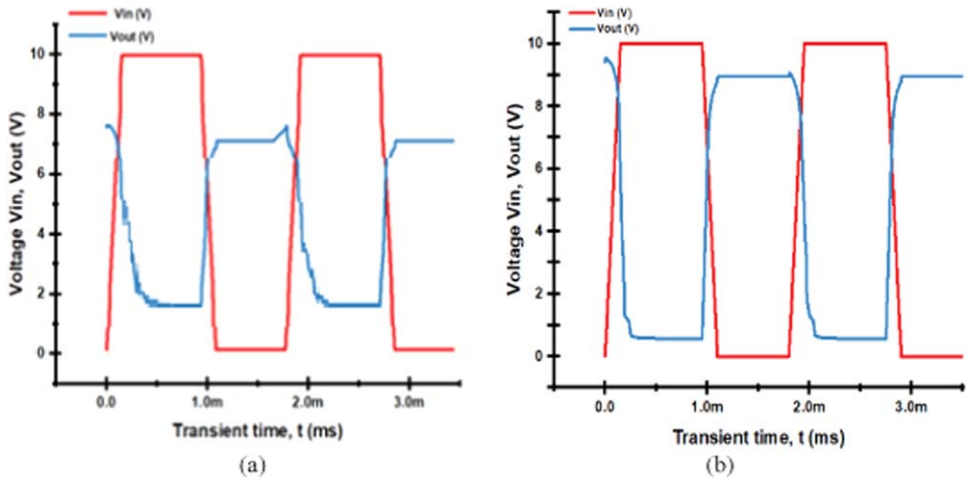


Figure 12. Transient response of (a) DLL and (b) ZVLL

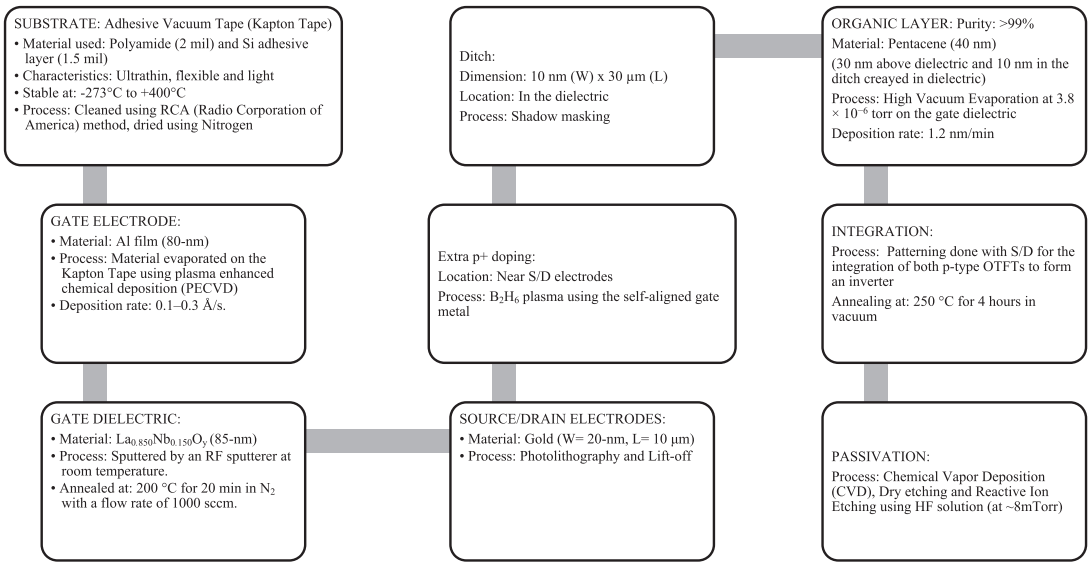


Figure 13. Suggestive steps for the fabrication of the simulated all-p type organic inverter in ZVLL mode using extra p+ doping and ditch based OTFT.

device has to be fabricated in future, then the steps reflected in figure 13 may be followed [36]. First, an 80 nm Al buffer layer needs to be deposited by plasma enhanced chemical vapor deposition (PECVD) with a deposition rate of 0.1-0.3 Å s^{-1} on a substrate made using adhesive vacuum tape (Kapton Tape) comprising 2 mil polyamides initially adhered to 1.5 mil Si adhesive cleaned using RCA (Radio Corporation of America) method and dried using nitrogen atmosphere. This can be selected as the substrate material owing to its flexibility, light weight, and ultrathin dimension [37]. This step may be then followed by the deposition of an 85 nm $La_{0.85}Nb_{0.15}O_y$ film by sputtering carried at room temperature by a Denton Vacuum LLC Discovery 635 RF sputterer by means of Nb and La2O3 targets in an Ar/O_2 (24/6 sccm) ambient, controlling its Nb content by applied erratic sputtering powers of the targets. The Annealing of this gate dielectric can be performed at 200 °C for 20 min in N_2 ambient. Then this layer needs to be crystallized by BLA and patterned to form active islands. The apt laser beam size can be 520 μm × 20 μm for BLA with a scan speed of 400 mm s^{-1} and laser power density of 110 kW cm^{-2}. Next, a 10 μm (L) and 20 nm (W) gold layer may be then deposited on the dielectric to serve as the contact source and drain electrodes for both p-type TFTs, leaving a gap of 30 nm in between these two electrodes. Photolithography and lift off processes may be used for this deposition which can be later followed by the insertion of extra p+ doping (10^{20} cm^{-3}) near the source and drain regions with B_2H_6 plasma using the self-aligned gate metal. Thereafter, a ditch of 10 nm (W) and 30 μm (L) needs to be created into the dielectric using

Table 12. Design of inverters based on different transistors as depicted in table 7.

Device	Combination for device design	V_{OH} (V)	V_{IH} (V)	V_{OL} (V)	V_{IL} (V)	NM_H (V)	NM_L (V)	Gain	τ_{PHL} (μs)	τ_{PLH} (μs)	τ_p (μs)
Inv 1	Pentacene-C_{60}	8.8	6.3	0.12	4.8	2.5	4.7	−6.7	65	59	62
Inv 2	CuPc -Fe_{16}CuPc	8.7	7.0	1.2	4.6	1.7	3.4	−2.0	121	80	101
Inv 3	Pentacene- a-Si:H	9.0	6.7	0.9	4.0	2.3	3.1	−4.5	120	111	116
Inv 4	Pentacene-ZnO	9.3	3.8	0.02	4.5	5.5	4.5	−8.6	116	50	83
Inv 5	Pentacene-Pentacene	8.8	7.4	0.5	6.0	1.4	5.5	−7.0	120	64	92
Inv using BGTC with proposed design	Pentacene-Pentacene	9.1	7.2	0.08	5.9	1.9	5.8	−5.8	126	61	94
Inv using dual gate BGBC with proposed design	Pentacene-Pentacene	8.9	4.6	1.1	4.8	4.3	3.7	−3.2	153	42	98
Inv Proposed	Pentacene-Pentacene	9.77	5.6	0.5	4.5	4.17	4	−7.2	103	64	84

the shadow masking technique. Creation of ditch may then be followed by the consecutive deposition of a 30 nm thick p-type pentacene layer above the gate dielectric and 10 nm pentacene layer in the ditch created, without breaking the vacuum to ensure a clean interface. The deposition of the p-type pentacene as the organic layer may be performed by High Vacuum Evaporation using Edwards Auto 306 at 3.8×10^{-6} torr at a deposition rate of 1.2 nm min^{-1} being scrutinized by a quartz- crystal oscillator. Then the patterning of both p-type TFTs by sputtering with source and drain metal layer may be carried out followed by the same process for the second p-type TFT. The device needs to be then annealed at 250 °C for 4 h in vacuum. Finally, to protect the semiconductor layer, passivation may be performed using Chemical Vapor Deposition (CVD) technique accompanied by the process of dry etching and Reactive Ion Etching at ∼8mTorr engaging a HF solution [17, 18, 35, 38].

4. Result and discussion

As indicated in various sections of the paper, as compared to BGTC, bottom contact structure exhibits mediocre performance owing to high contact resistance between the metal-OSC and small grain size of molecules, which in turn affects the contact barrier.

Also, when these TFTs are employed for the design of inverters and are validated for both DLL and ZVLL configurations, then as compared to DLL, ZVLL shows better performance. The following tables (tables 13–16) suggest the reported and simulated findings:

Table 13 formed the basis of choosing BGBC over BGTC TFTs. Furthermore, as proposed, when extra doping along with ditch was done on the BGBC TFT, the parameters that are critical in governing the performance changed to a considerable value as depicted in table 14.

Moreover, when the device was used to make inverter, then as compared to DLL, and ZVLL all-p inverters, the ZVLL inverter with ditch and extra doping incorporated OTFT, proved to be an outstanding selection as shown in the table 15:

For the ZVLL configuration all-p inverters, as illustrated in table 16, the parameters like noise margin, propagation delay and threshold voltage also witnessed a striking hike in all the values of the simulated device as compared to the reported ones.

These tables make it evident that the proposed all-p type inverter in ZVLL mode using a SG bottom gate bottom contact OTFT with extra p+ doping and a ditch in the dielectric may be employed in the flexible electronics-based applications.

Considering that high gain is a desirable characteristic in organic inverters, and the proposed device has a gain towards the lower side, yet it is sustainable because in spite of the low gains, the inverter is still able to keep the apparent signals '1' or '0' and deliver them to the next logic unit. In the current work we have tried to realize the circuits and have inferred that such inverters may be employed to construct wearable electronic circuits and biosensors where high gain requirements are secondary [19]. In the future work, we shall further improve the performance of the all-p type inverter by optimizing the processing technology. And shall try simulating and validating a memory cell using the proposed device as well. It can be noted from tables 12 and 15 that the noise margin of the proposed device is close to the desired $V_{DD}/2$ value. Also, the propagation delay of the proposed device is less than the other inverter configurations. We accept that the gain of the device is low yet the fact that all the materials chosen for the design of the inverter are flexible cannot be ignored.

It is well understood that the performance and switching speed of any organic device cannot be even compared with the gain of the inorganic counterparts, but our aim behind the work is to make a device that is

Table 13. Comparison of parameters of reported BGTC and BGBC OTFTs.

Parameters	BGTC	BGBC
Mobility	High	Low
Transconductance	High	Low
Current	High	Low
Contact Resistance	Low	High
Fabrication	Difficult	Easier

Table 14. Comparison of TFTs with and without extra doping.

Parameters	BGBC without extra doping	BGBC with extra doping
Mobility	Low	Increased by 21%
Transconductance	Low	Increased by 110%
Current	Low	Increased by 13.8%
Contact Resistance	High	Decreased by 21.57%
Sub-threshold Slope	Low	Increased by 23.68%

Table 15. Comparison of all-p Inverters in DLL, ZVLL configurations against the proposed ZVLL.

Parameters	All-p DLL	All-p ZVLL	All-p ZVLL Proposed
τ_{PHL} (μs)	32	141	103
τ_{PLH} (μs)	21	86	64
τ_{P} (μs)	27	94	84
V_{TH} (V)	−6.57	−5.36	−4.29

Table 16. Comparison of reported all-p Inverters against the proposed configuration.

Parameters	All-p	All-p proposed
NM_H	1.25	4.17
NM_L	3.22	4.00
Propagation Delay (μs)	94	84
V_{TH} (V)	−6.57	−4.29

flexible and is suitable for application areas having large area material and flexibility as major players. In future we aim at making a memory cell using the proposed circuit [20].

5. Conclusion

This work presented a single gate organic all-p inverter in Diode Load Logic and Zero V_{gs} Load Logic topologies. Upon observing the static and transient responses of the inverter in both DLL and ZVLL modes, noise margins and transition time were extracted. The paper discussed the obvious choices of constructing an OTFT and an inverter proving that as compared to BGTC, BGBC devices are better options. Moreover, the VTC for both suggested that though the drain current of DG was 33% more than that of a SG device having a value of 24 μs as compared to an 18 μs of SG, still SG is considered superior owing to lesser contamination of OSC layer and better economic viability. Besides, to enhance the performance further, $p+$ doping was additionally inserted near the S/D electrodes augmenting the performance of SG BGBC from 18 μs to 21 μs showing a percentage increase of 16.67%. Nonetheless, to supplement the output a ditch was incorporated into the organic semiconductor layer raising the I_D further by 30.9% and resulting in a drain current of 27.5 μs. Furthermore, an organic inverter was later designed incorporating this device. Converging to using an *all-p* inverter, different topologies viz. DLL and ZVLL were employed. Summarizing the dynamic response of both the configurations it was concluded that ZVLL configuration was a preferred choice over the counterpart. With respect to the

propagation delay, ZVLL experiences 211% greater propagation delay than the DLL which is 84 μs and 27 μs for the configurations respectively. Inferring from the transient response, the noise margin of ZVLL may be calculated as 4.7 and that of DLL comes out to be 3.9, which makes it apparent that despite the higher propagation delay, as ZVLL showed 20.5% better performance with respect to noise margin and robustness, it demonstrated to be the apparent choice. Considering the values of the various performance parameters it can be inferred that the proposed all-p type organic inverters may prove to be a trustworthy circuit for various applications of flexible electronics. And may also find their utilization in the areas of organic memory blocks, ring oscillators, biosensors, and wearable electronic devices.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

ORCID iDs

Poornima Mittal https://orcid.org/0000-0002-9479-8628

References

[1] Kumar B, Kaushik B K and Negi Y S 2014 Static and dynamic characteristics of dual gate organic TFT based NAND and NOR circuits *J. Comput. Electron.* **13** 627–38
[2] Schroeder R, Majewski L A and Grell M 2004 Improving organic transistor performance with Schottky contacts *Appl. Phys. Lett.* **84** 1004–6
[3] Halynne R L, Zachary J C, Rosemary R C, Nicholas T B, Cory S H, Adam J S and Benoît H L 2022 Chloro aluminum phthalocyanine-based organic thin-film transistors as cannabinoid sensors: engineering the thin film response *Royal Soc. of Chem.* **1** 1165–75
[4] Mittal P, Kumar B, Negi Y S, Kaushik B K and Singh R K 2012 Channel length variation effect on performance parameters of organic field effect transistors *Microelec. Journal* **43** 985–94
[5] Gupta S, Mittal P and Juneja P 2022 Performance improvement in single-gate organic transistors with contacts at top and bottom: additional p+ region near source and drain *Main Grp Chem.* **21** 243–61
[6] Omar S, Mandal S, Ashok A, Harish A R and Katiyar M 2011 Organic inverter: Theoretical analysis using load matching technique *Microelec. Reliability* **51** 2173–8
[7] Yu Y Y and Yang C H 2020 Preparation and application of organic-inorganic nanocomposite materials in stretched organic thin film transistors *Polymers* **12** 1058
[8] Li J, Zhang X W, Zhang L, Khizar H, Jiang X Y, Zhu W Q and Zhang Z L 2009 Improving organic transistor performance through contact-area-limited doping *Solid State Commun.* **149** 1826–30
[9] Mishra N, Mittal P and Kumar B 2019 Analytical modeling for static and dynamic response of organic pseudo all-p inverter circuits *J. Comput. Electron.* **18** 1490–500
[10] Klauk H *et al* 2005 Flexible organic complementary circuits *IEEE Trans. Electron Devices* **52** 618–22
[11] Mittal P 2021 Ditch and elevated organic thin film transistor-based improved common source voltage amplifier: Frequency response characteristics and analytical modeling *J. Soc. Inf. Disp.* **29** 642–58
[12] Bansal S and Jain P 2022 Simulation of organic thin film transistor (OTFT) having high carrier mobility *ICACITE IEEE*
[13] Rezaee A and Carrabina J 2023 Dual-gate organic thin-film transistor and multiplexer chips for the next generation of flexible EG-ISFET sensor chips *Sensors* **23** 6577
[14] Mittal P, Negi Y S and Singh R K 2015 An analytical approach for parameter extraction in linear and saturation regions of top and bottom contact organic transistors *J. Comput. Electron.* **14** 828–43
[15] Zschieschang U, Waizmann U, Weis J, Borchert J W and Klauk H 2022 Nanoscale flexible organic thin-film transistors *Sci. Adv.* **8** eabm9845
[16] Kumar B, Kaushik B K and Negi Y S 2013 Static and dynamic analysis of organic and hybrid inverter circuits *J. Comput. Electron.* **12** 765–74
[17] Koch N, Ghijsen J, Elschner A, Johnson R L, Pireaux J J, Schwartz J and Kahn A 2003 Conjugated organic molecules on metal versus polymer electrodes: demonstration of a key energy level alignment mechanism *Appl. Phys. Lett.* **82** 70–2
[18] Han C Y, Tang W M and Lai P T 2017 High-mobility pentacene organic thin film transistor with LaxNb(1-x)Oy gate dielectric fabricated on vacuum tape *IEEE Trans on Electronic Devices* **64** 1716–22
[19] Peng Z, Liu F, Sun H, Wang Y, Wang J and Jiang C 2021 Complementary-like inverter based on organic-inorganic heterojunction ambipolar transistors on flexible substrate *J. of Elec. Dev. Society* **9** 933–8
[20] Tokito S 2017 Flexible and printed OTFT devices for emerging electronic applications *IEEE Elec. Dev. Tech. and Manu. Conf.* 1–3
[21] Wang J, Wei B and Zhang J 2008 Fabricating an organic complementary inverter by integrating two transistors on a single substrate *Semicond. Sci. Technol.* **23** 055003
[22] Zabihipour M, Tu D, Strandberg J, Berggren M, Engquist I and Ersman P A 2021 Designing inverters based on screen printed organic electrochemical transistors targeting low-voltage and high-frequency operation *Adv. Mat.* **6** 2100555
[23] Rashid R B, Du W, Griggs S, Maria I P, Mcculloch I and Rivnay J 2021 Ambipolar inverters based on cofacial vertical organic electrochemical transistor pairs for biosignal amplification *Sc. Adv.* **7** eabh1055
[24] Raval H N, Tiwari S P, Navan, Mhaisalkar S G and Rao V R 2009 Solution-processed bootstrapped organic inverters based on P3HT with a high-k gate dielectric material *IEEE Electron Device Lett.* **30** 484–6
[25] Mittal P, Yadav S and Negi S 2021 Advancements for organic thin film transistors: structures, materials, performance parameters, influencing factors, models, fabrication, reliability and applications 133 *Material science in semiconductor processing* **133** 105975–6000
[26] Dodabalapur A, Laquindanum J, Katz H E and Bao Z 1996 Complementary circuits with organic transistors *J. Appl. Phys.* **69** 4227–9

[27] Seifaei M, Schillinger D, Prabhu P M, Kharkeshi M M, Zschieschang U, Cuignet X, Klauk H and Manoli Y 2023 Source-degenerated schmitt-trigger-based high-performance digital logic for flexible organic thin-film transistor technologies *TechRxiv* **preprint** 1–9
[28] Vusser S D, Geneo J and Heremans P 2006 Influence of transistor parameters on the noise margin of organic digital circuits *III Trans. Electron Devices* **53** 601–10
[29] Kneppe D, Talnack F, Boroujeni B K, Teixeira da Rocha C, Hoppner M, Than A, Mannsfeld S C B, Ellinger F, Leo K and Kleemann H 2021 Solution-processed pseudo-vertical organic transistors based on TIPS-pentacene *Materialstoday Energy* **21** 100697
[30] Ravipati D P, Kedia R, Santen V M V, Henkel J, Panda P R and Amrouch H 2022 FN-CACTI: advanced CACTI for FinFET and NC-FinFET technologies *IEEE Trans. Very Large Scale Integr. VLSI Syst.* **30** 339–52
[31] Verma S and Tripathi S L 2022 Effect of mole fraction and fin material on performance parameter of 14 nm heterojunction $Si_{1-x}Ge_x$ FinFET and application as an inverter *Silicon* **14** 8793–804
[32] Mo F, Spano C E, Ardesi Y, Roch M R, Piccinini G and Vacca M 2023 NS-GAAFET compact modeling: technological challenges in Sub-3-nm circuit performance *Electronics* **12** 1487
[33] Negi S, Baliga A K, Pandey Y, Mittal P and Kumar B 2016 Performance analysis of dual gate organic thin film transistor through analytical modeling *International Conference on Computing Communication and Automation (ICCCA- 2016)* 1533–8
[34] Myny K 2018 The development of flexible integrated circuits based in thin-film transistors *Nature Elect.* **1** 30–9
[35] Yao Z, Chang T J, Zentzlaff D and Rand B 2021 Benchmarking organic thin film transistor inverter design styles *Synth. Met.* **278** 116825
[36] Kim H, Jeong D Y, Lee S and Jang J 2019 A high-gain inverter with low-temperature poly-Si oxide thin-film transistors *IEEE Elect. Dev. Letters* **40** 411–4
[37] Ishikawa Y, Wada Y and Toyabe T 2010 Origins of characteristics differences BGTC and BGBC devices *J. Appl. Phys.* **107** 053709
[38] Gupta S, Mittal P and Juneja P 2021 Different materials, structures and parameters for organic thin film transistors *J. Phys. Conf. Ser.* **1714** 012017

This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2023.3325317

TCE-2023-04-0351

DREAM: Deep Learning-based Recognition of Emotions from Multiple Affective Modalities using consumer-grade body sensors and video cameras

Aditi Sharma, *Member, IEEE*, Akshi Kumar, *Senior Member, IEEE*

***Abstract*— Ambient smart cities exist on the intersection of digital technology, disruptive innovation and urban environments that now essentially augment affect empathy and intelligent interfacing for human computer interactions (HCI). This research puts forward a deep learning approach, DREAM, for recognition of emotions using three affective modalities (audio, video, physiological) to develop an empathetic HCI system using consumer electronic IoT sensors and cameras. Convolution network is used to train for physiological signals. VGG and ResNet have been used to pre-train the models for emotion recognition from video and audio signals. DREAM is then fine-tuned on the publicly available K-EmoCon dataset to accurately recognize emotion for each subject. K-EmoCon is annotated by seven persons for five discrete emotions, and two affect dimensions. Finally, a probability-based average decision-level fusion strategy is used for combining the outputs of all the modalities. Leave one out strategy is used to train and evaluate the model for subject specific accuracies. For discrete emotions highest accuracy of 81.7% and 82.4% is achieved for dimensional emotions. DREAM has performed better than existing state-of-the-art for both emotion models.**

***Index Terms*— Facial Emotion Recognition, Human Computer Interaction, IoT sensors, K-EmoCon, Multi Modalities, Transfer Learning.**

I. INTRODUCTION

AMBIENT Intelligence (AmI) enriches the urban spaces as it combines end-point hardware, software, user experience, machines/human machine interaction and learning. Smart technology, data, sensors, video camera, apps, and citizen science, all have the potential to influence and shape urban environment in an ambient smart city. Built upon the advances in sensors and sensor networks, cameras, pervasive computing, and artificial intelligence, an ambient smart city essentially includes systems characterized as sensitive, responsive, adaptive, transparent, ubiquitous, and intelligent [1,2]. But a city is not a machine and is rather made by citizen's actions and feelings. Augmented intelligence [3] (also known as intelligence amplification, or IA), a subsection of artificial intelligence (AI) has emerged as a symbiotic relationship between man and machine to enhance human intelligence within these ambient smart cities. IA has been part of the human-centered partnership model of people and AI for years. As the IoT and smart objects(camera) connectivity expands, we can expect to see IA in almost all aspects of life to enhance cognitive performance, including learning, decision making and new experiences that can lead the way to develop safe and reliable work place for humans in industry 5.0[4].

Emotions are an inseparable aspect of human intelligence and integral to decision making. Researchers are working in the field of human computer interaction (HCI) for more than a decade now, and emotion recognition and understanding of human senses (like taste and smell) remain critically important for HCI [7]. Chatbots and AI-powered conversational interfaces are reshaping the world of HCI. Moreover, most ambient-assisted applications in smart cities and industries still lack the understanding of human sentiments and emotions [5].

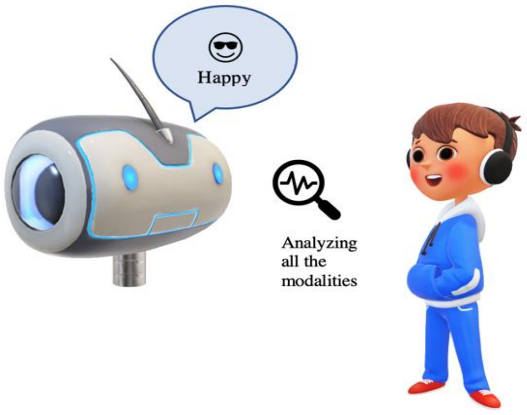


Fig.1. Human Computer Interaction

An obvious solution is to integrate the intelligent amplification of things (IAoT) with emotion sensing which can have a variety of applications such as, active, and assisted living, health care and industry. It aims to present an opportunity to developers for creating automated systems which can interrelate with the affective and cognitive state of humans to make better decisions [6]. For example, an IAoT-enabled bot can layer on adaptations using the user's state of

Corresponding author: Akshi Kumar.

Aditi Sharma was Research Scholar at the Delhi Technological University, New Delhi, India. She is now an Assistant Professor with Thapar Institute of Engineering and Technology, Patiala, India. (e-mail: aditi.sharma@thapar.edu).

Akshi Kumar is with the Department of Computing, Goldsmiths, University of London, United Kingdom. (e-mail: akshi.kumar@gold.ac.uk).

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:09:11 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2023.3325317

mind and have empathy to enact like humans to operate in any client-oriented business for example, smart hotel or hospital management system or call center and business process outsourcing (BPO) service provider [8], [9] using consumer grade cameras and sensors. Due to easy availability of these cameras and sensors in current era, it has the potential to revolutionize the way we understand and respond to human emotions, particularly in the field of consumer behavior. With the use of these sensors and cameras, it is possible to gather data on the emotional states of individuals in real-time, which can then be used to inform marketing strategies, product design, and customer service [10].

The role of empathy has come to prominence in HCI as the community increasingly engages with issues in medical, health and emotionally charged contexts [11]. As the natural human-human interaction (HHI) is fundamentally multimodal, this fosters the need to exploit multiple affective modalities to build an empathetic HCI (Fig.1).

Pertinent literature shows the use of a variety of modalities ranging from facial expressions, body gestures, speech, and physiological signals [12], [13], [14] for sensing affective states. Recent research also advocates that a reliable automatic affect recognition system should attempt to combine two or more modalities [15, 16].

This research puts forward a deep learning approach for recognition of emotions using three affective modalities. The approach can be integrated to both AI-powered conversational chatbots and sensors, steering to an emotion-aware ambient smart city. One of the key benefits of using consumer-grade sensors and cameras for the automatic recognition of emotions is that they are relatively inexpensive and easy to use. This means that businesses of all sizes can take advantage of these technologies to improve their understanding of consumer behavior. For example, a small business owner could use facial recognition software to track customer reactions to different products or promotions, while a larger corporation could use body sensors to gather data on how customers respond to different types of advertising. The proposed approach, DREAM is built using decision-level fusion of three different affective modalities, namely, audio (speech), video (face expressions) and physiological signals. Transfer learning is used to learn from existing data. VGG and ResNet have been used to pre-train the model, followed by fine tuning using CNN for audio signals. Video signals are processed through ResNet twice for facial emotion detection and is fine-tuned on K-EmoCon using convolution network. Physiological signals are processed with help of convolution models 1-D and 2-D. The output of all the modalities is combined using probability-based average decision fusion, it is one of the techniques of late fusion [15]. The publicly available dataset for emotion recognition in conversations, K-EmoCon is used to train and test the approach.

K-EmoCon is different from other available datasets in terms of annotation of the dataset. Most available datasets for emotion recognition are either self-annotated by the subject, or by an external observer like in case of IEMOCAP. But in K-EmoCon, each instance has been annotated by 7 people, one the person himself, second the opponent/partner in the debate, and lastly by 5 external observers. And all these seven annotations are taken both in terms of discrete emotion states, and the dimensional states, the two types of emotion models [16]. Processing the different output classes available is a complex task, most of the work carried out by researchers either use self-annotation as the output label, stating a person knows best but he or she is feeling [17]. Others have used annotation by the opponent, for interaction problems, how the other person is observing, should be observed by the robot for HCI [18]. But, to include every perspective, we have processed all seven emotion annotations in both discrete form as well as dimensional form for accurate emotion detection from all perspectives.

For understanding of human emotions, different researchers have used different types of data, like IEMOCAP, MELD, CASE, CK+ for facial expressions; Berlin, LEESD for emotion from Speech. Different Modalities has been explored by researchers for emotion recognition; some have focused on facial expressions, others on hand gestures, some have worked on emotion recognition from speech, while others have focused on physiological changes. Combination of two or more modalities has been employed by researchers on some datasets [15] [16]. Saha et al used kinetic sensors for recognizing hand gestures for emotion recognition into 5 emotion states using K-NN, SVM, NN [12]. Wei et al. used Attention based CNN for emotion identification from gestures in the video signals [13].

Bertero et al. in 2017, used TED talks for emotion identification from Speech signals using CNN [14]. Lalitha et al. used different perceptual features like MFCC, PLPC with deep neural network for speech emotion recognition from Berlin dataset [19]. Fan et al. in 2021 created pre-tuned models for emotion recognition from speech signals for HCI on LSSED dataset, the largest dataset available for speech emotion recognition [20]. Kumar et al. used experimented on IEMOCAP, MELD for emotion recognition using facial expressions [15]. Chowdhary et al. in 2021 used ConvNet on pre-trained models for identifying emotions from facial expressions in CK+ dataset [21]. Ahmed et al. and Zhang et al. have worked on emotion recognition from speech signals, by analysing the linguistic parameters, and the melogram pattern of the speech signals [32,33].

K-EmoCon dataset was published in 2020, only few researchers have worked on K-EmoCon till now, none of whom used all the modalities of the dataset for emotion recognition. Gupta et al. [22], Zitouni et al. [23], Alskafi. Et al. [24], and Dissanayake et al. [25] have used only physiological signals for identifying valance-arousal emotion state of the subject. Quan et al. used only audio and visual signals of 16 participants for emotion recognition [18]. Yang et al. has used CNN on K-EmoCon for identifying emotion recognition using physiological signals only, not the face expressions or linguistic patterns at all [33]. Alhussein used CNN on only linguistic and speech signals of K-EmoCon to identify the emotions in arousal and valance dimensional emotions, having state of the art f1-score of 82% [34].

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:09:11 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2023.3325317

The primary contribution of the proposed DREAM is:

- Three affective modalities (audio, video, and physiological) have been used for emotion recognition.
- Different deep neural networks trained for each modality of K-EmoCon.
- For accurate emotion identification, three annotations, namely- self, partner, & externals are used.
- The performance is evaluated for both emotion models; discrete emotion states and affect dimensions.

Next section of this paper contains description of the dataset. Section 3 contains methodology used, followed by discussion of result and analysis.

II. Dataset

K-EmoCon: A multimodal publicly available dataset for emotion recognition in conversations provided by Park et al. in 2020 [26]. The dataset contains audio, video and bio signals of 32 subjects, who participated in a debate task on a social issue in teams of 2, while wearing Physiological signal measuring devices, Emperica E4, NeuroSky, and Polar H7, along with Video cameras for recording the facial expressions and gestures of the participants. On average, 10 minutes of debate was conducted between each pair, for emotion recognition, audio and video was recorded of this debate, along with Bio-signals from 3 wearable devices. ACC (32Hz), BVP (64 Hz), EDA (4Hz), HR(1Hz), IBI, and Temp (4Hz) was measured through Emperica E4; Attention, Brainwave – EEG (125Hz) and Meditation through NeuroSky; and ECG (1 Hz) from Polar H7.

K-EmoCon is the first publicly available dataset that contains natural conversation between non-actors, while monitoring their physiological changes. Another unique feature of K-EmoCon is that each instance has been annotated by 7 people, one the person himself, other by the opponent/partner in the debate, and lastly by 5 external observers. Although it makes processing the dataset difficult, but it provides different perspective for accurate emotion recognition. Park et al. have used 20 different types of emotions for annotation, including dimensional emotional model (Arousal and Valance), Basic Emotions (Cheerful, Happy, Angry, Nervous, and Sad), along with both common and less common BROMP (Baker Rodrigo Ocumpaugh Monitoring Protocol) Affective categories [26].

A. *Data Selection*

K-EmoCon contains multiple files for each modality, not all the files are accessible or are having missing entities, data cleansing was performed. The audio recording of the dataset contains 16 files, each for the debate between the participants, so each file contains audio of 2 participants. Video Signals were recorded through separate cameras for each participant, but 11 of the participants didn't allow the data for public availability, we have only access of 21 participants visual signals. Bio-signal measuring devices too had some troubles. For our proposed model, we have taken audio, video, BVP, EDA, HR, IBI, Temp, and EEG signals for accurate Emotion recognition. As only 18 participant's complete data were available for all these signals, the empirical analysis has been conducted on 18 participants data only.

B. *Pre-processing*

For emotion recognition, classification of emotions has been taken into 7 types, 2 affective dimensions and 5 emotion states as shown in fig. 2. Emotion annotation of these 7 categories were ranked on scale 1-5 for dimensional, and 1-4 for emotions by all seven annotators.

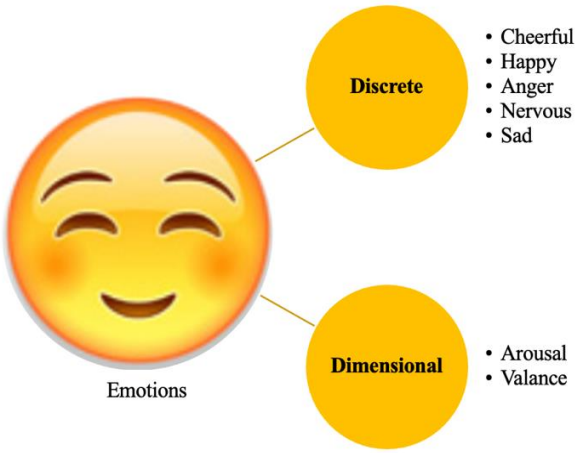


Fig. 2. Emotional Models

For our analysis we divided each emotion and dimension into two parts, low arousal if average score of all 7 annotators between 1-3, and high arousal if between 3-5, similarly very cheerful for average score of 2-4, and less cheerful for 1-2.

The experiment was evaluated for both Affective dimensions, and Emotion states separately, with 4 and 10 output classes respectively. To create time synchronous data for proposed work, the 18 participants' all signals were mapped with output (Dimensions, and Emotions), so each signal is annotated at 5 second interval with one of the four affective dimensions for first experiment, and similarly with one of the ten emotion state for second experiment.

All input files of every modality are mapped with the averaged annotated emotion. Speech signals of 22Khz sampling is cut at interval of 5 seconds and mapped to corresponding annotated emotion state. Similarly, visual signals sampled with 30 fps with frames resized into 112*112 pixels is mapped to 5 second annotated emotion by mapping 150 (30*5 frames) to single emotion instance. The bio-signals after pre-processing, is mapped to output emotion class, by using statistical feature of mean for taking 5 second data as an instance.

The bio-signals and visual signals are recorded for each participant separately, but the audio signals are recorded in sharing for the team of 2 participants. The audio signals of the opponent can trigger certain emotions in the subject, but it cannot reflect the emotions of the subject, so while pre-processing, we manually generate audio file for each participant from original file. First, a duplicate copy of the file was created for the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:09:11 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2023.3325317

opponent, and then manually the voice of the opponent was identified and removed from the file (sections) and replaced with the voice signals of the participant at start of the debate. Similar process was followed for opponent copy. For bio-signals, in pre-processing, missing values were replaced by means of the next two consecutive instance values.

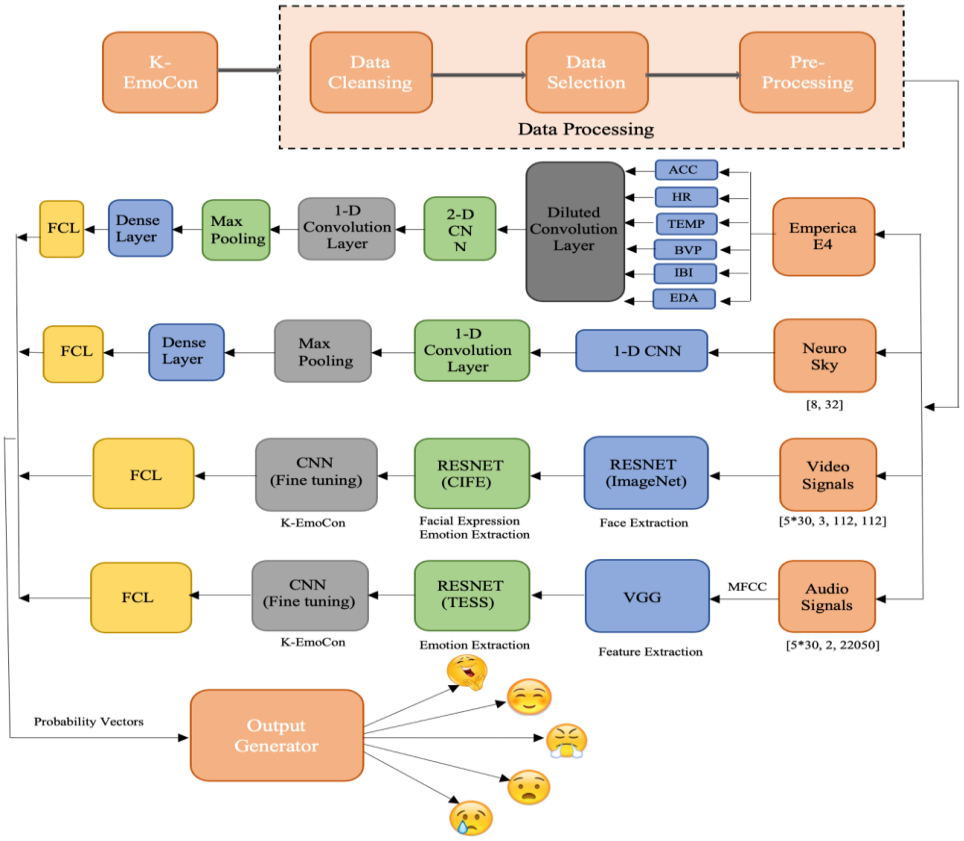


Fig.3. Architecture of DREAM

III. DREAM: Proposed Model

To create a human computer interaction system using consumer grade camera and sensors, we proposed a model with 3 modalities- DREAM. The architecture of DREAM is shown in fig 3. Initially nine input files from K-EmoCon dataset were provided to proposed model after pre-processing. Six out of nine files were of Emperica E4, Only EEG file taken from NeuroSky, Polar_H7's data had many missing values, and Heart rate has already been received from E4, so no file from Polar_H7 was processed. Video file of each participant, and Separate audio file generated after pre-processing for the participant was used as the initial input to the model.

A. Transfer Learning

The use of deep learning models has advanced many application areas as they provide state-of-the-art results. The models are especially useful for feature extraction owing to the ability to learn representations that cannot be modelled manually. But these models require large training datasets to perform well in real-time and building a model from scratch is time consuming and expensive in terms of data collection/labelling, privacy and training time. Transfer learning has emerged as a solution for developing and training deep neural models with less data and compute power. Transfer learning is about "transferring" the learnt representations to another problem. With transfer learning, we basically try to exploit what has been learned in one task to improve generalization in another.

For the task of emotion recognition, we follow an inductive transfer learning approach. A general representation is pre-trained on a large unlabelled dataset and is then adapted to a supervised target task using the labelled data available. The pre-trained model can also be made available for use to others who can fine-tune it for their target downstream tasks. Thus, while pre-training on a large dataset is computationally expensive, it only needs to be done once. Compared to it, downstream fine-tuning is much cheaper.

B. Audio Modality

After generating separate audio files for each participant, the speech signals of sampling rate 22KHz, was provided as an input ([T, C, F]), where T represents time i.e., 5s, C represents Channel i.e., 2, F represents frequency 22050Hz. Mel-

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:09:11 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2023.3325317

Frequency Cepstral Coefficients (MFCC) waveform encoding was used to extract audio features. For mapping audio to visual signals at same time instance, we reframed the audio signal size to that of visual signals, so temporal dimension after MFCC of audio signal became 150 (30*5). These extracted features are then provided as an input to VGG a pre-trained model [8] followed by ResNet [27] for pre-training the model for emotion recognition on TESS (Toronto Emotional Speech set) dataset. VGG creates a simple but unified deepen structure of the network, ResNet uses residual learning, that ease the training process [28]. These pre-trained models are then fine-tuned by using K-EmoCon for mapping the output to 10 emotion states. The output of fine-tuned layer is then passed to a fully connected layer (FCL).

C. Video Modality

After mapping emotional state at each segment of 5 second interval, the size of individual video clip became [5 x 30, 3, 112, 112] represented as ([T x F, C, W, H]), where T is time, F for frames per second, C for color channels, W and H stands for width and height respectively.

This input feature vector is feed to pre-trained ResNet on ImageNet for face extraction. The output of this ResNet model is then given to second pre-trained ResNet model on CIFE for emotion mapping of facial expressions. These pre-trained models have different with output classes, so it is then fine-tuned on the dataset itself. These visual signals can be used to identify the body gestures such as hand movements, but as the movement of participants is limited (seated on a chair), we just took facial expressions into consideration for emotion classification. Finally, the output from fine-tuned model is feed into a fully connected layer to predict the emotion state.

D. Physiological Modality

Seven Physiological signals have been measured by 2 wearable sensor devices Emperica E4, and NeuroSky. The band pass filtering was applied on EEG signals (8 bathes, 32 Hz), with a window size of 5 seconds to map the emotion state [29]. The input is then provided to a convolution layer with kernel of size 3 * 1, with 64 kernel and the pooling size 3*3. The output of pooling layer is again provided to 2D convolution layer, followed by max pooling.

The output of this pooling layer is provided to dense layer followed by a fully connected layer. For 6 bio signals of Emperica E4 wrist band, as the size of each signal is different, like ACC is 3*126 (for each participant), and HR 1*126, initially a diluted convolution layer is applied to all 6 bio-signals, then after generating the same size feature map, it is provided as an input to two 1D convolution layers similar to EEG signal, followed by Dense layer, and fully connected layer to produce emotion state.

E. Convolution Neural Network

To project the input sequences into a semantic space with a unified dimension, a 1D convolution layer or 2D convolution layer are applied on feature dimensions with varying kernel size as described in previous subsections. In the CNN layers and fully connected layers, the activation function of the Rectified Linear Unit (ReLU) is set, which handles a threshold of 0 for the negative values. This $ReLU(x)$ function is calculated as equation (1):

$$f(x) = \max(0, x_i) \tag{1}$$

The max-pooling layers are alternated between the CNN layers because of convolutional region that can increase the robustness of the features and reduce the dimensionality of the different signals vector. As a regularization technique to decrease the overfitting in the layers of the neural network, the dropout with a value of 0.5 is added. The output layers of the fully connected network are configured with the SoftMax classifier, with the purpose that the hidden layers verify the probability of predicting the emotion [35]. During the supervised training, the loss is minimized with the Root Mean Square Propagation (RMSProp) optimizer, since it adjusts the learning rate adaptively. Initially, the learning rate is set to 0.001. Once the model is executed, the knowledge base is consolidated between the vector of physiological features and the class vector. Then, to evaluate the emotion recognition, in the fully connected layer the cross-entropy loss function is used.

Each 1D CNN contains a sequence of temporal data for the recognition of local patterns, which can be learned from the physiological, video, and audio signals morphology. The functionality of the CNN layers is given by the convolution kernel that obtains the local patches and the Max pooling extracts the windows from the feature vectors to generate the down sampling output vector.

F. Average Decision Fusion

The output of the 4 fully connected layers of the three modalities are given as an input to an output generator with weighted decision fusion. The final output generator provided output by taking in emotion state produced by 4 fully connected layers and providing output by using average weighted decision fusion model. In this method, we take output of all 4 sub-models produced by the SoftMax function resulting in probabilities for all 10 emotion states. Suppose the 4 sub-models are represented by E_V, E_A, E_{EEG}, E_{E4}, Where E_V stands for output vector of Emotions produced by Visual modality, E_A stands for output vector of Emotion of Audio modality, E_{EEG} for output vector of EEG signal, and lastly E_{E4} gives the output vector produced by 6 bio-signals of Emperica E4 device.

Each output vector will contain probabilities of all 10 classes produced for that instance, say P_{v1}, P_{v2}, P_{v3}, P_{v4}, P_{v5}, P_{v6}, P_{v7}, P_{v8}, P_{v9}, P_{v10}, where P_{v1} represents, the probability of the output to be less cheerful and P_{v6} represents probability of belonging to very cheerful for visual modality model. This Probability value is generated by SoftMax activation function displayed in equation (2), employed after fully connected layers in each sub-model.

$$SoftMax\ (P_{Vi}) = \frac{\exp(P_{Vi})}{\sum_i exp(P_{Vi})} \tag{2}$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:09:11 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2023.3325317

Then in final output generator, probability of corresponding class will be added with probabilities of same class output from different sub-model as shown in equation (3).

$$P_i = P_{Vi} + P_{Ai} + P_{EEGi} + P_{E4i} \tag{3}$$

After calculating probability of each output class, the counter classes' probabilities are compared with each other, the one with more value will be provided as output, resulting in 5 outputs for a single instance.

$$E = max(P_i, P_{n-i}), i \in (1, \ldots, n) \tag{4}$$

From equation (4), it can be observed that for "n" output classes, there will be n/2 outputs.

For all the fully connected layers in four sub-models, Adam optimizer has been used, and loss is calculated by categorical cross-entropy as shown in equation (5).

$$Loss = -\sum (y'_{i1} log(y_{i1}) + y'_{i2} log(y_{i2}) + \cdots + y'_{in} log(y_{in})) \tag{5}$$

y_{i1}, y_{i2}, y_{in} are internal node labels, y'_{i1} , y'_{i2} , y'_{in} are the output layer nodes, produced by SoftMax function.

IV. Result & Discussion

The proposed model was executed twice once for output categorized as Dimensional Affect (low valance, high arousal, low arousal, high valance), and second execution for ten emotion states. For both executions, the proposed model works exactly same. The model was evaluated on AWS, Amazon EC2 G4dn instance having T4 GPU, with 100Gbps networking. The dimensional model presupposes that emotions are not independent and that there exists a relation between them hence the need to place them in a spatial space, whereas discrete emotion models categorize the emotions based upon the feelings of the user. Both have their own importance, to have an empathetic HCI, root should understand the actual discrete emotion, but at the same time should be able to understand the impact of the emotion using dimensional state. With rigorous experimentation, tuned hyper-parameter values were identified for most efficient results shown in table 1.

For performance evaluation the training and testing of the model is performed using "leave one out". The size of the dataset is limited, having all the signals recorded only for 10 minutes for each of the 32 participants, generating total of 320 minutes' data approximately. But only taken 18 participants data was taken for model execution, resulting in almost half the total data. Therefore, to have large training set, leave one out strategy was employed, implying that first model will be trained on 17 participants complete data, will be tested for performance evaluation of 18th participant. Then the same process will be repeated leaving another participant data for testing. So, the model will be executed 18 times for both types of output categories.

TABLE 1
HYPER-PARAMETERS

	LEARNING RATE	EPOCHS	BATCH SIZE	DROPOUT	ACTIVATION	OUTPUT
VISUAL	0.001	60	32	0.5	ReLU	Soft Max
AUDIO		56	48	0.5		
EEG		120	32	-		
E4		68	16	-		

The execution of the model has been performed on AWS's Amazon EC2 instance through MacBook Pro. To evaluate the performance of proposed multimodal, F1-score has also been computed along with Accuracy.

TABLE 2
MODEL PERFORMANCE FOR EMOTION STATES

	P_1	P_4	P_8	P_9	P_{10}	P_{11}	P_{15}	P_{18}	P_{19}	P_{20}	P_{21}	P_{22}	P_{23}	P_{24}	P_{25}	P_{26}	P_{29}	P_{30}
F1-SCORE																		
LESS CHEERFUL	57.9	71.2	69.8	61.7	64.3	66.2	68.7	70.1	65.2	67.2	**72.3**	59.4	65.4	68.2	67.9	63.6	61.8	68.1
VERY CHEERFUL	70.5	69.3	72.3	**79.7**	76.1	73.5	68.5	69.3	70.4	61.9	78.4	73.7	62.7	69.5	69.8	66.3	70.3	75.8
LESS HAPPY	73.8	67.3	69.4	78.4	61.1	54.9	66.9	64.5	77.9	72.5	83.6	44.7	70.2	65.8	73.5	71.9	74.9	**79.1**
VERY HAPPY	81.9	74.9	78.5	80.3	72.8	75.1	**83.4**	73.9	71.8	72.9	51.4	78.2	68.9	72.3	70.3	68.4	63.7	75.3
LESS ANGRY	68.7	64.5	70.3	68.4	72.7	69.3	71.6	74.8	75.4	65.2	80.1	75.4	67.3	**82.5**	67.1	71.9	66.9	80.7
VERY ANGRY	75.4	58.4	68.4	67.6	78.3	72.6	**87.8**	51.9	70.3	81.9	71.8	71.7	76.4	72.1	74.9	72.3	71.5	57.9
LESS NERVOUS	56.2	65.8	71.2	75.4	74.1	79.4	75.5	66.3	73.9	72.6	79.5	72.8	79.8	77.9	64.8	80.1	78.4	**82.9**
VERY NERVOUS	69.2	72.5	75.3	76.1	63.8	68.4	**82.7**	78.2	67.3	69.8	64.9	81.6	54.9	76.8	75.8	73.6	40.6	79.4
LESS SAD	79.3	68.5	72.4	73.4	80.4	81.9	69.5	68.7	64.5	55.8	81.6	68.2	**82.3**	79.5	69.4	78.9	63.5	69.9
VERY SAD	74.6	**83.6**	76.7	72.5	75.8	78.4	78.2	66.1	62.6	79.3	76.7	79.6	74.8	73.1	77.2	82.4	79.9	73.6
ACCURACY																		
LESS CHEERFUL	68.4	63.5	75.3	71.2	69.9	68.1	64.5	74.7	69.9	64.3	70.4	66.7	71.7	73.2	71.9	66.8	**78.9**	68.7

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:09:11 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2023.3325317

VERY CHEERFUL	66.1	68.7	51.9	68.4	59.8	70.2	62.6	**76.1**	73.5	68.5	68.3	49.8	68.7	71.8	72.3	72.6	74.1	57.6
LESS HAPPY	70.8	71.8	66.3	70.2	75.3	65.7	68.4	68.1	72.7	70.8	66.1	72.5	75.1	63.8	64.5	67.6	**78.3**	72.6
VERY HAPPY	69.5	75.4	72.6	69.3	68.7	71.8	67.6	**78.3**	72.6	58.7	63.7	59.8	69.3	61.7	62.6	74.8	66.2	71.9
LESS ANGRY	64.5	75.1	68.5	70.3	76.1	69.3	72.3	74.7	75.1	62.5	66.9	69.3	**80.2**	66.2	63.6	61.8	68.9	72.3
VERY ANGRY	68.9	69.3	71.9	73.9	74.9	72.4	63.9	70.3	**79.4**	67.4	71.5	75.3	76.1	67.8	74.7	65.4	54.9	71.7
LESS NERVOUS	67.3	72.6	64.8	67.3	64.5	70.6	70.4	74.9	46.7	69.5	**81.7**	72.4	73.4	51.7	70	62.7	72.3	64.5
VERY NERVOUS	72.1	64.5	70.6	72.6	68.9	68.4	62.1	73.9	72.5	68.3	45.8	**76.9**	71.8	72.9	68.4	70.2	69.4	62.6
LESS SAD	**73.2**	62.6	68.4	68.1	67.3	66.9	54.7	67.3	59.4	65.4	68.2	76.1	73.5	68.5	44.3	68.9	63.6	61.8
VERY SAD	71.6	48.7	69.2	63.5	76.4	62.3	66.3	70.2	68.7	71.2	**73.8**	63.6	61.8	69.3	72.6	67.3	71.8	72.9

TABLE 3
MODEL PERFORMANCE FOR DIMENSIONAL AFFECTS

	P_1	**P_4**	**P_8**	**P_9**	**P_{10}**	**P_{11}**	**P_{15}**	**P_{18}**	**P_{19}**	**P_{20}**	**P_{21}**	**P_{22}**	**P_{23}**	**P_{24}**	**P_{25}**	**P_{26}**	**P_{29}**	**P_{30}**
								F1-SCORE										
LOW VALANCE	**89.6**	87.6	83.6	83.1	80.1	79.8	68.7	73.9	71.8	72.9	68.4	79.2	69.4	68.2	77.2	82.4	79.9	73.6
HIGH VALANCE	84.3	79.7	78.9	91.7	76.1	81.3	68.5	69.3	70.4	61.9	78.4	73.7	62.7	71.5	75.3	76.1	**92.6**	83.7
LOW AROUSAL	79.8	**91.7**	81.3	81.4	84.2	83.5	66.9	64.5	77.9	80.5	83.6	81.4	70.2	65.8	73.5	71.9	74.9	84.1
HIGH AROUSAL	83.2	78.9	78.5	80.3	**90.7**	86.7	83.4	73.9	71.8	72.9	80.6	78.2	68.9	72.3	70.3	71.4	81.5	75.3
								ACCURACY										
LOW VALANCE	75.4	58.4	68.4	72.3	74.7	75.1	82	57.8	77.2	**82.4**	79.9	73.6	76.4	72.1	74.9	72.3	71.5	68.3
HIGH VALANCE	56.2	65.8	71.2	**81.4**	73.7	79.4	75.5	66.3	73.9	72.6	79.5	72.8	79.8	77.9	64.8	80.1	78.4	78.1
LOW AROUSAL	69.2	72.5	75.3	77.8	79.7	**80.2**	82.7	78.2	67.3	77.2	79.4	79.9	73.6	76.8	77.2	78.1	76.9	70.6
HIGH AROUSAL	**79.8**	69.6	76.3	70.7	76.4	72.1	74.9	72.3	76.3	77.3	74.9	72.4	74.7	65.4	72.5	75.3	76.1	76.4

The model was evaluated on both accuracy as well as F1-score, as both are good evaluation metrics for machine learning models, although accuracy is more suitable for balanced dataset, while F1-score is suitable for both balanced as well as imbalanced dataset. The K-EmoCon is 50-50 balanced, but it doesn't fall in imbalanced dataset as well. We have employed both strategies because, some researchers only use F1-score, while others use accuracy. Authors expect the DREAM model to act as a baseline for future empathic HCI emotion recognition models, so we have provided both the evaluation metrics for the researchers. The model was executed twice to identify both types of emotion Dimensional as well as discrete, for the similar reason, so every researcher working on K-EmoCon, will be able to compare their work with DREAM results. Also authors have developed model for both types of emotions recognition, because identifying an emotion type depends upon the situation, for real-time application of empathic HCI emotion recognition, emotion should be both discrete and dimensional. Dimensional emotion gives us the amplitude of the emotion, whereas discrete emotion allows us to understand the actual emotion.

A. Result

The proposed work has three modalities, each executed separately in synchronous manner, combining the output of final layers of each model into Output layer. The performance evaluation of proposed model for both output categories are shown in table 2 and table 3. For emotional states highest F1-score was obtained for very angry emotion for P_{15}, and highest accuracy of 81.7 for P_{21} participant for Less Nervous emotion. For Dimensional Affect Arousal and Valance, the highest accuracy of 82.4 and F1-score of 92.6 was achieved. DREAM model have performed better than the SOTA which was achieved earlier by Alhussein et. al. [34], where they have achieved 82% and 83% F1-score for arousal and valance emotion respectively. Alhussein has not performed the discrete emotion recognition, so we are only able to compare the model on dimensional emotional, where DREAM has outperformed the SOTA model [34].

B. Discussion

The proposed multimodal have resulted in state-of-the-art results for both emotion categories. The comparison of the model is not possible, since this is the first work that incorporate three different modalities on K-EmoCon dataset and for classifying that many different emotions. Although the model proposed by Alhussein et al. [34] have produced the earlier state of the art on

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:09:11 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2023.3325317

dimensional emotions, but our model has surpassed their model performances.

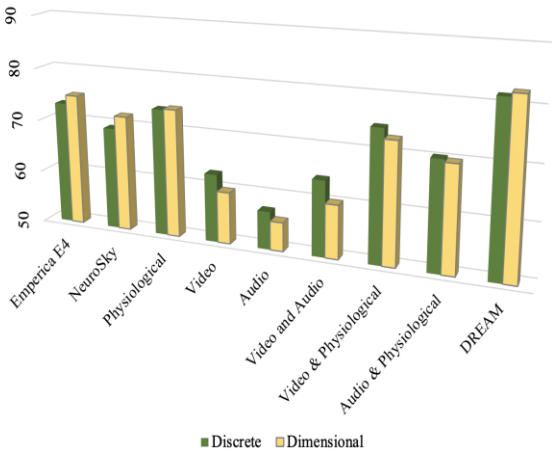


Fig.4. Ablation Study Analysis

The ablation study on the proposed model was performed as well, to analyze the impact of each modality. The Physiological modality performs best individually, followed by video modality, and the accuracy achieved by only speech signals for emotion detection was the least of all three. The best accuracy achieved by each modality in the ablation study is shown in figure 4. The variation in results in each individual modality is impacted by the amount of the data available as well, and the bio-signals were available for more participants in comparison to video or audio signals. For future work, researchers should work on making the model more secure and fast, rather than just enhancing the performance of the models.

V. Conclusion

Human computer interaction using consumer grade camera and sensors can enable the use of robots for various domains like, healthcare, hotel management, smart industries, supply chain manager, Human resources, and many more. With availability of smart wearable sensors, it is possible to track the activities of a person along with their audio, video, and bio-signals. We proposed multimodal emotion recognition using K-EmoCon dataset, containing audio, video, and physiological signals. No other work has used all the modalities for emotion recognition. The performance of the model was also tested on two types of emotion classification. In both classifications, DREAM has produced state of the art results. For improving the model further researchers can embed other modalities like linguistic by generating the transcripts of audio signals, from video signals, researchers can also use hand movements (gestures), for more accurate emotion recognition.

References

[1] D. J. Cook, J. C. Augusto, & V. R. Jakkula, “Ambient intelligence: Technologies, applications, and opportunities”. Pervasive and Mobile Computing, 5(4), 277-298, 2009.

[2] D. Ayata, Y. Yaslan, Y., & M. E. Kamasak, (2018). Emotion based music recommendation system using wearable physiological sensors. IEEE transactions on consumer electronics, 64(2), 196-203.

[3] J. Ploennigs, J. Cohn, & A. Stanford-Clark, “The future of IoT”. IEEE Internet of Things Magazine, 1(1), 28-33, 2018.

[4] A. Zanella, N. Bui, A. Castellani, L. Vangelista, & M. Zorzi, “Internet of things for smart cities”. IEEE Internet of Things journal, 1(1), 22-32, 2014.

[5] Fei Yan, Abdullah M. Iliyasu, and Kaoru Hirota. "Conceptual framework for quantum affective computing and its use in fusion of multi-robot emotions." Electronics 10.2 (2021): 100.

[6] Rui, Yong. "From artificial intelligence to augmented intelligence." IEEE MultiMedia 24.1 (2017): 4-5.

[7] M. Obrist, C. Velasco, N. Ranasinghe, A. Israr, A. Cheok, & K. P. Gopala, “Sensing the future of HCI: touch, taste, and smell user interfaces. interactions,” 23(5), 40-49, 2016.

[8] M. Chowdary, Tu Kalpana, N. Nguyen, and D. Jude Hemanth. "Deep learning-based facial emotion recognition for human–computer interaction applications." Neural Computing and Applications, 1-18, 2021.

[9] P. Gupta, S. A. Balaji, S. Jain, and R. K. Yadav. "Emotion Recognition During Social Interactions Using Peripheral Physiological Signals." In Computer Networks and Inventive Communication Technologies, pp. 99-112. Springer, Singapore, 2022.

[10] J. S. Park, J. H. Kim, & Y. H. Oh, (2009). Feature vector classification based speech emotion recognition for service robots. IEEE Transactions on Consumer Electronics, 55(3), 1590-1596.

[11] A. Guhn, L. Merkel, L. Hübner, I. Dziobek, P. Sterzer, & S. Köhler, “Understanding versus feeling the emotions of others: how persistent and recurrent depression affect empathy”. Journal of psychiatric research, 130, 120-127, 2020.

[12] S. Saha, S. Datta, A. Konar and R. Janarthanan, "A study on emotion recognition from body gestures using Kinect sensor," 2014 International Conference on Communication and Signal Processing, 2014, pp. 056-060

[13] Jie Wei, Xinyu Yang, and Yizhuo Dong. "Time-Dependent Body Gesture Representation for Video Emotion Recognition." In International Conference on Multimedia Modeling, pp. 403-416. Springer, Cham, 2021

[14] D. Bertero and P. Fung, "A first look into a Convolutional Neural Network for speech emotion detection," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 5115-5119.

[15] A. Kumar, K. Sharma, and A. Sharma, "MEmoR: A multimodal emotion recognition using affective biomarkers for smart prediction of emotional health for people analytics in smart industries." Image and Vision Computing (2022): 104483.

[16] A. Sharma, K. Sharma, and A. Kumar, "Real-time emotional health detection using fine-tuned transfer networks with multimodal fusion." Neural Computing and Applications(2022): 1-14.

[17] P. S. Gupta, A. Balaji, S. Jain, and R. K. Yadav, "Emotion Recognition During Social Interactions Using Peripheral Physiological Signals." In Computer Networks and Inventive Communication Technologies, pp. 99-112. Springer, Singapore, 2022.

[18] J. Quan, Y. Miyake, and T. Nozawa, "Incorporating Interpersonal Synchronization Features for Automatic Emotion Recognition from Visual and Audio Data during Communication." Sensors 21, no. 16 (2021): 5317.

[19] S. Lalitha, S. Tripathi, & D. Gupta, “Enhanced speech emotion detection using deep neural networks”. International Journal of Speech Technology 22, 497–510 (2019).

[20] W. Fan, X. Xu, X. Xing, W. Chen, and D. Huang, "LSSED: a large-scale dataset and benchmark for speech emotion recognition." In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 641-645. IEEE, 2021.

[21] M. Chowdary, Tu Kalpana, N. Nguyen, and D. Jude Hemanth, "Deep learning-based facial emotion recognition for human–computer interaction applications." Neural Computing and Applications (2021): 1-18.

[22] P. Gupta, S. A. Balaji, S. Jain, and R. K. Yadav. "Emotion Recognition During Social Interactions Using Peripheral Physiological Signals." In Computer Networks and Inventive Communication Technologies, pp. 99-112. Springer, Singapore, 2022.

[23] M. Zitouni, Cheul Sami, Young Park, Lee Uichin, Leontios Hadjileontiadis, and Ahsan Khandoker. "Arousal-Valence Classification from Peripheral Physiological Signals Using Long Short-Term Memory Networks." In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 686-689. IEEE, 2021.

[24] F. A. Alskafi, A. H. Khandoker, and H. F. Jelinek. "A Comparative Study of Arousal and Valence Dimensional Variations for Emotion Recognition

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:09:11 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2023.3325317

Using Peripheral Physiological Signals Acquired from Wearable Sensors." In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 1104-1107. IEEE, 2021.

[25] V. Dissanayake, S. Seneviratne, R. Rana, E. Wen, T. Kaluarachchi, and S. Nanayakkara. "SigRep: Toward Robust Wearable Emotion Recognition With Contrastive Representation Learning." IEEE Access 10 (2022): 18105-18120.

[26] C. Y. Park, N. Cha, S. Kang, A. Kim, A. H. Khandoker, L. Hadjileontiadis, A. Oh, Y. Jeong, and U. Lee, "K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations." Scientific Data 7.1 (2020): 1-16.

[27] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.

[28] S. Kriman, S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, and Y. Zhang. "Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions." In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6124-6128. IEEE, 2020.

[29] A. Kumar, K. Sharma, and A. Sharma, "Hierarchical deep neural network for mental stress state detection using IoT based biomarkers." Pattern Recognition Letters 145 (2021): 81-87.

[30] M. Z. Uddin, J. J. Lee, and T-S. Kim. "An enhanced independent component-based human facial expression recognition from video." *IEEE Transactions on Consumer Electronics* 55.4 (2009): 2216-2224.

[31] U. Ahmed, J. C. W. Lin, and G. Srivastava. "Emotional intelligence attention unsupervised learning using lexicon analysis for irony based advertising." *ACM Transactions on Asian and Low-Resource Language Information Processing* (2022).

[32] Y. Zhang, G. Srivastava, "Speech emotion recognition method in educational scene based on machine learning", *EAI Endorsed Transactions on Scalable Information Systems, EAI, 2022*. DOI: 10.4108/eai.10-2-2022.173380.

[33] K. Yang, B. Tag, Y. Gu, C. Wang, T. Dingler, G> Wadley, & J. Goncalves, "Mobile emotion recognition via multiple physiological signals using convolution-augmented transformer". In Proceedings of the 2022 International Conference on Multimedia Retrieval (pp. 562-570).

[34] G. Alhussein, M. Alkhodari, A. Khandoker, & L. Hadjileontiadis, "Novel Speech-Based Emotion Climate Recognition in Peers' Conversations Incorporating Affect Dynamics and Temporal Convolutional Neural Networks", 2023.

[35] A. Kumar, K. Sharma, & A. Sharma, A. "Empirical Analysis of Psychological Well-Being of Students During the Pandemic with Rebooted Remote Learning Mode". In Proceedings of Data Analytics and Management: ICDAM 2022 (pp. 13-29). Singapore: Springer Nature Singapore, 2023.

Dr. Aditi Sharma (Member, IEEE) is working as a Regular Assistant Professor at Thapar Institute of Engineering and Technology. She has near 5 years of experience in Academia. She has previously worked with Jaypee University of Information Technology, Waknaghat as Assistant Professor (Grade- II) in Department of Computer Science and Engineering/Information Technology. Prior to Jaypee Group, Ms. Sharma was associated with Delhi Technological University, and DIT University.

Dr. Aditi has received her Ph.D. from Delhi Technological University in the field of Affective Computing. She earned her Post-graduation, Masters in Software Engineering from Delhi Technological University in 2017. She graduated from Punjabi University Patiala in 2015 with Bachelor of Technology in Computer Science and Engineering. With a top score of 717 and an All India Rank of 571 in 2015, she has qualified GATE four times. In 2018, She also qualified UGC-NET for Assistant Professor. Affective Computing, Machine Learning, Predictive Healthcare, and Text Summarization are some of her research interests. She has 14 publications, five of which are in high impact SCI/SCIE journals. In 2022 and 2023, Dr. Sharma has received "Commendable Research Award for Excellence in Research" for her research work from Delhi Technological University. She holds professional memberships in IEEE, ACM, CSI and IAENG.

Dr. Akshi Kumar (Senior Member, IEEE) is a Senior Lecturer, and Director-PGR in the Department of Computing at the Goldsmiths, University of London, London, United Kingdom. She is a Post-doc from Federal Institute of Education, Science and Technology of Ceará, Fortaleza, Brazil and a PhD from Faculty of Technology, University of Delhi, India. She has worked as an Associate Professor at the Manchester Metropolitan University, Manchester, United Kingdom and Netaji Subhas University of Technology (NSUT), New Delhi, India and as an Assistant Professor at the Delhi Technological University (DTU), New Delhi, India.

Dr. Kumar has been endorsed by the Royal Academy of Engineering, United Kingdom as an Exceptional promise in the field of Artificial Intelligence/Data Science in 2022. She has received 9 research awards for Excellence in Research from various National and International organizations. Her name has been included in the *"Top 2% scientist of the world"* list by Stanford University, USA in 2023, 2022 and 2021. Based on the list, her current world rank within the field of Artificial Intelligence is 1415. She has published more than 95 peer-reviewed journal papers including 55 SCIE publications, 65+ conference papers with 5 best paper awards and 2 patents with Indian Patent Office. She has successfully guided 6 doctorates, 33 Master thesis candidates. She has been serving as an Associate editor and guest editor in various high impact journals with reputed publishers. Her research interests are in affective computing, social network and media analytics, NLP, and AI for pervasive healthcare.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:09:11 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3322947

Edge-based computation of super-resolution Superlet spectrograms for real-time estimation of heart rate using an IoMT based reference signal less PPG sensor

Pankaj, *Member*, *IEEE*, Ashish Kumar, *Member*, *IEEE*, Manjeet Kumar, *Member, IEEE*, Rama Komaragiri, *Sr. Member*, *IEEE*

***Abstract*— Cardiovascular disease (CVD) is one of the leading causes of the mortality rate increase. To effectively analyze wearable sensor data for providing accurate and reliable estimation of vital signs such as heart rate (HR), the use of Artificial Intelligence (AI) in wearable devices is increasing. The use of AI in designing healthcare wearable sensors is crucial to the uprising of the Internet of Medical Things (IoMT).**

The appearance of IoMT sensor technologies lets the healthcare industry shift from vis-a-vis consulting to telemedicine. IoMT sensors have transformed the healthcare industry by improving patient safety and reducing healthcare costs. This paper proposes a Photoplethysmogram (PPG) enabled wearable device in an edge-IoMT computing environment that enables users to monitor their real-time health status. A deep learning approach for automatic feature extraction is proposed in this work. The deep learning algorithm learns features from a super-resolution spectrogram computed using superlet transform. In the proposed system, a PPG signal is input, and the output layer provides information on HR. The proposed framework uses two publicly available PPG datasets to train and test the proposed edge-assisted model. The model is further evaluated using an in-house acquired PPG signal dataset. The proposed framework obtained a mean absolute error of 0.76 BPM, 1.01 BPM,1.46 BPM, and 1.79 BPM for IEEE SPC training, IEEE SPC test, BAMI-I, and BAMI-II datasets, respectively. The proposed edge-based IoMT framework satisfactorily predicts HR in real-time using reference signal less PPG sensor signal.

***Index Terms*— Edge computing, Heart Rate Estimation, Healthcare devices, IoMT sensor, Photoplethysmogram (PPG), Super-Resolution, wearable device.**

Submitted on 18 May. 2023.

Pankaj is with the Department of Electronics and Communication Engineering, Bennett University, Greater Noida, 201310, India and also with the Panipat Institute of Engineering and Technology, Panipat, 132102, India. (e-mail: er.pankaj08@gmail.com).

Ashish Kumar is with the School of Computer Science Engineering and Technology, Bennett University, Greater Noida, 201310, India. (e-mail: akumar.1june@gmail.com).

Manjeet Kumar is with the Department of Electronics and Communication Engineering, Delhi Technological University, Bawana, Delhi, 110042, India. (e-mail: manjeetchhillar@gmail.com).

Rama Komaragiri is with the Department of Electronics and Communication Engineering, Bennett University, Greater Noida, 201310, India. (e-mail: rama.komaragiri@gmail.com.).

"Copyright (c) 20xx IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org."

I. INTRODUCTION

Cardiovascular disease (CVD) is primarily a chronic disease and the leading cause of death worldwide, per the World Health Organization (WHO) reports. Out of the deaths due to CVD, four out of five are due to heart attacks and heart strokes [1].

Several effective strategies suggest reducing CVD through a healthy lifestyle and proper medication. However, the approaches to early detection of CVD risk can help society reduce the trauma of mortality and health costs. Hence, regular monitoring of cardiac health status is vital for early diagnosis and timely treatment of cardiovascular disease. Therefore, continuous and reliable monitoring of vital body signs such as heart rate (HR) is crucial in designing wearable healthcare systems [2]. Estimation of HR in real-time using wearable devices has been steadily growing. An electrocardiogram (ECG) signal is used as a gold standard technique to measure HR. However, challenges associated with long-term usage to monitor cardiac health have resulted in the acceptance of the Photoplethysmogram (PPG) signal as an alternative approach [3]. The demand for wearable health monitoring devices based on PPG is ever-increasing due to its non-invasive nature, easy-to-use, simple operation, and low cost.

There are two different ways to measure vital signs using a PPG signal. They are contact PPG and contactless PPG [4]. In contactless PPG, light focused from a distance records the PPG signal. A contact PPG measurement uses a light-emitting diode (LED) and a photodetector to record the signal. The contact PPG provides continuous real-time monitoring of vital body signs, which is impossible in contactless PPG measurements. This paper uses a contact PPG to record the PPG signal to estimate HR. The wrist is a commonly preferred site to record PPG signals using a wearable device [5]. The design of a PPG-enabled wearable health monitoring device focuses on factors such as ease of use and cost from the user perspective and computational complexity from the design perspective. It is a reality that motion artifact noises exist in every PPG signal acquired using a wearable device [6].

Motion artifacts degrade the quality of a PPG signal. Thus, for accurate and reliable estimation of vital signs such as HR, the motion artifacts spectrum must be separated from the recorded PPG signal [7]. Thus, denoising techniques suppress motion artifact components from the acquired PPG signals. The methods to remove motion artifacts from a PPG signal are divided into two categories. The first method requires a reference motion signal simultaneously recorded using

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:45:00 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3322947

additional hardware, and the second does not require a reference signal [8]. The PPG sensors available in the market use the first type to reduce the effect of motion artifacts on the PPG signal. The performance of the PPG health monitoring devices available in the market depends on the accuracy of the additional accelerometer sensor signal used to provide a reference noise signal. However, using an additional sensor to acquire a reference motion signal increases the complexity and energy requirements of the device [9]. The proposed method transforms a PPG signal into a time-frequency super-resolution spectrogram using superlet transform to improve detection accuracy. The superlet transform separates the HR and motion artifact components. The proposed method does not require additional accelerometers or sensors to eliminate motion artifacts.

The features of a super-resolution spectrogram are learned through a deep convolutional neural network (CNN) model to estimate HR in real-time. Recently, interest in personalized wearable health monitoring systems has been gaining importance due to the association of edge computing and sensor technologies [10]. Wearable devices allow users to monitor HR with daily routine activities and act as emerging tools to detect cardiac abnormalities at an initial stage. The proposed device shows the potential of wearable PPG signal analysis to track CVD in real-time. To the best of the author's knowledge, the proposed framework is the first of its kind, showing a complete model from data acquisition to HR estimation in real-time.

The rest of the proposed work is organized as the following. Section II describes the background and objective. Section III provides details on the architecture of the proposed system. The proposed methodology highlighting the potential of the super-resolution PPG spectrogram is detailed in Section IV. Model optimization is described in section V. The results analysis, comparison with state-of-the-art approaches, results and discussion are described in Section VI. Finally, the conclusion and future work are discussed in section VII.

II. BACKGROUND AND OBJECTIVE

The main hurdle in properly utilizing PPG-based health monitoring devices is the suppression of motion artifacts to estimate HR accurately. Many signal processing techniques, such as independent component analysis [11], empirical mode decomposition [12], adaptive filtering [13] [14], short-time Fourier transform [15], and wavelet transform [16], have been proposed to suppress motion artifacts from the PPG signal. The state-of-the-art methods comprise a three-step process: suppressing motion artifacts, HR estimation, and HR tracking [18], [19]. The above mentioned methods require a simultaneously recorded reference signal to remove noise [20]. The need for predefined heuristic threshold values or activity-based parameter optimization limits their usage in real-time monitoring devices.

The IEEE Signal Processing Cup 2015 (IEEE SPC) dataset is a benchmark dataset. The IEEE SPC dataset contains simultaneously recorded reference accelerometer signals, two PPG sensor signals, and a gold standard ECG signal [17]. The IEEE SPC dataset aims to provide motion artifact-contaminated PPG signal recordings to estimate HR. The ground truth HR (HR_{true}) calculated for an eight-second window from a simultaneously recorded ECG signal is also available with every recording of the IEEE SPC dataset. The mean absolute errors (MAE) of the above mentioned methods evaluated using the IEEE SPC dataset ranges between 1-3 BPM. The above mentioned methods use traditional power spectral density to estimate HR. The reliable and accurate HR tracking assumes that the probability of a significant change in HR in consecutive windows is minimal. Thus, referencing the HR value with the immediately prior HR window is used to track HR in those windows severely affected by motion artifacts [21]. In [22], a hybrid approach combining CNN and LSTM estimates HR in real-time. However, subject-specific training makes this approach complex and incompatible for implementation in personalized real-time monitoring devices. A fully binarized network that reduces computational complexity [23] is proposed to estimate HR. In [24], an end-to-end deep learning approach that computes the Fourier transform spectrum of input PPG and accelerometer signals is proposed to predict HR. The use of additional preprocessing steps increases the computational complexity of the method. The work proposed in [25] introduced a new dataset called BAMI.

Approaches present in the literature require an additional sensor and a preprocessing unit to suppress motion artifacts from the acquired PPG signal to estimate HR accurately. The additional sensors and processing step adversely affect the computational complexity and energy requirements of the device [26]. Deep learning-based methods have recently been used to extract relevant features without additional signal-processing steps to suppress motion artifacts. However, the deep learning-based automatic learning approach requires reference accelerometer signals to estimate HR. The use of additional hardware poses limitations on the energy requirement, a key hurdle while designing wearable devices.

A single PPG sensor-based method that eliminates the additional accelerometer sensor to estimate HR is proposed in [27]. However, the requirement of sparse training for different hardware specifications limits the use of the method in real-time monitoring. In [28], a clean PPG signal is generated using ECG data from a given dataset and trains the model with noisy PPG and clean PPG signal. Estimated HR is calibrated by using spectrum analysis. A calibration parameter is required to refine the estimated HR value, which is constant for a complete dataset. Deep Convolutional Neural Networks (CNN) play a significant role in machine learning (ML). CNN can process images like a human brain. A fully automatic learning principle of CNN allows the network to extract relevant features of an input image across different layers. In the early stages of CNN-based HR estimation studies, the proposed models are trained only using the IEEE SPC dataset. However, training models with a small dataset cannot assure accurate estimation in real-time. Thus, a recent study based on deep learning used additional publicly available datasets to improve the accuracy. In [29], a method is proposed to improve the reliability of a deep neural network to estimate HR. Conventional time and frequency-based signal processing methods are inaccurate and require fewer operations to process a signal at high rates.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:45:00 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3322947

Extracting HR using conventional time-frequency approaches is challenging as these are designed with predefined conditions. The healthcare industry needs generalized approaches to estimate HR in real-time. Due to the proliferation of real-time health monitoring wearable devices, implementing the Internet of Medical Things (IoMT) sensor at the edge layer is becoming a new phenomenon. Integrating IoMT with the wearable device allows the user to monitor HR during busy hours and be alert about changes in health conditions. Monitoring health conditions helps take timely preventive measures to save a life. Most denoising methods require motion artifact filtering and manual feature selection, which increase the computational complexity of the device. Hence, it cannot be implemented using edge-assisted IoMT sensors. In [30], a temporal convolutional network-based framework is proposed to estimate the HR using a PPG sensor-enabled wearable device. This work requires a post-processing stage to refine the estimated value of HR. The protocol used in the post-processing stage is fixed and the same for all subjects. Fixing a parameter value based on limited subjects does not indicate the work's generalized capability. Moreover, this method requires a reference accelerometer signal to compute HR. In [31], a deep neural network architecture search-based HR estimation approach is proposed.

This work shows extensive performance in estimating HR using the PPG signal but requires a reference signal. The need for a three-axis accelerometer sensor for reference signals increases the hardware and computational complexity of the framework. Further, maintaining a correlation between the accelerometer and PPG signals is challenging. In [32], the HR estimation framework is proposed using cascaded CNN and LSTM networks. The performance of the proposed network depends on three input signals: the PPG signal, the first derivative of the PPG signal, and the second derivative of the PPG signal. Correlating three signals to suppress the effect of motion artifacts is challenging in real-time. Moreover, the proposed work uses a GPU-based workstation to deploy the framework for HR estimation. Hence, deploying the method in [A.3] is not feasible on edge-based hardware.

In [33], a centralized state sensing algorithm-based HR estimation framework is proposed. The proposed framework uses a raw PPG signal to predict HR using CNN-based models. A GPU-based workstation is used to train and test the model for HR prediction. These GPU-based workstations increase the power consumption of the device. Wearable devices require low power consumption to increase the extended battery life. The proposed work conceptualizes the potential of IoMT with edge computing to estimate HR in real-time using PPG signals. This work proposes an end-to-end automatic feature extraction technique based on a deep learning approach. The proposed work trains the deep neural network model with the superlet transform based super-resolution time-frequency spectrogram to increase the accuracy of the model. Additionally, the proposed work, compressed and optimized the proposed deep neural network model using TFLite, ensures efficient deployment on memory-constrained wearable edge devices. This process enables faster inference, low power consumption, and processes PPG signal near the subject (as a wearable device). The lightweight version of the proposed model requires less memory and computation time with identical performance to the uncompressed model at the edge device.The proposed method computes a super-resolution spectrogram of an input PPG signal using superlet-transform to estimate HR. The proposed method neither requires a reference accelerometer signal nor multiple PPG sensors.

Further, the proposed method inherently filters motion artifacts and does not require additional signal preprocessing steps. The proposed model also does not require manual feature extraction steps. The proposed work is a hybrid approach of superlet transform and deep learning-based lightweight HR estimation method. The proposed work is compared with a reference HR extracted from a simultaneously recorded ECG signal, used as a true label of the PPG spectrogram. The reference ECG signals are used as a model to learn the correlation between spectrogram and true HR value. To the best of our knowledge, the proposed framework, for the first time, introduces the development and design of a deep learning-based, edge-assisted model for real-time HR estimation.

The key highlights of the proposed work are the following:

1. The proposed framework is used to design an edge-based remote health monitoring device for real-time HR assessment.
2. For the first time in literature, this work introduces a computationally efficient deep control algorithm residing at the edge layer to process the subject's health in real-time.
3. The proposed framework is tested with various dimensions of the input spectrogram to evaluate the variation in the accuracy of the model.
4. An optimized two-dimensional (2-D) CNN-based deep model, centrally trained using two publicly available datasets, is transferred to the edge device for real-time HR estimation.

III. SYSTEM ARCHITECTURE

This section describes various blocks required to realize the proposed wearable device. Fig. 1 shows the proposed device as an edge device. This device proposes an edge-assisted wearable device that can sense, process, and configure the signal in real-time.

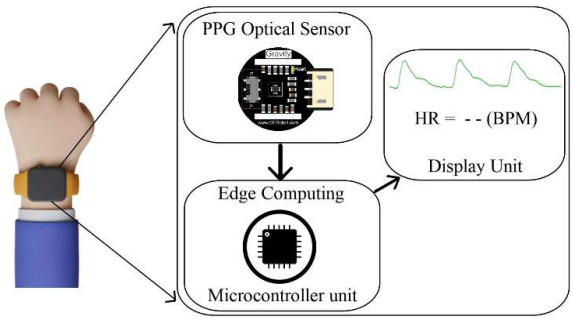


Fig. 1. Block diagram representation of the proposed framework

A PPG optical sensor is used to acquire a PPG signal in real-time. The analog PPG signal is digitized using a signal converter. The digitized PPG signal is first transformed into a time-frequency super-resolution spectrogram to separate the clean PPG signal component and motion artifact component from the raw PPG signal. The proposed device is designed to

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:45:00 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3322947

handle edge computation at the device level. As shown in Fig. 1, preprocessing of the signal, deep learning-based feature extraction and regression layer-based vital sign estimation are accomplished in the edge layer. The Edge resides close to the sensor. The PPG signal from the sensor is processed and analyzed in the Edge in real-time, eliminating network latency. The display connected to the edge layer displays the estimated vital signs. The important information on vital signs can be further shared with a medical practitioner and others through the cloud. The deep learning CNN model is trained using the super-resolution spectrogram.

Core components: The edge hardware is realized using a quad-core Cortex-A72 64-bit processor operating at 1.5 GHz. The following peripherals and functions associated with the signal conditioning unit complete the hardware setup.

- A PPG optical sensor to acquire real-time PPG signal.
- Analog to digital converter to digitize acquired analog PPG signal.
- Recording and storing of real-time PPG signal
- Algorithm design to preprocess the acquired signal
- Conversion of preprocessing data to superlet transform-based 2-D spectrogram
- Training and testing the proposed model using the spectrograms obtained from publicly available dataset
- Design an algorithm to process the acquired data through edge-assisted hardware
- Estimation and analysis of HR using design hardware.

A 64 GB memory fulfills the memory requirements of the proposed device. The complete prototype is shown in Fig. 2.

IoT sensor description: The first block is the front-end IoMT sensor module used to acquire real-time PPG signals and to provide reference HR value for performance analysis. The GRAVITY SEN0203™ PPG sensor is attached to the right-hand wrist to acquire a real-time PPG signal. The PPG sensor interfaces with the edge unit to preprocess the acquired raw data. The sensor consists of an integrated circuit, SON1303, that uses reflective mode to record the PPG signals in real-time.

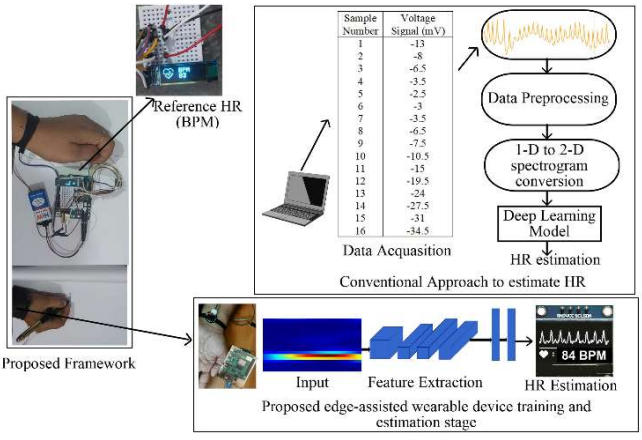


Fig. 2. Block-wise representation of the proposed framework

The transmitter module consists of a double green LED of wavelength 570nm. The receiver consists of a photo-diode, low-noise preamplifier with a built-in optical amplifier with a peak sensitivity wavelength of 570 nm. To the left hand, a Gravity MAX30102 heart rate sensor is attached to acquire a reference HR value to validate the performance of the proposed method. The sensor is interfaced with the processing unit to provide HR (BPM) in real-time, as shown in Fig. 2.

Edge layer: The edge hardware is realized using a quad-core Cortex-A72 64-bit processor. Fig. 2 shows the main processing blocks of the proposed wearable device on a system architecture level and depicts the step-by-step process performed after data acquisition on the edge device. The processing of the edge layer is divided into three parts. First, the acquired PPG signal is subjected to a preprocessing block. In the second part, in the Edge layer, the 1-D PPG signal is converted into a 2-D spectrogram through preprocessing. The third layer consists of a deep CNN model for real-time HR estimation.

IV. METHODOLOGY

The proposed algorithm is tested using the IEEE SPC dataset and BAMI dataset. The algorithm is then implemented on the edge device. The PPG signals are recorded, and HR is estimated in real-time using the proposed hardware setup. A complete description of the methodology is as follows.

Dataset: The IEEE SPC and the BAMI datasets are used to train and test the performance of the proposed algorithm. Both datasets contain multichannel PPG sensor signals and a simultaneously recorded three-axis accelerometer signal measured using a wearable device. Both datasets include reference HR named HRtrue measured for a window of 8 seconds using a simultaneously acquired ECG signal. The sampling frequency is 125 Hz in the IEEE SPC dataset and 50 Hz in the BAMI dataset. The proposed framework is tested using a small in-house recorded dataset, including activities close to real-life activities, to check the generalized capability of the proposed method.

Eight subjects participated in the data collection: two female and six male subjects, aged 21 – 40. With the wearable hardware setup shown in Fig. 2, it is possible to collect data during daily life activities. Each subject follows a data collection protocol to maintain uniformity in data collection. The PPG signal is recorded using the right wrist PPG sensor. Each record is acquired for a total duration of four minutes. For subjects one to five, during the initial thirty seconds, the subject is at rest.

During the next 120 seconds, the subject writes something on paper. During the next sixty seconds, the subject bends down the finger to touch the finger base joint. During the last thirty seconds, the subjects are at rest. The subjects six to eight are at rest for the initial thirty seconds. During the next thirty seconds, the subject slightly rolls the arm in a 360-degree format. During the next sixty seconds, the subject bends the arm. During the next thirty seconds, the subjects are at rest. During the next sixty seconds, the subjects walk inside the lab. The subjects step up for the next thirty seconds with the left and right feet. This process continues to imitate a running pose in our lab for sixty seconds. The HR value obtained with the PPG signal acquired using the left hand is used as a ground truth HR. The left-hand remains stable to provide reference HR and further evaluate the proposed method. The sampling frequency of both the PPG signals is 100 Hz.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:45:00 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3322947

Preprocessing: During the preprocessing stage, the recorded PPG signal from the dataset is divided into consecutive eight-second windows. After the first window, each consecutive PPG signal window shifts by two seconds. Thus, after the first window, a window of eight-second duration has a six-second overlap with the previous window. The segmented eight-second PPG signal windows are subject to a fourth-order Butterworth bandpass filter with cutoff frequencies of 0.4 Hz to 4 Hz [13]. The filtering stage filters out the signal components that lie outside the range of cardiac activity. The filtered signals are normalized to zero mean and unit variance to extract meaningful information from the amplitude-limited data.

The superlet transform spectrogram: PPG-based HR estimation is unfortunately affected by motion artifacts, thus reducing the utility of PPG as an HR estimation device in real-time. The accurate and reliable estimation of HR in real-time is the main feature of the proposed devices. Separation of heart and motion artifacts components is challenging as both PPG signal and motion artifacts signal components possess overlapping frequency spectra between 0.4 Hz to 4 Hz. Due to spectral overlap between heart and motion artifacts components, estimation of HR may result in false information. A superlet represents the motion artifact contaminated PPG signal in the time-frequency domain by separating the peaks related to HR and motion artifacts. Thus, the superlet provides a robust approach to accurately estimating HR in real-time. The wavelet is normalized to study different properties of the PPG signal in the time-frequency spectrogram. Normalization enables detecting self-similar features across scales (compressed or dilated). Normalization ensures that the features receive the same normalized amplitude in the representation if the feature has the same shape and peak amplitude for a different order of wavelet [34]. With normalization, a wavelet tends to concentrate at the dominant frequency in the signal, represented by a high-intensity lobe. The high-intensity lobe is where the scaled wavelet frequency coincides with the signal's local frequency.

Superlet transform involves multiple wavelets with a fixed center frequency to provide super-resolution by utilizing different cycles.

$$SL_{f,g} = \left\{ \psi_{s,d_k} \middle| d_k = d_1, d_2, \cdots, d_g \right\} \quad (1)$$

The mathematical representation of a superlet is given by Eq. (1). Here, g represents the number of wavelets used in the superlet (order). The superlet transform decreases the redundancy of the representation for the higher frequency by increasing the order of the superlet. d_1 is the first wavelet base cycle used for the analysis. In superlet, the value of different cycles defining the order of the wavelet can be considered using a multiplicative or additive rule [35]. This work uses the multiplicative rule given by $d_i = i \times d$, $i = 1, 2, \cdots, g$.

With multiple wavelets of different cycles, the response of the superlet to the preprocessed PPG signal $S[n]$ is given by the geometric mean of the individual wavelet response represented in Eq. (2).

$$R\left[SL_{f,g} \right] = \sqrt[g]{\prod_{i=1}^{g} R\left[\psi_{s,d_i} \right]} \quad (2)$$

The geometric mean of the time-frequency information of the PPG signal obtained with multiple wavelets represents a bright red peak. The bright red peak indicates a large amplitude used to estimate the HR. A dark blue or yellow peak will represent the component that does not have the same feature with multiple wavelets, such as noise.

Fig. 3 represents a 3-D spectrogram of a PPG signal obtained using a superlet plotted in the time-frequency domain. The color represents the amplitude. A bright red indicates a large amplitude, and a dark blue color depicts a small amplitude.

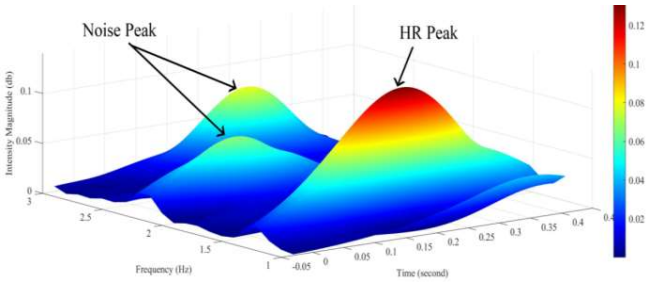


Fig. 3. Separation of noise peak and dominant HR peak using superlet transform

As observed from Fig. 3, the separation of the HR peak from a motion artifact peak in a corrupted PPG signal is evident. A corresponding 2-D view of the super-resolution spectrogram in the time-frequency domain is shown in Fig. 4. Fig. 4 is a 2-D representation of a 3-D spectrogram in which the amplitude (third dimension) is represented by color intensity. HR modulates the PPG signal; hence, HR information is concentrated as lobes in the super-resolution spectrogram, centered on a frequency corresponding to the HR.

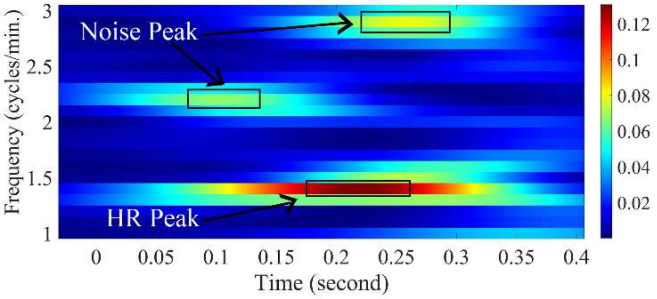


Fig. 4. 2-D representation of superlet transform spectrogram

This 2-D spectrogram is subjected through a CNN using the true HR value as the predicted target.

During the training phase of the proposed work, the relationship between the change in color intensity of an eight-second window spectrogram of a PPG signal and the true HR computed from the ECG is learned. Instead of mapping the fixed image region with true HR value, the proposed work introduces a self-adaptive color intensity region selection framework by fusing high-intensity peak regions of different spectrograms for HR estimation. Once trained based on given inputs, the model predicts the HR for the test PPG dataset. The obtained HR prediction values from the trained model are verified using the true HR value and the mean absolute error, estimated in BPM over the number of observations.

The core idea of the proposed superlet transform-based spectrogram is to distinguish the periodicity of HR peaks and motion artifact peaks in the time-frequency domain to train the model for better accuracy. The 2-D-CNN layer in the deep

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:45:00 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3322947

learning technique learns the feature from the 2-D spectrogram. If the input spectrogram data is sufficient, the deep CNN model can predict the HR reliably and accurately. Fig. 5 shows the proposed framework as an edge device.

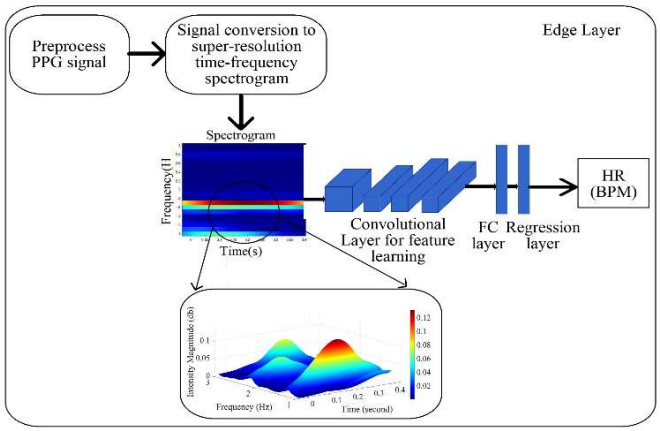


Fig. 3. Detailed flow diagram of proposed framework as an edge device.

The proposed device eliminates the latency issue by introducing an edge platform more suitable for real-time human health monitoring. Edge computing technology brings the process data close to the user to reduce latency.

V. MODEL ARCHITECTURE

The potential of CNN has been shown in many fields as CNN uses the both temporal and spatial correlation of input data. A convolutional layer slices the input into small parts (receptive fields). A convolutional layer extracts the required features. The convolutional layer's parameters consist of kernels or learnable filters. Each filter produces a 2-D activation map convolved across the input image volume, computing the dot product between filter and input entries. The activation map shows the features detected by learnable filters for a given input spectrogram.

A pooling layer allows the network to extract a combination of similar features extracted by the convolutional layer. Thus, combining the convolutional layer and pooling layer makes the CNN automatically extract features. The proposed framework acts as a supervised framework by converting the PPG signal into a superlet transform 2-D time-frequency spectrogram using deep neural networks. CNN models consist of 2-D layers characterized by convolutional filters that slide over the input spectrogram as per the assigned weights. The proposed CNN model consists of four convolutional layers. The kernel size for each layer is 3 × 3. After each convolutional layer, a maximum pooling layer is inserted with a 2 × 2 kernel size. A dropout layer is inserted after two convolutional layers. The input to a convolution layer is the previous layer's output, as shown in Fig. 6. During the training phase, the proposed model is executed for 250 epochs with batch size 32.

An early stopping functionality is imposed to avoid the model's over-fitting during training. For a given number of epochs, if the model does not improve in metric with patience 4, the model reduces the learning rate by a factor of 0.5. All layers except the last layer use ReLU as an activation function. The last layer is the estimation unit, based on a single neuron-dense layer performing a regression task. During the training phase, the proposed CNN model learns the high-intensity peak pattern as HR patterns using the super-resolution 2-D superlet spectrogram to estimate the HR. The testing phase uses a super-resolution spectrogram from an independent testing dataset.

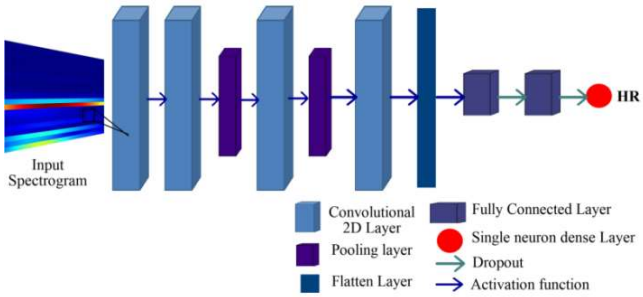


Fig. 4. Deep CNN layer architecture of the proposed CNN model

The feature extracted from the convolutional layer is passed to the dropout layer with a dropout ratio of 0.2. This means that 20% of the neurons are ignored in the subsequent layer, thus making the system less sensitive to overfitting.

The output of the last convolutional layer is fed to the two fully connected layers with 128 and 64 neurons. The output of the fully connected layer is subjected to the regression task consisting of a single neuron to predict the HR value. The parameters of the CNN layer, such as spectrogram size, learning rate, filter size, and the number of fully connected layers, are optimized to maximize the performance of estimated HR efficiency on the training and test datasets.

MAE is used to confirm the performance of the proposed model as a loss function. MAE is the mean of the absolute difference between the HRtrue and the $HR_{estimated}$ using the proposed framework given by Eq. (3).

$$MAE = \frac{1}{W}\sum_{k=1}^{W}\left|HR_{true}(k) - HR_{estimated}(k)\right| \quad (3)$$

Here, W is the length of the total 8-second windows, and k is the window number. Root mean square error is also assessed to study the accuracy of the proposed model. The performance of the proposed method is analyzed by calculating these metrics.

VI. RESULTS AND DISCUSSION

This section presents the performance evaluation of the proposed framework on a PPG signal acquired in real-time. The HR is estimated using an edge-assisted hardware setup. The framework is trained and tested using the IEEE SPC and BAMI datasets. This work uses a single PPG sensor data. Thus, it does not require additional accelerometer or other sensor data for HR estimation. The proposed framework is extensively trained using signals from two datasets, making the model effectively learn physiology and physical exercises. The dataset used in the testing phase is removed from the training phase to estimate how our model performs on any new dataset. A 10-fold cross-validation approach is used to split the train and test datasets. It shows a generalized framework of the proposed work to overcome the overfit and underfit problem of model training. Table I shows the performance comparison of the proposed

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:45:00 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3322947

TABLE I
PERFORMANCE MATRICES COMPARISON WITH STATE-OF-THE-ART TECHNIQUES

Method	Dataset	Algorithm	MAE
TROIKA [17]	IEEE SPC Training IEEE SPC Test	Singular spectrum analysis	2.34 BPM
JOSS [18]	IEEE SPC Training IEEE SPC Test	spectral subtraction	1.28 BPM
Khan [12]	IEEE SPC Training	Signal decomposition	1.02 BPM
SpaMa [19]	IEEE SPC Training IEEE SPC Test	spectral filtering	0.89 BPM 3.36 BPM
Temko [13]	IEEE SPC Training IEEE SPC Test	Wiener filtering	1.02 BPM 3.01 BPM
Nathan [26]	IEEE SPC Training	Particle filter	1.56 BPM
FSM [20]	IEEE SPC Training	Wiener filtering	0.99 BPM
Motion [36]	IEEE SPC Training IEEE SPC Test	Recursive Wiener filtering	1.02 BPM 1.85 BPM
Arun [14]	IEEE SPC Training IEEE SPC Test	Adaptive filtering	1.03 BPM 1.89 BPM
FDM [8]	IEEE SPC Training IEEE SPC Test BAMI-I BAMI-II	Fourier decomposition and FFT	1.66 BPM 1.87 BPM 1.33 BPM 1.45 BPM
CNN [24]	IEEE SPC Training PPG-Dalia	CNN	4 BPM 7.65 BPM
PPGNet [34]	IEEE SPC Training IEEE SPC Test	Inception + LSTM	3.36 BPM 12.48 BPM
CorNet [22]	IEEE SPC Training IEEE SPC Test	CNN+LSTM	4.67 BPM 5.55 BPM
BinCor Net [23]	IEEE SPC Training IEEE SPC Test	Bin. CNN+LSTM	6.78 BPM 7.32 BPM
Chung [25]	IEEE SPC Test IEEE SPC Test BAMI-I BAMI-II	CNN+LSTM	0.67 BPM 0.86 BPM 1.39 BPM 1.46 BPM
Deep Heart[28]	IEEE SPC Training	DnCNN	1.61 BPM
Deep Pulse [29]	IEEE SPC Training IEEE SPC Test BAMI-II	TCNBest	2.76 BPM 5.05 BPM 2.38 BPM
Proposed work	IEEE SPC Training IEEE SPC Test BAMI-I BAMI-II Measured data	Super-resolution and CNN	**0.76 BPM** **1.01 BPM** **1.46 BPM** **1.79 BPM** **1.27 BPM**

work with state-of-the-art techniques. The Pearson correlation coefficient and the Bland-Altman plot analysis are considered to assess the efficacy of the proposed framework, as shown in Table I. A correlation value of 0.997 indicates a high similarity between $HR_{estimated}$ and HR_{true}. Upper limits of agreement (LOA) and lower LOA are used to compare $HR_{estimated}$ and HR_{true}. Table II compares the proposed work with state-of-the-art HR estimation algorithms using MAE as a metric. The MAE values of the proposed framework using IEEE SPC training, IEEE SPC test, BAMI-I, and BAMI-II datasets are 0.76 BPM, 1.01 BPM, 1.46 BPM, and 1.79 BPM, respectively.

The proposed Edge-based IoMT framework outperforms all previously reported methods in Table II. The advantage of the proposed Edge computing IoMT sensor-enabled wearable device to estimate HR in real-time is that the proposed IoMT framework used the TFLite tool and post-quantization technique to compress and optimize the deep neural network model into the lightweight model.

The proposed framework demonstrated MAE for IEEE SPC training, IEEE SPC test, BAMI-I, and BAMI-II dataset, exceeding the MAE reported in [25].

However, the work reported in [25] uses a reference accelerometer sensor signal to generate a reference signal to suppress the effect of motion artifact from the acquired PPG signal. The use of additional accelerometer sensors increases the computational complexity of the system. Moreover, the algorithm's accuracy proposed in [25] depends on the correlation between the PPG signal and the reference accelerometer sensor signal. Maintaining a high correlation between both signals in real-time is challenging and increases overall system complexity and energy consumption. The efficacy of the work proposed in [8], [12] – [14], [17] - [20], [26] and [36] depends on the post-processing stage used to refine the value of HR estimation. The need for a post-processing stage increases the computational complexity of the algorithm.

TABLE II
COMPARISON OF MAE VALUES ESTIMATED USING THE PROPOSED TECHNIQUE WITH OTHER EXISTING METHODS

Method	Pearson correlation	Upper LOA (BPM)	Lower LOA (BPM)
TROIKA [17]	0.992	4.79	−7.26
JOSS [18]	0.993	5.41	−5.94
Nathan [26]	-	4.45	-4.75
FSM [20]	0.997	2.96	-2.60
Motion [36]	0.992	7.30	-7.29
Arun [14]	0.990	7.8	− 8.4
FDM [8]	0.993	7.79	−6.78
Chung [25]	0.996	4.66	-4.81
Deep Heart [28]	0.990	7.30	−6.68
Proposed Work	0.997	2.34	-2.23

Moreover, optimizing the protocol in the post-processing stage is challenging for PPG signals acquired in real-time. The literature listed in Table II lack generalized capabilities. None of these algorithms are deployed in a wearable device.

This paper presents an edge-assisted complete deep learning framework for HR estimation in real-time using PPG signal only. The proposed deep neural model is sequentially arranged with a 2-D convolutional layer, a max pooling layer, and a fully connected layer, which are incorporated to estimate HR. No additional layer structure is required to suppress motion artifacts. The performance of the super-resolution spectrogram-based HR estimation approach with traditionally proposed time-frequency approaches is further analyzed. The proposed framework is trained and tested separately with 2-D time-frequency spectrograms created using superlet, wavelet, and short-time Fourier transform. Fig. 7 compares the $HR_{estimated}$ obtained using all three trained models with the HR_{true} value of a randomly selected subject from the IEEE SPC dataset.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:45:00 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3322947

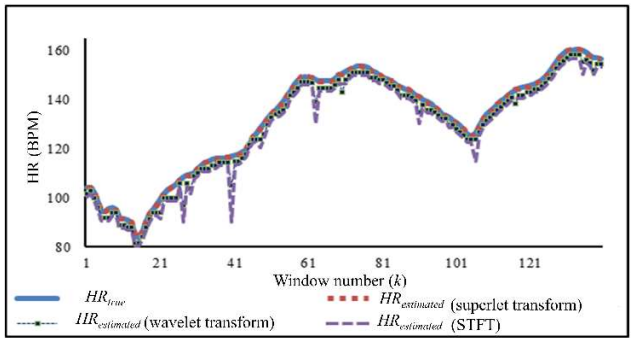


Fig. 5. Comparison between HR_{true} and $HR_{estimated}$ evaluated using state-of-the-art wavelets and STFT approach.

Fig. 7 shows that estimated HR using superlet transform is accurate for most windows. Hence, the performance of the proposed framework is superior to conventional motion artifact suppression signal processing techniques. To further improve the results of the proposed framework, increase the number of fully connected layers of various depths and widths but find no changes in the performance metric. Most of the techniques in the literature require a PPG sensor and a reference accelerometer signal. However, for a real-time monitoring device, using an accelerometer signal to provide a reference signal to estimate HR has no real significance as the accelerometer signal intensity need not correlate with a person's daily life activity.

Nevertheless, in our results from IEEE SPC training, IEEE SPC test, BAMI-I, and BAMI-II dataset, the HRs have estimated accurately during the process of motion artifacts affected windows, and even HRs estimation results accurate during the windows when true HR for those windows is randomly increased. None of the state-of-the-art techniques investigate the efficacy of their proposed model in real-time as the acceleration sensor signal need not necessarily correlate with the exercises. The key highlight of the proposed device is that the model's accuracy in estimating HR does not depend upon any motion artifacts suppression signal processing algorithm and additional accelerometer sensor data.

All the above aspects are essential when designing an energy-efficient Edge computing IoMT sensor-enabled wearable device to estimate HR in real-time. Deep learning shows superior performance in vital sign estimation. However, high-power requirements due to multiple processing show challenges in designing the wearable device due to operating life. Reference [37] suggested that shifting the processing task to the cloud can reduce the computation burden of the device. However, the latency issue makes this approach impractical as the device has to monitor the HR in real-time. Therefore, the proposed framework introduces a completely energy-efficient, Edge-assisted module to estimate HR in real-time.

In addition, the proposed framework downsizes the spectrogram sizes to reduce the computational parameter further. We test the model at different downsizes 320×320, 296 ×296, 224×224, 196×196, 144 ×144, 128 ×128. The MAE value remains the same for 224×224 and higher image sizes. However, the loss function value drastically increases when the model tests for lower image sizes. Hence, 224×224 is considered to estimate HR for a different dataset. The proposed work shows the advantage of image size selection during the design of the proposed device.

Fig. 8 shows a box plot of the HR_{true} and $HR_{estimated}$ using the data processed using the proposed method. The boxes and median value of HR_{true} and $HR_{estimated}$ are almost identical, demonstrating the potential of the proposed framework.

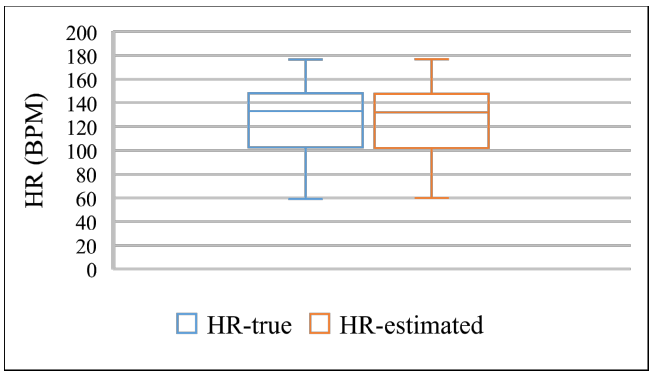


Fig. 6. Comparison between HR_{true} and $HR_{estimated}$ using the data processed using the proposed method.

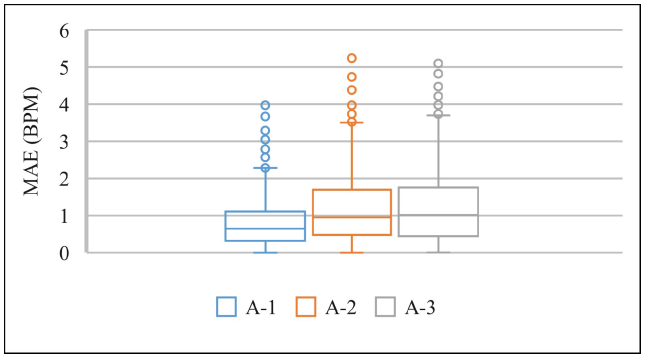


Fig. 7. Box plot representing MAE of different types of exercises.

The box plot in Fig. 9 shows the MAE of proposed frameworks under different exercises. In the A-1 exercise database, the subjects run on a treadmill at various speeds.

In exercises A-2, subjects performed various forearm and upper arm exercises such as push-ups, shaking hands, and stretching. In exercises A-3, subjects performed intensive forearm and upper arm movements such as boxing. Fig. 9 shows similar performance in all three types of exercise. The results show that the generalization of the proposed framework encompasses all types of signals recorded with different exercises without a significant deviation in the performance.

The proposed model is also tested on the PPG data acquired in real-time. The HR value obtained using a reference HR device is considered HR_{true} to verify the performance of the proposed edge-assisted hardware. A total of one hundred and twenty window spectrograms are used to estimate the HR in BPM ($HR_{estimated}$). For the same duration, the HR value of the reference device is considered as HR_{true}. Fig. 10 shows the comparison of $HR_{estimated}$ with HR_{true}. HRestimated accurately tracks the HRtrue for most windows, irrespective of the physical motion occurring while recording the dataset. One of the limitations of this step is that no standard ground truth HR is available to compare the obtained results. Thus, in the future, we aim to involve an ECG device under the supervision of

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:45:00 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3322947

medical personnel to record ground truth HR using ECG simultaneously.

Fig. 8. Comparison between HR_{true} and $HR_{estimated}$ recorded with the proposed wearable device

VII. Conclusion

The proposed framework uses two publicly PPG datasets to estimate HR and one short-duration in-house recorded dataset. While recording these datasets, the exercises are close to real-life activities. Moreover, the proposed work significantly outperforms the state-of-the-art methods, showing the potential of super-resolution spectrogram-based HR estimation. The proposed framework for HR estimation incorporates the potential of both deep learning and signal processing approaches. The proposed device demonstrates better computational complexity and memory requirement performance. The proposed trained model is embedded in a virtualized edge layer connected to the physical layer. An edge-assisted health monitoring device processes the data on the wearable device itself. The edge-assisted wearable device doesn't transfer the recorded PPG data to any centralized server or cloud for processing. Thus, the latency is reduced as the computation is done on the device.

The proposed framework shows the proposed method's efficiency in estimating HR during the presence of motion artifacts during real-life activities. The proposed framework demonstrated compelling performance in HR estimation. The proposed device allows the work to be embedded in a user-centric, personalized healthcare device. Finally, the computationally efficient edge-assisted device is a novel contribution, providing features towards an embedded application to monitor the real-time CVD state.

References

[1] World Health Organization. (2022). Health topics: cardiovascular diseases. Fact Sheet. Available online: https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(CVDs) (accessed on 11 December 2022). identification of the CVD.

[2] Pankaj, A. Kumar, R. Komaragiri, and M. Kumar, "A Review on Computation Methods Used in Photoplethysmography Signal Analysis for Heart Rate Estimation," Archives of Computational Methods in Engineering, vol. 29, no. 2, pp. 921-940, May 2021.

[3] S. Nabavi and S. Bhadra, "A Robust Fusion Method for Motion Artifacts Reduction in Photoplethysmography Signal," in IEEE Transactions on Instrumentation and Measurement, vol. 69, no. 12, pp. 9599-9608, Dec. 2020.

[4] C. H. Cheng, K. L. Wong, J. W. Chin, T. T. Chan, and R. H. Y. So, "Deep learning methods for remote heart rate measurement: A review and future research agenda," *Sensors*, vol. 21, pp. 1–32, Sept. 2021.

[5] P. Mohapatra, P. Sreeletha and M. Sivaprakasam, "A Yellow–Orange Wavelength-Based Short-Term Heart Rate Variability Measurement Scheme for Wrist-Based Wearables," IEEE Transactions on Instrumentation and Measurement, vol. 67, no. 5, pp. 1091-1101, May 2018.

[6] Pankaj, A. Kumar, R. Komaragiri, and M. Kumar, "Analysis of photoplethysmogram signal to estimate heart rate during physical activity using fractional Fourier transform–A sampling frequency independent and reference signal-less method". Computer Methods and Programs in Biomedicine, 229, 107294, Jan. 2023.

[7] A. John, S. J. Redmond, B. Cardiff and D. John, "A Multimodal Data Fusion Technique for Heartbeat Detection in Wearable IoT Sensors," in IEEE Internet of Things Journal, vol. 9, no. 3, pp. 2071-2082, Feb. 2022.

[8] Pankaj, A. Kumar, R. Komaragiri, and M. Kumar, "Reference signal less Fourier analysis-based motion artifact removal algorithm for wearable photoplethysmography devices to estimate heart rate during physical exercises," *Computers in Biology and Medicine*, vol. 141, p. 105081, Feb. 2022.

[9] S. Sakib, M. M. Fouda, and Z. M. Fadlullah, "A Rigorous Analysis of Biomedical Edge Computing: An Arrhythmia Classification Use-Case Leveraging Deep Learning," *IoTaIS 2020 - Proceedings: 2020 IEEE International Conference on Internet of Things and Intelligence Systems*, pp. 136–141, Jan. 2021.

[10] S. Dewanto, M. Alexandra, and N. Surantha, "Heart rate monitoring with smart wearables using edge computing," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 3, pp. 141–148, Jan. 2020.

[11] B. S. Kim and S. K. Yoo, "Motion artifact reduction in photoplethysmography using independent component analysis," *IEEE Transactions on Biomedical Engineering*, vol. 53, pp. 566–568,April 2006.

[12] E. Khan, F. al Hossain, S. Z. Uddin, S. K. Alam, and M. K. Hasan, "A Robust Heart Rate Monitoring Scheme Using Photoplethysmographic Signals Corrupted by Intense Motion Artifacts," *IEEE Transactions on Biomedical Engineering*, vol. 63, pp. 550–562, Aug. 2016.

[13] A. Temko, "Accurate Heart Rate Monitoring during Physical Exercises Using PPG," *IEEE Transactions on Biomedical Engineering*, vol. 64, pp. 2016–2024, March 2017.

[14] K. R. Arunkumar and M. Bhasker, "Heart rate estimation from wrist-type photoplethysmography signals during physical exercise," *Biomedical Signal Processing and Control*, vol. 57, p. 101790, March 2020.

[15] D. Zhao, Y. Sun, S. Wan, and F. Wang, "SFST: A robust framework for heart rate monitoring from photoplethysmography signals during physical activities," *Biomedical Signal Processing and Control*, vol. 33, pp. 316–324, Dec. 2016.

[16] A. Biswas, M. Singha Roy, and R. Gupta, "Motion artifact reduction from finger photoplethysmogram using discrete wavelet transform," In Recent trends in signal and image processing, vol. 727, pp. 89-98, Jan. 2019.

[17] Z. Zhang, Z. Pi, and B. Liu, "TROIKA: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise," IEEE Transactions on Biomedical Engineering, vol. 62, pp. 522–531, Feb. 2015.

[18] Z. Zhang, "Photoplethysmography-based heart rate monitoring in physical activities via joint sparse spectrum reconstruction," IEEE Transactions on Biomedical Engineering, vol. 62, pp. 1902–1910, Aug. 2015.

[19] S. M. A. Salehizadeh, D. Dao, J. Bolkhovsky, C. Cho, Y. Mendelson, and K. H. Chon, "A novel time-varying spectral filtering algorithm for reconstruction of motion artifact corrupted heart rate signals during intense physical activities using a wearable photoplethysmogram sensor," Sensors, vol. 16, Dec. 2015.

[20] H. Chung, H. Lee, and J. Lee, "Finite State Machine Framework for Instantaneous Heart Rate Validation Using Wearable Photoplethysmography during Intensive Exercise," IEEE Journal of Biomedical and Health Informatics, vol. 23, pp. 1595–1606, Sept. 2018.

[21] K. R. Arunkumar and M. Bhaskar, "CASINOR: Combination of adaptive filters using single noise reference signal for heart rate estimation from PPG signals," Signal, Image and Video Processing, vol. 14, pp. 1507–1515, Nov. 2020.

[22] D. Biswas et al., "CorNET: Deep Learning Framework for PPG-Based Heart Rate Estimation and Biometric Identification in Ambulant Environment," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 13, pp. 282–291, Jan. 2019.

[23] L. G. Rocha *et al.*, "Binary CorNET: Accelerator for HR Estimation from Wrist-PPG," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 14, pp. 715–726, June 2020.

[24] A. Reiss, I. Indlekofer, P. Schmidt, and K. Van Laerhoven, "Deep PPG: Large-scale heart rate estimation with convolutional neural networks," *Sensors*, vol. 19, July 2019.

[25] H. Chung, H. Ko, H. Lee, and J. Lee, "Deep Learning for Heart Rate Estimation from Reflectance Photoplethysmography with Acceleration Power

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:45:00 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3322947

Spectrum and Acceleration Intensity," *IEEE Access*, vol. 8, pp. 63390–63402, March 2020.
[26] V. Nathan and R. Jafari, "Particle Filtering and Sensor Fusion for Robust Heart Rate Monitoring Using Wearable Sensors," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, pp. 1834–1846, Dec. 2017.
[27] M. Panwar, A. Gautam, D. Biswas, and A. Acharyya, "PP-Net: A Deep Learning Framework for PPG-Based Blood Pressure and Heart Rate Estimation," *IEEE Sensors Journal*, vol. 20, pp. 10000–10011, April 2020.
[28] X. Chang, G. Li, G. Xing, K. Zhu, and L. Tu, "DeepHeart: A Deep Learning Approach for Accurate Heart Rate Estimation from PPG Signals," *ACM Transactions on Sensor Networks*, vol. 17, pp. 1–18, Jan. 2021.
[29] D. Ray, T. Collins, and P. V. S. Ponnapalli, "DeepPulse : An Uncertainty-aware Deep Neural Network for Heart Rate Estimations from Wrist-worn Photoplethysmography," July 2022.
[30] Burrello, A., Pagliari, D. J., Risso, M., Benatti, S., Macii, E., Benini, L., & Poncino, M. (2021). Q-ppg: Energy-efficient ppg-based heart rate monitoring on wearable devices. IEEE Transactions on Biomedical Circuits and Systems, 15(6), 1196-1209.
[31] Song, S. B., Nam, J. W., & Kim, J. H. (2021). NAS-PPG: PPG-based heart rate estimation using neural architecture search. IEEE Sensors Journal, 21(13), 14941-14949.
[32] Ngoc-Thang, B., Nguyen, T. M. T., Truong, T. T., Nguyen, B. L. H., & Nguyen, T. T. (2022). A dynamic reconfigurable wearable device to acquire high quality PPG signal and robust heart rate estimate based on deep learning algorithm for smart healthcare system. Biosensors and Bioelectronics: X, 12, 100223.
[33] Lee, E., & Lee, C. Y. (2021). PPG-based smart wearable device with energy-efficient computing for mobile healthcare applications. IEEE Sensors Journal, 21(12), 13564-13573.
[34] V.V.Moca, H. Bârzan, A.Nagy-Dăbâcan, & R. C. Mureșan, Time-frequency super-resolution with superlets". *Nature Communications*, Vol. *12*, pp.1-18, Jan. 2021.
[35] Pankaj, A. Kumar, R. Komaragiri, and M. Kumar. "STSR: Spectro-Temporal Super-Resolution Analysis of a Reference Signal Less Photoplethysmogram for Heart Rate Estimation During Physical Activity," IEEE Transactions on Instrumentation and Measurement, 71, 1-10, July 2022.
[36] M. A. Motin, C. K. Karmakar, and M. Palaniswami, "PPG Derived Heart Rate Estimation during Intensive Physical Exercise," *IEEE Access*, vol. 7, pp. 56062–56069, May 2019.
[37] S. A, V. Ravichandran, P. SP, J. Joseph, and M. Sivaprakasam, "PPGnet: Deep Network for Device Independent Heart Rate Estimation from Photoplethysmogram," *arXiv Computer Science*, pp. 1899–1902, March 2021.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:45:00 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

J Mater Sci: Mater Electron **(2023) 34:1999**

Effect of sensitizer on the luminescence of thermally stable Eu^{3+}-activated metasilicate phosphor for solar cell applications

Deepali[1], and M. Jayasimhadri[1,*]

[1] *Luminescent Materials Research Lab, Department of Applied Physics, Delhi Technological University, Bawana Road, Delhi 110 042, India*

Received: 28 May 2023
Accepted: 1 October 2023
Published online: 16 October 2023

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

ABSTRACT

Solid-state reaction methodology has been opted to synthesize microcrystalline pure phase Sm^{3+}/Eu^{3+} co-doped $Na_4Ca_4Si_6O_{18}$ (NCMS) phosphors. Diffraction peaks of synthesized samples have shown the well-identical profile with the standard diffraction pattern and no impurity has been found in the sample. The particle size and morphology of the samples have been captured using field emission scanning electron microscope. Photoluminescence studies of Sm^{3+} activated and Sm^{3+}/Eu^{3+} co-activated NCMS phosphor have been thoroughly investigated and discussed. The energy transfer mechanism in NCMS: Sm^{3+}/Eu^{3+} phosphors have been studied using PL and lifetime data to facilitate the emission in red region under n-UV excitation. Dexter's and Reisfeld's approximation reckoned the type of interaction taking place among dopant and co-dopant ions. The chromaticity coordinates of NCMS: Sm^{3+}/Eu^{3+} phosphors show the color tunability from orange to red region excited with n-UV/blue light as the concentration of activator ion increases. Moreover, improved luminescence and superior thermal stability of Sm^{3+}/Eu^{3+} co-activated NCMS phosphor demonstrates the viability of the phosphor to be utilized in solar cell applications.

1 Introduction

In the modern era, due to overwhelming global population growth and their constant energy consumption, there has been an intense upsurge in the demand of energy. Solar energy has been ascended as a leading renewable energy source to accomplish the massive demand through amalgamating this energy source into technology [1, 2]. The extensive and incessant availability of solar energy have drawn the attention of the researchers and scientists to develop the device with capability of converting solar energy into electrical energy and initiated the step towards bridging the gap between nature and technology [3, 4]. Recently, the silicon based solar cells are most widely used at industrial horizon due to their benefits such as high stability, ease of massive fabrication and environmental friendliness [5]. However, the commercially available silicon-based solar cells utilize merely a small proportion of solar spectrum indeed have minimal propensity to

Address correspondence to E-mail: jayaphysics@yahoo.com

https://doi.org/10.1007/s10854-023-11413-7

convert solar energy into electrical energy. The theoretically calculated conversion efficiency of silicon-based solar cell has been reported with the band gap of 1.1 eV as 30%, whereas remaining 70% has been dissipated due to the solar cell response that does not match with the solar spectrum which is widely known as spectral mismatch [6, 7]. In general, spectral mismatch caused due to two substantial losses: predominant loss (i.e. transmittance loss) arises when the photon with lower energy transmitted through the solar cell without generating electron-hole pair as commercially available silicon has higher band gap energy. Moreover, secondary or thermalization loss occur due to relaxation of charge carriers in the form of thermal energy that absorbed by high-energy photons and leads to declining the conversion efficiency of solar cell [8–10]. Thus, use of the luminescent materials could be a viable approach to reduce the spectral mismatch caused by losses and increase the conversion efficiency of solar cells. The luminescent materials especially downconversion/downshifting (DC/DS) materials, which absorbs energy in UV/n-UV region and releases energy in the visible region (without heat dissipation) have been discovered to utilize as a solar spectral convertor for photovoltaic applications [11]. However, these luminescent materials or phosphors could be efficiently used in numerous other applications such as display devices, optical sensors, security ink, solid state lighting (w-LED) and bioimaging etc. [12].

Trivalent rare-earth (RE) ions listed as Eu^{3+}, Ce^{3+}, Ho^{3+}, Sm^{3+}, Tm^{3+}, Tb^{3+} and Dy^{3+} etc. doped into various host matrices mentioned as silicates, aluminates, phosphates, tungstates and germanates etc., have shown remarkable performance in terms of photovoltaic and w-LED applications [13]. It is a crucial to choose suitable host matrix and doped with appropriate RE ions for better luminescence results. Therefore, selecting the metasilicate as a host matrix seems to be a wise decision as it shows magnificent thermal and chemical stability with ease of synthesis under ambient condition. Singly rare earth ions doped phosphor is excited with UV/n-UV light, reflects an intense single color emission. For instance, Eu^{3+} ion-doped materials may be employed as a red component in the generation of white light with improved CRI and CCT values [14]. However, bi-activated phosphors illustrate the transfer of energy occur from sensitizer (donor) to activator (acceptor), which may extend excitation levels to improve the luminescence of the phosphor materials. Sm^{3+}/Eu^{3+} codoped zinc phosphate barium titanate (ZPBT) glasses have illustrated the enhanced luminescence intensity of Eu^{3+} ion via incorporating the Sm^{3+} as a sensitizer [15]. Thus, double ion doped or co-doped phosphors are quiet approachable practically in solar cell applications to achieve excellent results.

In the current report, a series of microcrystalline Sm^{3+}/Eu^{3+} co-activated $Na_4Ca_4Si_6O_{18}$ (NCMS) phosphors have been prepared through conventional synthesis methodology (i.e. solid state reaction route) and delved their structural, morphological, optical and photoluminescence behaviour. The emission intensity through activator ion has been enhanced by the incorporation of sensitizer ions and investigated in detail using photoluminescence spectra. Furthermore, the type of interaction among the dopant and co-dopant ions can be revealed by employing the Dexter's and Reisfeld's approximation. The admirable thermal stability of the material has been acknowledged using temperature dependent photoluminescence spectra. The decay analysis and CIE chromaticity coordinates demonstrate the energy transfer as well as color tunability behaviour of as-prepared NCMS phosphor. The above stated research outcomes from Sm^{3+}/Eu^{3+} co-activated NCMS phosphor showcase its potential to utilise in solar cell application.

2 Materials and methods

A series of Sm^{3+} activated and Eu^{3+} co-activated NCMS phosphors have been synthesized by well-established solid state reaction route. Precursors viz. Na_2SiO_3, CaO, SiO_2, Sm_2O_3 and Eu_2O_3 with high purity have been used. All the precursors collectively weighed in stoichiometric amount, dispersed in an acetone and ground for an hour. The crushed white powder was loaded in an alumina crucible, which further placed in a furnace for calcination at 650 °C for an hour then increased the temperature up to 950 °C and hold for 3 h [20]. Since, the calcination temperature of host matrix has been optimised thereafter subsequent series of NCMS with dopant (Sm^{3+}) and co-dopant (Eu^{3+}) ions have been synthesized by keeping regular ambient conditions. Furthermore, the incorporated concentration of Sm^{3+} as dopant ion (i.e. 1.0, 2.0, 3.0, 4.0 and 5.0 mol%) in NCMS host matrix are named as NS1 to NS5. The Eu^{3+} as co-dopant ions have been added to the NS3 and are identified as NSE1, NSE2, NSE4,

NSE6 and NSE8 for 1.0, 2.0, 4.0, 6.0 and 8.0 mol% concentration of Eu^{3+} ion.

The crystallinity and phase purity of host and bi-activated NCMS phosphor have been recorded using Smartlab high resolution X-ray diffractometer from Rigaku. To estimate the optical band gap, diffuse reflectance spectra has been monitored using UV–Visible/NIR spectrophotometer from Jasco (Model no: V-770). FE-SEM has been executed using NOVA Nano SEM450 to capture the size and morphology of powder samples. The photoluminescence properties have been investigated using JASCO spectrofluorometer (FP-8300 series). The decay time has been accomplished with FL920 Edinburgh spectrofluorophotomer instrument using pulsed Xenon lamp. The temperature dependent photoluminescence has been carried out using detachable optical fibre based Ocean optics (FLAME-S-XR1-ES) spectrophotometer.

3 Results and discussion

3.1 Structural, phase purity and strain measurement studies

Figure 1a depicts the X-ray diffraction (XRD) patterns of undoped, Sm^{3+} doped and Sm^{3+}/Eu^{3+} co-doped NCMS materials. Diffraction peaks of all the samples have shown the good agreement to the standard diffraction peaks of $Na_4Ca_4Si_6O_{18}$ (JCPDS card no: 75-1687) materials that signified the absence of any impurity and confirmed the preparation of single phase crystalline hexagonal NCMS host matrix. The as-prepared NCMS crystal with pure hexagonal phase belongs to $P3_2$ space group and have unit cell parameters as $a = b = 10.46$ Å; $c = 13.17$ Å and $V = 1249.42$ $Å^3$. However, the insertion of dopant (Sm^{3+}) and co-dopant (Eu^{3+}) ions into the host lattice may supplement the photoluminescence intensity without changing the elementary crystal structure. Since the XRD profile of Sm^{3+} activated and Sm^{3+}/Eu^{3+} co-activated NCMS phosphors have shown the well-matched diffraction peaks with the standard and undoped pattern of NCMS structure indicate the effective incorporation of dopant and co-dopant ions into the host lattice. Moreover, as the ionic radii of Sm^{3+} (0.96 Å) and Eu^{3+} (0.95 Å) having coordination number six is quite closer to the ionic radii of Ca^{2+} (1.00 Å) and can be easily replaced with these two dopant and co-dopant ions in the NCMS crystal framework. Furthermore, the average crystallite size (D_{hkl}) can be evaluated using illustrious Debye-Scherrer equation given below [16]:

$$D_{hkl} = \frac{k\lambda}{\beta_{hkl}\mathrm{Cos}\theta_{hkl}} \quad (1)$$

each parameter in the above equation such as k, λ, β and θ are designated. k is constant with value 0.94, λ is wavelength of incident X-rays with the value of 1.54 Å, β is full width at half maxima (FWHM) and θ is diffraction angle. Using above equation, the average

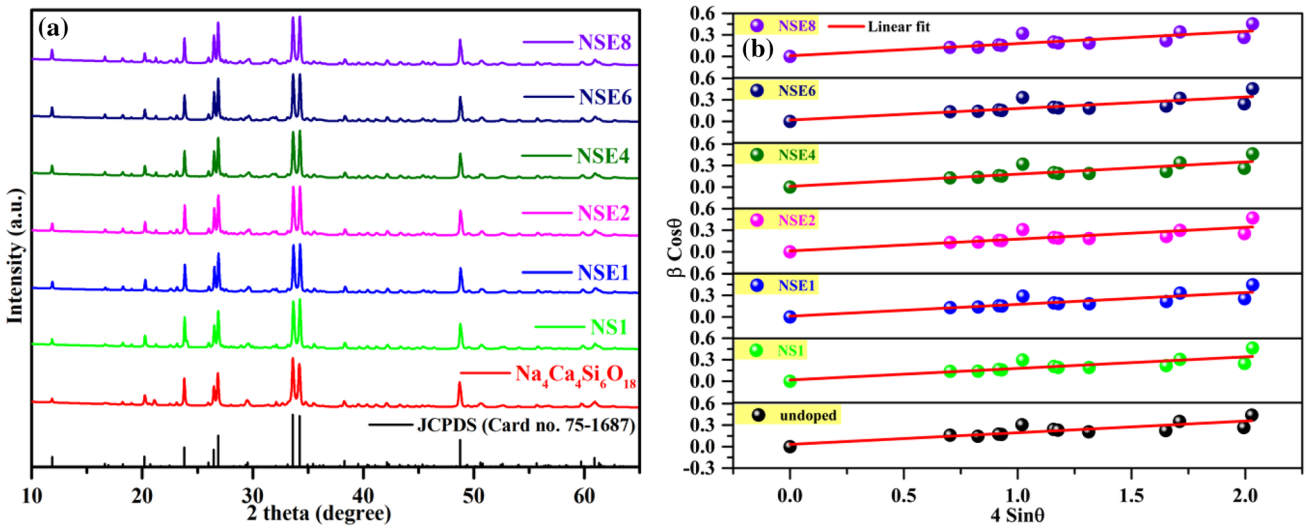


Fig. 1 **a** X-ray diffraction patterns and **b** Williamson–Hall (W–H) of undoped, Sm^{3+} (1.0 mol%) and Sm^{3+}/Eu^{3+} (1.0, 2.0, 4.0, 6.0 and 8.0 mol% of Eu^{3+} ions) co-activated $Na_4Ca_4Si_6O_{18}$ phosphor

crystallite size are computed for all the as-synthesized samples and varies in a range of 44–49 nm [17]. The average crystallite size (D) and strain (ϵ) can also be estimated using Williamson–Hall (W–H) plot equation given below [18]:

$$\beta \mathrm{Cos}\theta = \frac{k\lambda}{D_{\mathrm{hkl}}} + 4\epsilon\ \mathrm{Sin}\theta, \tag{2}$$

where most of the parameters such as D, k, λ, β and θ have already defined earlier and ϵ is the strain. The strain has been estimated using slope and intercept value after plotting a graph between $\beta\mathrm{Cos}\theta$ and $4\mathrm{Sin}\theta$, as represented in Fig. 1b. Using W–H plot, strain for the samples would lie in a range of 0.163–1.710 [19]. Moreover, the positive values of the slope reveal the tensile strain which also elevate with rise in the concentration of dopant and co-dopant ions in the samples.

3.2 Band gap measurement

To compute the optical band gap, diffuse reflectance spectra of NCMS: $x\mathrm{Sm}^{3+}/y\mathrm{Eu}^{3+}$ (x = 3.0 mol% & y = 1.0 and 8.0 mol%) phosphors have been measured in UV, n-UV and visible regions (200–600 nm), as demonstrated in Fig. 2a and b. The spectra show few characteristic reflectance peaks are appeared in n-UV and visible region due to f–f transition among dopant and codopant ions [20]. Moreover, the band gap (E_g) energy have been reckoned using Kubelka–Munk (K–M) relation as represented in Eq. 3 [21]:

$$F(R) = \frac{(1-R)^2}{2R} = \frac{K}{S}, \tag{3}$$

where $F(R)$ refers to K–M function, R (reflectance) tends to the reflectance of sample to reference sample (Standard: $BaSO_4$) ratio, K and S symbolizes the absorption and scattering coefficient. Further, by equating Tauc equation (Eq. 4) with K–M relation provides the optical band gap (E_g) energy by following the value of n which elaborate the kind the electronic transitions through equation below [22]:

$$\alpha h\nu = A(h\nu - E_g)^n \tag{4}$$

$$(F(R)h\nu) = B(h\nu - E_g)^n \tag{5}$$

Thus, the band gap energy for NCMS: $x\mathrm{Sm}^{3+}/y\mathrm{Eu}^{3+}$ (x = 3.0 mol% & y = 1.0 and 8.0 mol%) phosphor have been calculated by extrapolating the line to x-axis in $[F(R)h\nu]^2$ vs. $h\nu$ plot, exposed in the inserted plot of Fig. 2a and b. The direct allowed band gap with n = 1/2 for NSE1 and NSE8 have been computed and found to be 4.41 and 4.40 eV, respectively [23].

3.3 Morphological studies

The shape, size and morphology of the undoped $Na_4Ca_4Si_6O_{18}$ (Fig. 3a–c) and $Na_4Ca_4Si_6O_{18}$: $3.0\mathrm{Sm}^{3+}/1.0\mathrm{Eu}^{3+}$ (Fig. 3d–f) phosphors have been unveiled by field emission scanning electron microscopy (FE-SEM) technique. FE-SEM micrographs divulge the irregular and agglomerated particles in micron range which was observed due to high calcination temperature and taken at different resolutions, as depicted in Fig. 3a–f [24]. Moreover, energy-dispersive X-ray analysis (EDAX) has also been done along with FE-SEM, which express the presence of elements with weight and atomic percentage in the material, as represented in Fig. 3g and h.

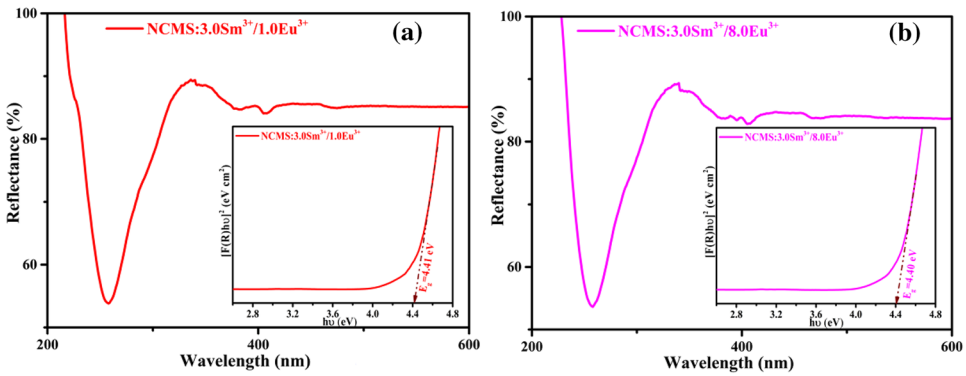
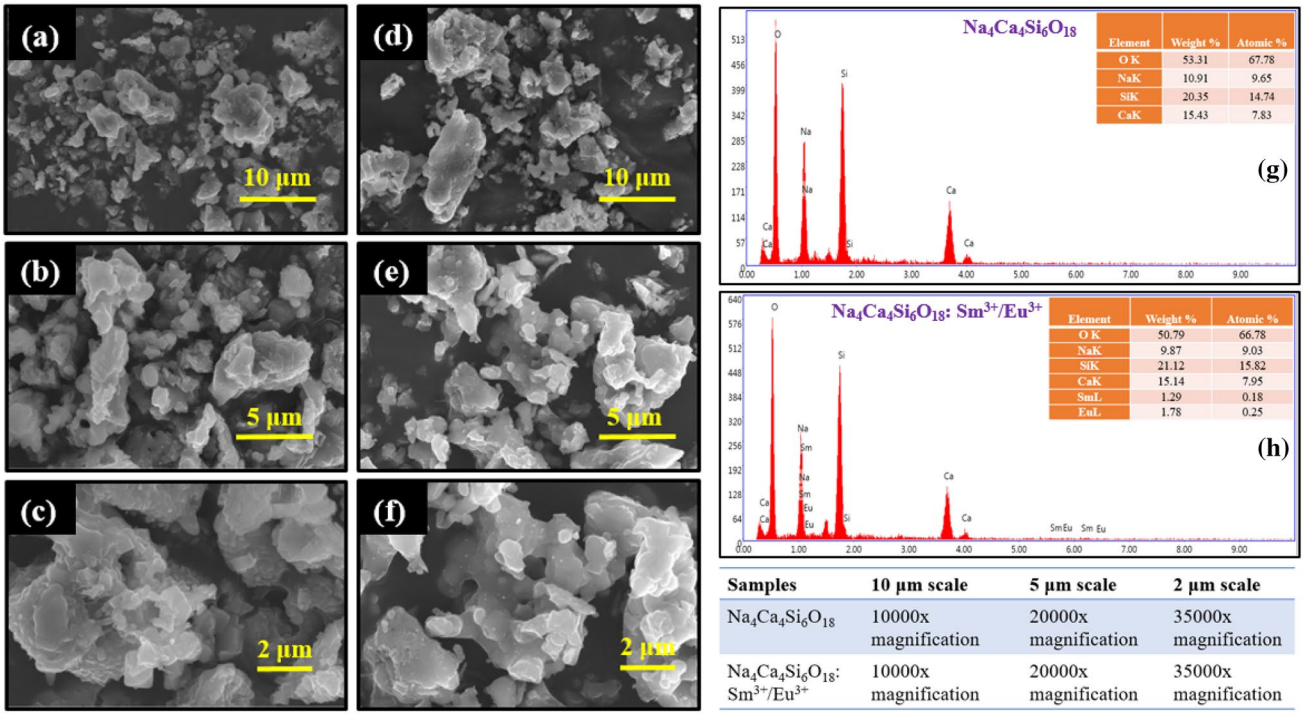


Fig. 2 **a**, **b** Diffuse reflectance spectra of Sm^{3+}/Eu^{3+} co-activated $Na_4Ca_4Si_6O_{18}$ phosphors and inset shows the band gap plot of associated NCMS: Sm^{3+}/Eu^{3+} phosphors



Samples	10 µm scale	5 µm scale	2 µm scale
$Na_4Ca_4Si_6O_{18}$	10000x magnification	20000x magnification	35000x magnification
$Na_4Ca_4Si_6O_{18}$: Sm^{3+}/Eu^{3+}	10000x magnification	20000x magnification	35000x magnification

Fig. 3 **a–f** SEM micrograph of undoped and NCMS: 3.0 Sm^{3+}/1.0Eu^{3+} phosphor at different resolution mentioned in the table shown above. **g**, **h** Energy dispersive X-ray analysis (EDAX) of undoped and NCMS: 3.0 Sm^{3+}/1.0Eu^{3+} phosphor

3.4 Luminescent studies of NCMS: Sm^{3+}/Eu^{3+} phosphor

The photoluminescence (excitation and emission) spectra of Sm^{3+} doped $Na_4Ca_4Si_6O_{18}$ phosphors were captured in 300–500 nm excitation and 500–750 nm emission range, as represented in Fig. 4a and b, respectively. The recorded excitation spectra express multiple peaks at fixed emission wavelength of 602 nm. These multiple peaks were observed in excitation range having electronic transition from lower state $^6H_{5/2}$ to various excited states as $^4P_{1/2}$ (315 nm), $^4H_{9/2}$ (344 nm), $^4D_{3/2}$ (361 nm), $^6P_{7/2}$ (375 nm), $^4G_{11/2}$ (388 nm), $^6P_{3/2}$ (403 nm), $^4M_{19/2}$ (416 nm), $^4G_{9/2}$ (441 nm), $^4M_{17/2}$ (450 nm), $^4I_{13/2}$ (467 nm) and $^4I_{9/2}$ (488 nm) [25]. Among all the excitation peaks, the highest intense peak is observed at 403 nm which attributes to $^6H_{5/2} \rightarrow {}^6P_{3/2}$ and have been chosen to record the emission spectra. Moreover, the monitored emission spectra affirm peak position at 564, 602, 650 and 714 nm implies to the transitions of Sm^{3+} ions from $^4G_{5/2}$ to $^6H_{5/2}$, $^6H_{7/2}$, $^6H_{9/2}$ and $^6H_{11/2}$, respectively. In addition to the intense emission peak at 602 nm, two less strong peaks identified at 564 and 650 nm obeys selection rule and ensues the magnetic dipole (MD) ($\Delta J = 0, \pm 1$) and electric dipole (ED) ($\Delta J \leq 6$; $\Delta J = 2$, 4 and 6) transitions, respectively, where J is the total angular momentum [26]. The MD transition resembles to $^4G_{5/2} \rightarrow {}^6H_{5/2}$, and does not effect from crystal field environment wherein hypersensitive forced ED transition attributed to $^4G_{5/2} \rightarrow {}^6H_{9/2}$ is highly influenced with crystal field environment. Most effectively intense peak observed at 602 nm ($^4G_{5/2} \rightarrow {}^6H_{7/2}$) due to both partially MD and ED transitions by following the selection rule (where, $\Delta J = \pm 1$) [27, 28]. The information regarding the symmetric and asymmetric nature of local environment of the dopant ions in the host matrices is represented by the ratio of ED to MD transitions [29]. Lower intensity ratio indicates lesser distortion from the inversion symmetry. Hence, the estimated value for intensity ratio of ED to MD transition is 0.81 for optimized concentration and designates lesser distortion from the inversion symmetry or higher environmental symmetry [30]. Furthermore, the emission intensity for distinct Sm^{3+} ion concentrations incorporated into the host matrix increases up to 3.0 mol%, however, further increase in concentration leads to decrease in emission by the cause of concentration quenching mechanism, as represented in Fig. 4c [31, 32].

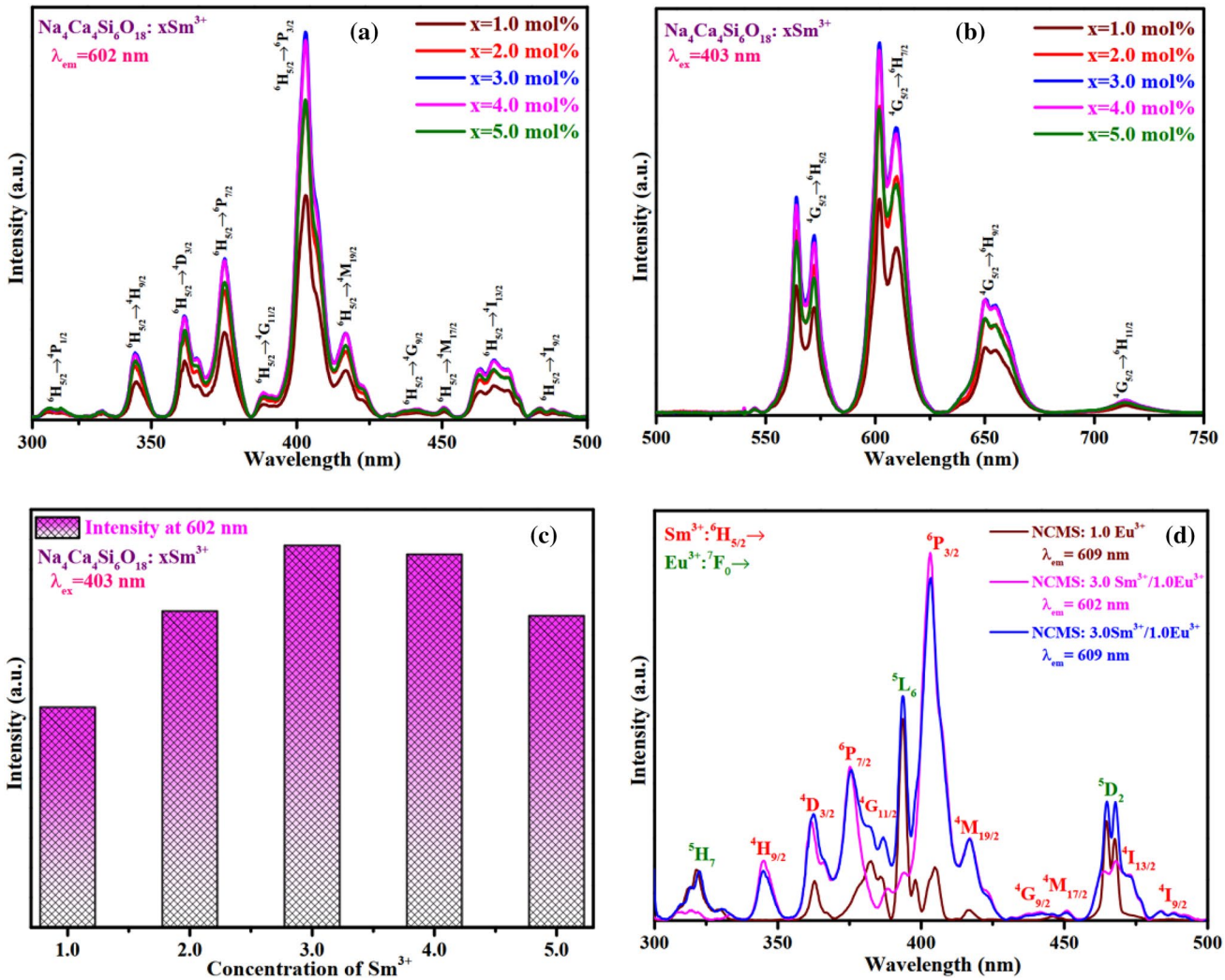


Fig. 4 **a**, **b** PLE and PL spectra of Sm^{3+} (1.0, 2.0, 3.0, 4.0 and 5.0 mol%) doped $Na_4Ca_4Si_6O_{18}$ phosphors monitored with the $\lambda_{em}=602$ nm and $\lambda_{ex}=403$ nm, respectively. **c** Bar plot for intensity versus concentration of Sm^{3+} ions. **d** Compiled excitation plot of 1.0 mol% of Eu^{3+} doped & 3.0 Sm^{3+}/1.0 Eu^{3+} co-doped NCMS phosphor under $\lambda_{em}=609$ nm and $\lambda_{em}=602$ nm

Initially, the optimization of concentration of dopant (Sm^{3+}) ions in the as-prepared host matrix has been done to act as a sensitizer. Furthermore, the Eu^{3+} ions (activators) added in the optimized sample in order to enhance the intensity. As a result, sensitizer has been proved to be essential in boosting the intensity of activator ion as shown in the excitation plot of Sm^{3+} and Sm^{3+}/Eu^{3+} co-activated NCMS phosphor in Fig. 4d [33]. The excitation spectra of 1.0 mol% of Eu^{3+} ion activated and 3.0 mol% of Sm^{3+} doped with 1.0 mol% of Eu^{3+} co-activated NCMS phosphor have been taken under 602 and 609 nm emission wavelength. Under both emission wavelengths several peaks are obtained due to Sm^{3+} and Eu^{3+} dopant ions located at 344, 361, 375, 388, 403, 416, 441, 450, 467 and 488 nm wavelength attributed to distinct levels from $^6H_{5/2}$ to $^4H_{9/2}$, $^4D_{3/2}$, $^6P_{7/2}$, $^4G_{11/2}$, $^6P_{3/2}$, $^4M_{19/2}$, $^4G_{9/2}$, $^4M_{17/2}$, $^4I_{13/2}$, $^4I_{9/2}$ (for Sm^{3+} ions) and 321, 393 and 464 nm resembles from $^7F_0$ to $^5H_7$, $^5L_6$, $^5D_2$ state (for Eu^{3+} ions) [34]. It can be espied that the augmented excitation peaks of Eu^{3+} ions at 393 ($^7F_0 \rightarrow {}^5L_6$) and 467 ($^7F_0 \rightarrow {}^5D_2$) nm have been achieved due to the insertion of Sm^{3+} ions whose sensitization effect has emerged onto the Eu^{3+} in Sm^{3+}/Eu^{3+} co-doped NCMS host matrices, as represented in Fig. 4d.

Figure 5a and b expresses the excitation spectra of Sm^{3+} activated and Eu^{3+} co-activated $Na_4Ca_4Si_6O_{18}$ phosphor at intense emission wavelength of Sm^{3+}

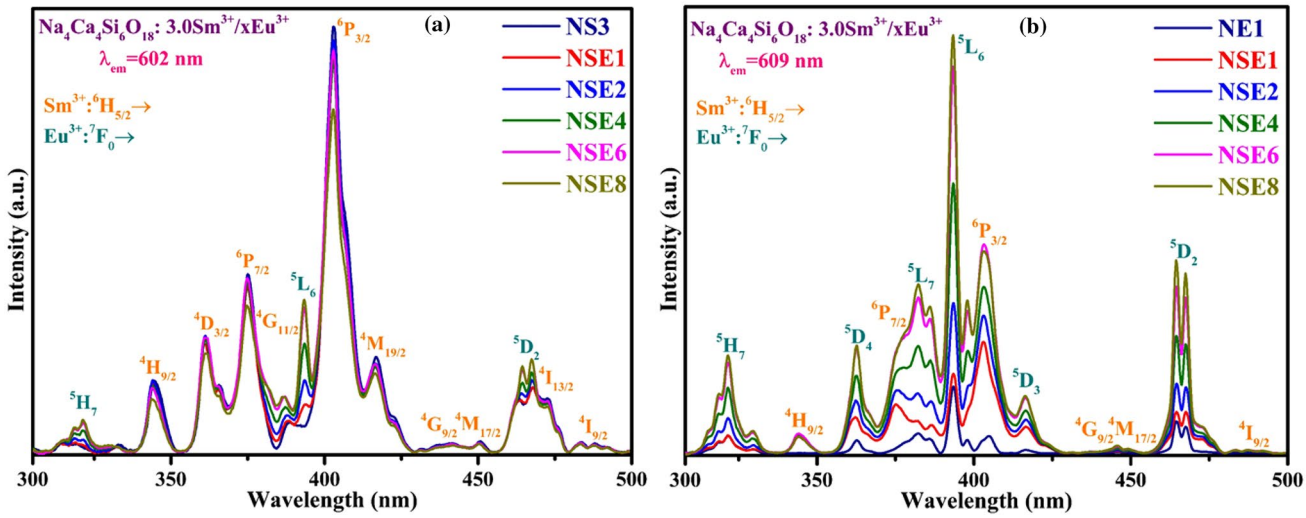


Fig. 5 **a, b** Excitation spectra of Sm^{3+}, Eu^{3+} activated and Sm^{3+}/Eu^{3+} co-activated NCMS phosphor by monitoring the emission wavelength of $\lambda_{em} = 602$ and 609 nm

(602 nm) and Eu^{3+} (609 nm) ions in order to elucidate the behaviour of peaks with incorporation of Eu^{3+} as a co-dopant. In the excitation spectra under 602 and 609 nm, mixed peaks of Sm^{3+} and Eu^{3+} ions are observed in a range of 300–500 nm, as discussed earlier in Fig. 4d. Considering the intense peaks enumerated at 375, 403, 467 and 393 nm in the excitation spectra in Fig. 5a and b, the emission profiles have been monitored in the visible spectral region as catalogued in Fig. 6a–d. The emission spectra recorded under 375 and 403 nm excitation wavelengths are quite similar as both the excitation wavelengths are of pure Sm^{3+} ions. Emission peaks positioned at 564, 602, 650 and 714 nm are attained due to the presence of Sm^{3+} ion acts as sensitizer in the host matrix and their intensity drops with simultaneous rise in the Eu^{3+} ions concentration. However, peaks obtained at 579, 591, 609, 653, 686 and 705 nm are due to insertion of Eu^{3+} as an activator ion and upsurges as followed with the concentration of Eu^{3+} ions. The peaks observed in a range of 635–675 nm due to Sm^{3+} ions are showing the random intensity behaviour and unable to distinguished which is because of the combined effect of Sm^{3+} and Eu^{3+} ions. Figure 6c depicts the emission spectra of NCMS: Sm^{3+}/Eu^{3+} phosphor at 467 nm excitation wavelength, which is second utmost intense as well as nearly common peaks emerged in the excitation range. It can be observed that the Sm^{3+} ion peaks are dropping down wherein the peaks position obtained due to Eu^{3+} ions are rising with its concentration. Figure 6d manifests that the emission spectra under 393 nm excitation wavelength for NCMS: Sm^{3+}/Eu^{3+} phosphor wherein the emission peaks of Eu^{3+} ions are rising with increase in the Eu^{3+} ions concentration. Moreover, inset of Fig. 6d illustrates the emission peaks of NSE1 is elevated in comparison with NE1 phosphor indicated that the Sm^{3+} ion has aided to improve the emission intensity. It can extricate from the graph that few peaks positioned at 564 and 602 nm are appeared in emission spectra of NSE1 phosphor which signifies the presence of Sm^{3+} ions due to the fact that the excitation energy of Sm^{3+} and Eu^{3+} is nearby, which further vanishes with increase in Eu^{3+} ion concentration [35].

Dexter and Schulman express the ubiquitous phenomenon of concentration quenching effect during energy transfer from sensitizers (Sm^{3+}) to activators (Eu^{3+}) [36]. In general, there are two viable mechanisms by which energy can be transferred among rare earths, namely exchange and multipolar interaction. Predominantly, the exchange interaction appears if the critical distance (R_c) is smaller than the 5 Å, whereas the multipolar interaction occurs when the R_c is larger than 5 Å. According to Blasse, the critical distance (R_c) between sensitizer and activator ions can be evaluated using below equation [37]:

$$R_c = 2\left(\frac{3V}{4\pi x_c N}\right)^{\frac{1}{3}} \quad (6)$$

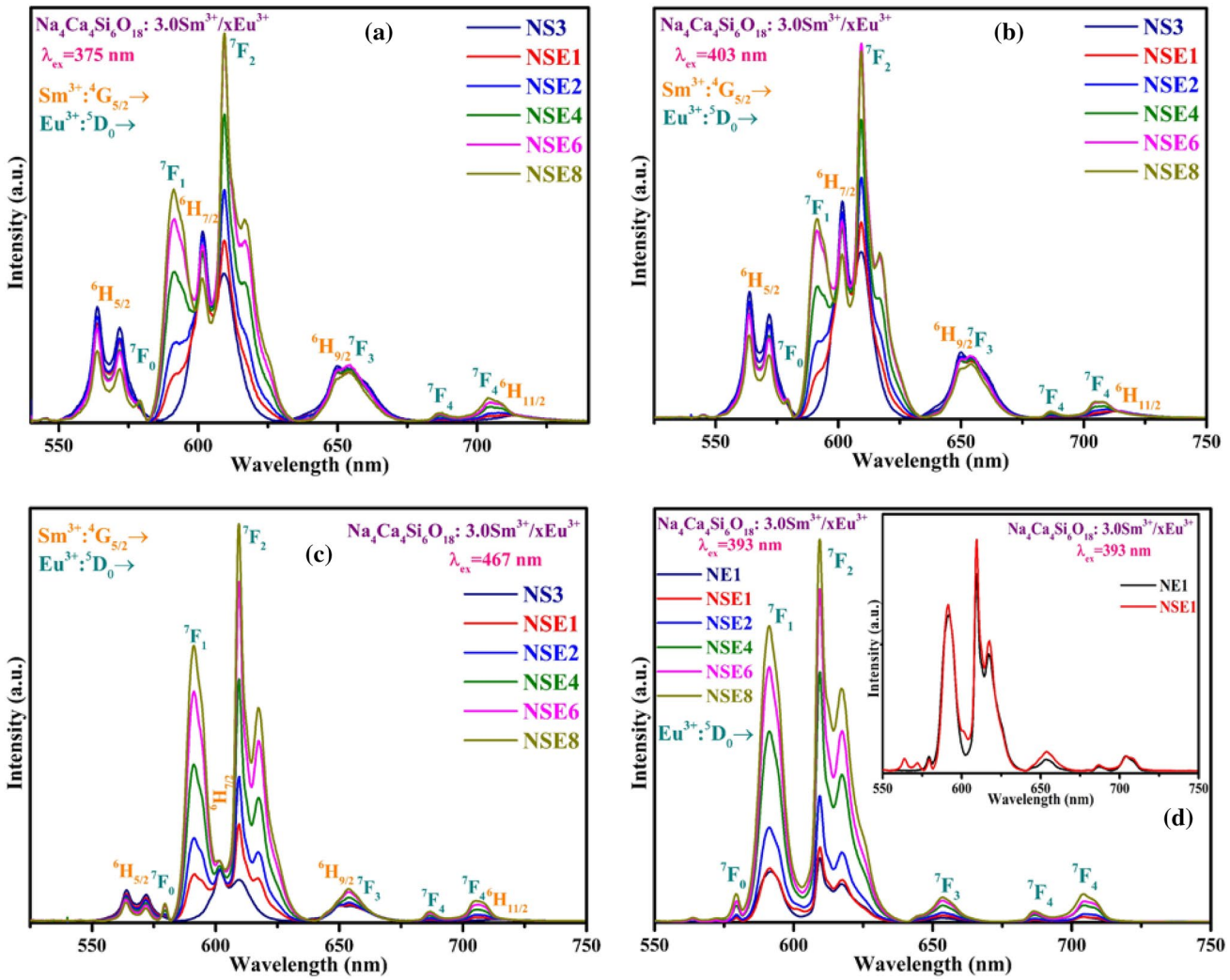


Fig. 6 **a** PL spectra of Sm^{3+} activated and Sm^{3+}/Eu^{3+} co-activated NCMS phosphor under the excitation wavelengths $\lambda_{ex}=375$ nm **b** $\lambda_{ex}=403$ nm **c** $\lambda_{ex}=467$ nm and **d** $\lambda_{ex}=393$ nm with inset representing the comparison of emission spectra of 1.0 Eu^{3+} with $3.0Sm^{3+}/1.0Eu^{3+}$ co-doped NCMS phosphor

In the above-mentioned equation, various significant parameters such as the critical concentration (x_c) of sensitizer (Sm^{3+}) and activator (Eu^{3+}), volume (V) and number of cations in the unit cell (N). For $Na_4Ca_4Si_6O_{18}$ matrix, the value of N is 3 and V is 1249.42 $Å^3$ inclusively, the critical concentration is taken where the emission intensity of sensitizer is approaches to half of their own intensity i.e. 0.11 mol by taking Sm^{3+} as fixed at 3.0 mol% and Eu^{3+} as 8.0 mol%. Finally, the critical distance is evaluated as 19.34 Å, which is higher than 5 Å, prone to allow multipolar interactions that includes dipole-dipole, dipole-quadrupole and quadrupole-quadrupole (i.e. d–d, d–q and q–q) interactions [38]. Thus, in order to inspect the sort of interaction in the Sm^{3+}/Eu^{3+} co-doped NCMS phosphor sample, Dexter's and Reisfeld's approximation are employed according to the below equation [39]:

$$\frac{\eta_{s0}}{\eta_s} \alpha\, C^{\frac{n}{3}}_{Sm^{3+}+\,Eu^{3+}} \tag{7}$$

where η_{s0} is the luminescent quantum efficiency of sensitizer (no activator), η_s is the luminescent quantum efficiency of activator ions and C signifies as the sum of concentration of sensitizer and activator ions with the power of $n/3$, where n represents the type of interaction taking place depending on the values 6 (d–d), 8 (d–q) and 10 (q–q). The η_{s0}/η_s ratio can be

equated to the ratio of associated luminescence intensity and represented as below [40]:

$$\frac{I_{s0}}{I_s} \alpha C_{Sm^{3+}+\ Eu^{3+}}^{\frac{n}{3}} \quad (8)$$

Similar to the Dexter's and Reisfeld's approximation, here I_{s0} and I_s is the intrinsic emission intensity of Sm^{3+} ion (without Eu^{3+} ion) and Sm^{3+} ion (with Eu^{3+} ion), respectively. Figure 7 depicts that the value of n obtained after best linearly fitted I_{s0}/I_s versus $C_{Sm^{3+}+Eu^{3+}}^{n/3}$ plot is 10 for bi-activated NCMS phosphor under 403 nm excitation. This signifies that the non-radiative energy transfer from donor to acceptor (sensitizer → activator) ions is caused by the quadrupole–quadrupole multipolar interaction [41].

The partial energy level diagram clearly illustrates the energy transfer mechanism among sensitizer and activator ions, as represented in Fig. 8. Initially the excited energy is absorbed by the Sm^{3+}/Eu^{3+} doped NCMS phosphor samples, which stimulates the ion from lower state to various excited states. Further, the ion attained to higher state later falls swiftly to the lower excited state through non-radiative relaxation represented by the dashed lines and then relaxes radiatively to the ground states. The lower excited state of Sm^{3+} and Eu^{3+}, i.e. $^4G_{5/2}$ and $^5D_0$ attribute to energy 17900 and 17286 cm^{-1}. The lower excited energy level

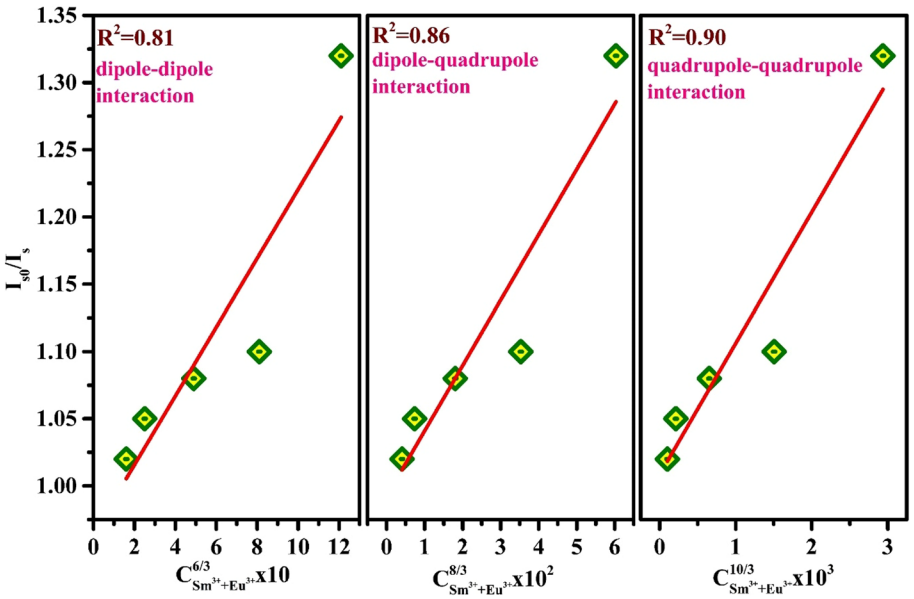


Fig. 7 I_{s0}/I_s versus $C_{Sm^{3+}+Eu^{3+}}^{n/3}$ plot revealing the types of interactions among dopant and co-dopants ions in bi-activated NCMS phosphor

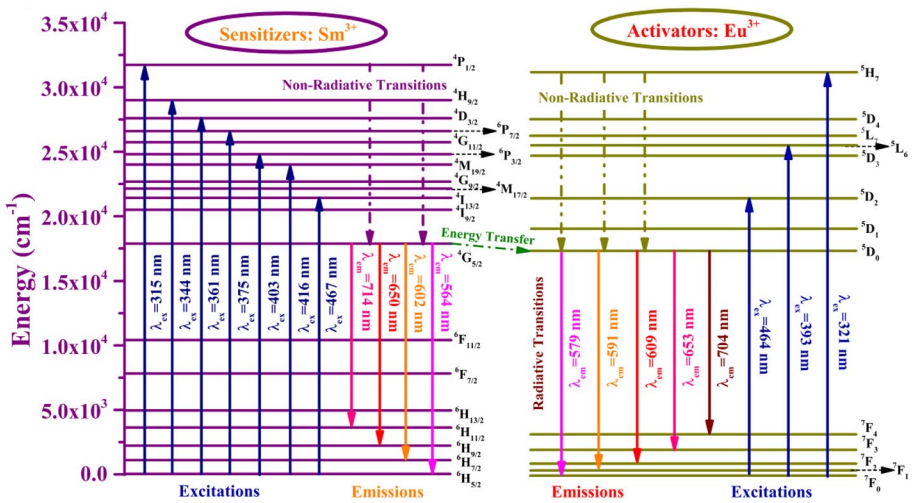


Fig. 8 Partial Energy level diagram representing the energy transfer process in Sm^{3+}-doped and Sm^{3+}/Eu^{3+} co-doped NCMS phosphor

of Sm^{3+} is nearly 614 cm^{-1} above than the lower excited level of Eu^{3+} ions. Hence, probably there may be a chance of transferring a part of energy from Sm^{3+} ions which leads to decline of emission obtained due to sensitizers, as represented in Fig. 8. Indeed, the minor distinction between the lower excited energy levels of sensitizer and activator ions allow the possible energy transfer from Sm^{3+} to Eu^{3+} ions. This transition acclaims that Sm^{3+} ions must behave as a sensitizer (donor) for activator (acceptor) ions in the current bi-activated NCMS crystal lattice [42].

3.5 Lifetime and energy transfer efficiency/ probability analysis

Decay analysis for sensitizer and activator ions in NCMS samples have been done under the excitation of 403 nm (n-UV), which attributes to the emission transition of $^4G_{5/2} \rightarrow {}^6H_{7/2}$ at 602 nm as shown in Fig. 9a–d. The decay profile of samples represents the declining trend with rise of Eu^{3+} ions concentration in the host matrix which implies the strong evidence of migration of partial energy from sensitizer to activator ions. The exponential decay curves were well fitted with bi-exponential equation after adopting different exponential equations. The optimized bi-exponential equation is as shown below [43]:

$$I = I_o + A_1 e^{\left(\frac{1}{\tau_1}\right)} + A_2 e^{\left(\frac{1}{\tau_2}\right)} \quad (9)$$

where I_o and I are denoted as the emission intensities at time t is 0 and t, respectively. The fitting constants are A_1 and A_2 and the lifetimes are represented as τ_1 and τ_2 which indicate slow and fast lifetimes. Furthermore, the average lifetime can be evaluated from bi-exponential fitted data as given below [44]:

$$\tau_{avg} = \frac{A_1\tau_1^2 + A_2\tau_2^2}{A_1\tau_2 + A_2\tau_2} \quad (10)$$

The evaluated average lifetime is in the range of 2.18–3.15 µs for undoped and co-doped phosphors expressed in Fig. 9a–d. The energy transfer efficiency and probability in terms of decay lifetime can be evaluated by the following equations [45]:

$$\eta_{T\left(Sm^{3+}\rightarrow Eu^{3+}\right)} = 1 - \left(\frac{\tau_d}{\tau_{d_o}}\right) \quad (11)$$

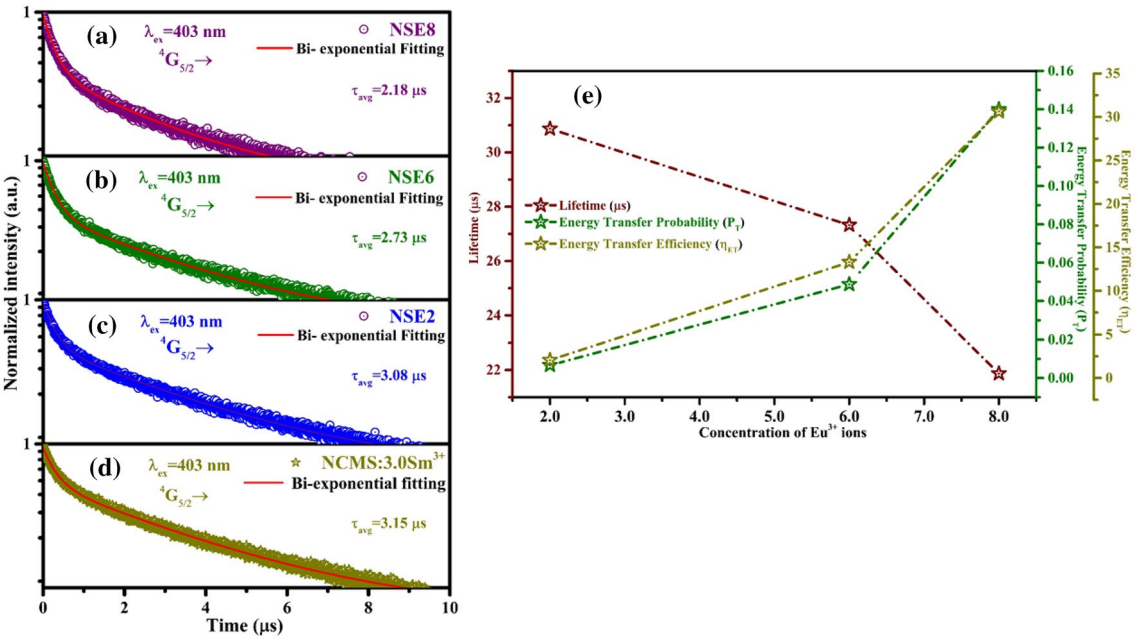


Fig. 9 **a–d** Lifetime decay plot of $3.0Sm^{3+}/xEu^{3+}$ (x = 2.0, 6.0 and 8.0 mol%) and 3.0 mol% of Sm^{3+}-doped NCMS phosphor for $^4G_{5/2}$ under 403 nm excitation wavelength. **e** Lifetime, energy transfer efficiency and energy transfer probability with increasing the concentration of Eu^{3+} ion at λ_{ex} = 403 nm and λ_{em} = 602 nm

$$P_{T\left(Sm^{3+}\rightarrow Eu^{3+}\right)} = \left(\frac{1}{\tau_d}\right) - \left(\frac{1}{\tau_{d_o}}\right) \quad (12)$$

The parameters in above equations such as τ_{d_o} and τ_d are referred as the average decay time of Sm^{3+} without Eu^{3+} and with Eu^{3+}, respectively. The estimated values of energy transfer efficiency in terms of percentage for NSE2, NSE4 and NSE8 is found to be 2.06, 13.30 and 30.61% whereas the values of energy transfer probability are 0.006, 0.048 and 0.139, respectively. The lifetime values, energy transfer efficiencies and energy transfer probability are comprehensively revealed in Fig. 9e. The assessed energy transfer efficiency and probability are increasing with rise in concentration of activator (Eu^{3+}) ions concentration, which signifies the transfer of energy from Sm^{3+} to Eu^{3+} ions in the host matrix.

3.6 CIE color coordinates

Commission Internationale de L'Eclairage (CIE) coordinates manifest luminescent color performance of the as-synthesized NCMS: Sm^{3+}/Eu^{3+} co-doped phosphors under distinct excitations, as represented in Fig. 10a–c. On the basis of emission spectra of Sm^{3+}/Eu^{3+} bi-activated NCMS phosphor (NS3, NSE1, NSE2, NSE4, NSE6 and NSE8) samples, the CIE color coordinates are determined under the 375, 403 and 467 nm excitation wavelength as summarized in Table 1. The chromaticity plot of bi-activated NCMS phosphors under 375, 403 and 467 nm excitation wavelength expresses the shift from orange to red luminescent region with the addition of Eu^{3+} ions concentration in the host lattice. In the NCMS: Sm^{3+}/Eu^{3+} bi-activated phosphor, the red

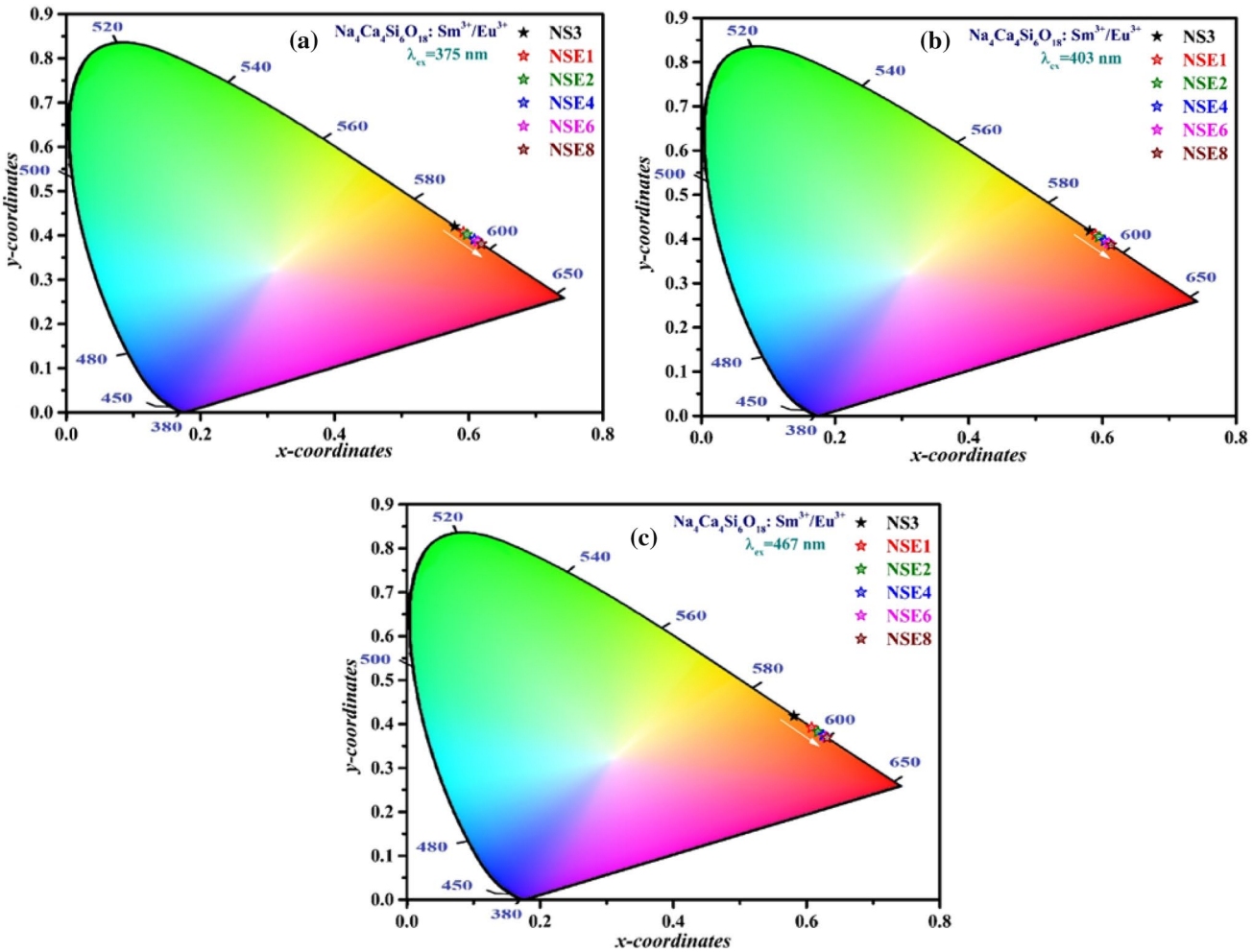


Fig. 10 **a** CIE chromaticity coordinates of Sm^{3+} doped NCMS phosphor (optimized) and Sm^{3+}/Eu^{3+} biactivated NCMS phosphor under excitation wavelengths $\lambda_{ex} = 375$ nm **b** $\lambda_{ex} = 403$ nm **c** $\lambda_{ex} = 467$ nm

Table 1 CIE chromaticity coordinates of singly doped NCMS: Sm^{3+} and doubly doped NCMS: Sm^{3+}/Eu^{3+} phosphors under different excitation wavelength

Sample	CIE chromaticity coordinates different excitation wavelength		
	403 nm	375 nm	467 nm
NS3	(0.580, 0.418)	(0.579, 0.420)	(0.581, 0.418)
NSE1	(0.589, 0.409)	(0.591, 0.407)	(0.607, 0.391)
NSE2	(0.594, 0.404)	(0.597, 0.401)	(0.617, 0.382)
NSE4	(0.603, 0.395)	(0.608, 0.390)	(0.625, 0.374)
NSE6	(0.609, 0.390)	(0.613, 0.386)	(0.628, 0.370)
NSE8	(0.613, 0.386)	(0.618, 0.380)	(0.630, 0.368)

shift under 375, 403 and 467 nm excitation wavelength occur due the appropriate energy transfer from Sm^{3+} to Eu^{3+} with increase in the concentration of Eu^{3+} ions. The CIE chromaticity coordinates for NSE8 phosphor found to be close to the available commercialized Y_2O_3: Eu^{3+} (0.647, 0.343) [46]. Furthermore, Fig. 11a demonstrates spectral distribution of sunlight (300–1200 nm) at air mass 1.5 Global (AM 1.5G) with the excitation and emission spectrum of NSE8 phosphor at emission wavelength of 602 nm and excitation wavelength at 403 nm. Since, the down-converting materials absorb the shorter wavelength (300–500 nm) and reemit into higher wavelength, thus down-converting NCMS:Sm^{3+}/Eu^{3+} phosphor may apply directly onto the front surface of solar cell to utilize the unused part of solar spectrum and can enhance efficiency of solar cell, as represented in Fig. 11b [47].

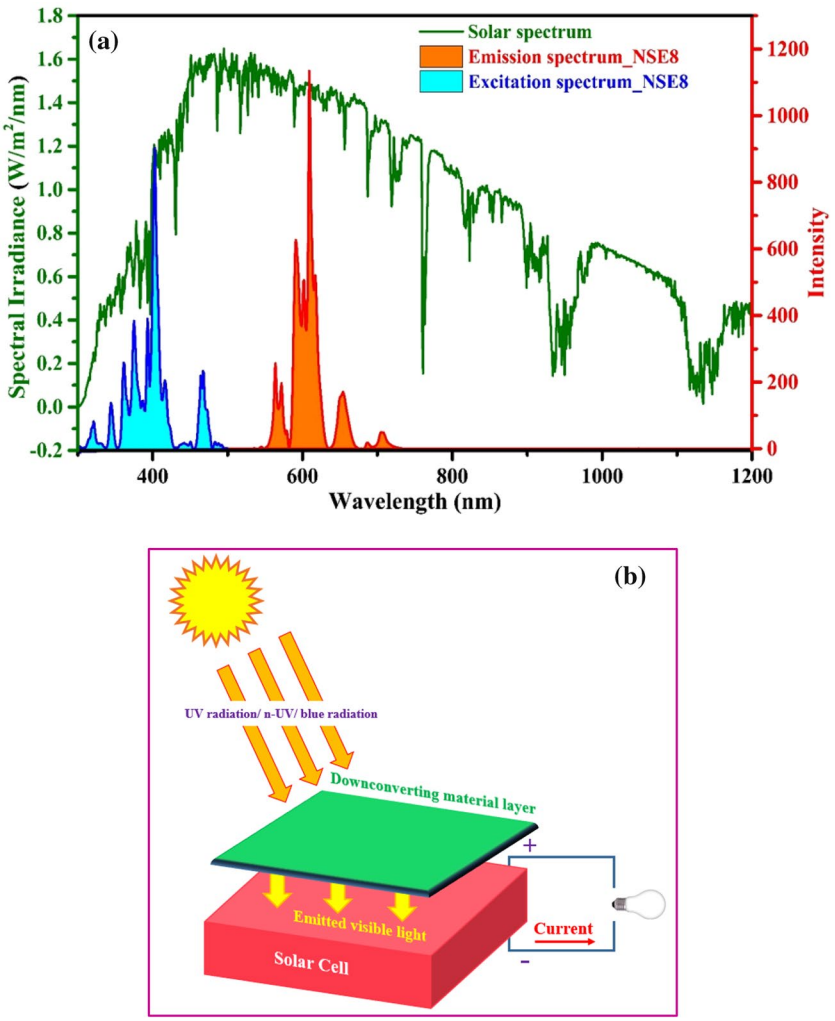


Fig. 11 **a** The comparative plot for AM 1.5 solar spectrum, excitation and emission spectrum of NSE8 under $\lambda_{em} = 602$ nm and $\lambda_{ex} = 403$ nm, respectively; **b** Schematic representation of solar cell with downconverting material

3.7 Thermal stability of NCMS: Sm^{3+}/Eu^{3+} phosphor

In order to fabricate devices with high power, exploration of thermal stability of the material is required. Figure 12a depicts the emission plot of NSE1 phosphor with respect to temperature (ranging from 303 to 463 K) measured under 403 nm excitation wavelength in a range of 550–750 nm. The emission intensity tends to decline with rise in temperature, and can be observed the emission intensity persist to 77.6% at 423 K (150 °C) of initial intensity obtained at room temperature (i.e. 30 °C) as shown in the graph inserted in Fig. 12a. Thus, the greater sustainability of integrated emission intensity at 423 K temperature indicate that the phosphor possesses excellent thermal stability [48]. Moreover, the activation energy (ΔE_a) of as synthesized NSE1 phosphor have been

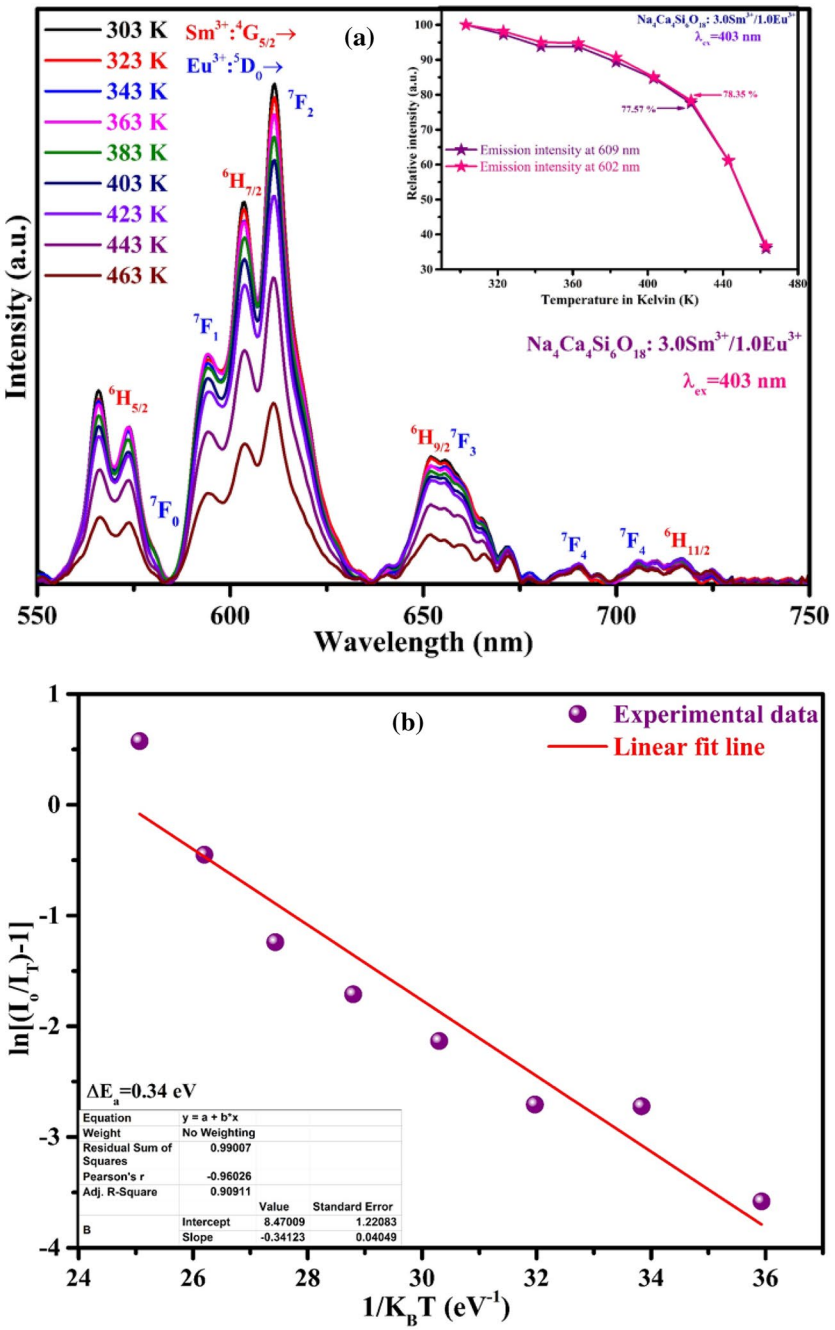


Fig. 12 **a** Temperature dependent PL spectra of NCMS: $3.0Sm^{3+}/1.0Eu^{3+}$ phosphor under 403 nm excitation wavelength; inset shows the relative intensity vs. temperature plot at emission wavelength 602 and 609 nm for NSE1 phosphor under 403 nm excitation wavelength **b** Plot of $\ln[(I_0/I_T)-1]$ vs. $1/K_BT$ (eV^{-1}) for NSE1 phosphor

evaluated using the well-known Arrhenius equation [49]:

$$I_T = \frac{I_o}{1 + \alpha \exp\left(-\frac{\Delta E_a}{K_B T}\right)} \quad (13)$$

In the above mentioned formula, α and K_B (Boltzmann's constant = 8.617×10^{-5} eV/K) are the constants, I_o and I_T denotes to the initial emission intensity and emission intensity at specific temperature (T) in K. Figure 12b represents the evaluated activation energy (ΔE_a) as 0.34 eV by plotting $\ln[(I_o/I_T)-1]$ versus $1/K_BT$ and fitted linearly. In addition, the achieved activation energy of NSE1 phosphor is fairly higher than other phosphors reported earlier [50–52]. Thus, the as-prepared phosphor has shown tremendous thermal stability and motivated to implement in solar cell applications.

4 Conclusions

In this report, Sm^{3+} activated and Sm^{3+}/Eu^{3+} bi-activated microcrystalline NCMS phosphors have been synthesized by following the high temperature solid state reaction approach. Hexagonal pure phase has been obtained at high temperature and confirmed through X-ray diffraction technique, in which all diffraction patterns are well agreed with the standard JCPDS diffraction pattern. Emission spectra of Sm^{3+} doped NCMS phosphors have been recorded with n-UV excitation and optimized Sm^{3+} ions concentration in the host matrix is found to be 3.0 mol%. Further, PL spectra of Sm^{3+}/Eu^{3+} bi-activated phosphors under 375, 403 and 467 nm excitation wavelengths indicated that the intensity of emission peaks originated from Sm^{3+} ions is declining, whereas the intensity related to Eu^{3+} ions is increasing with rise in the concentration of Eu^{3+} ion. The incorporation of Sm^{3+} and Eu^{3+} into the host material enhances the efficiency of energy conversion process as well as provide the pathway to tailor the emission wavelength in a specific red region required for solar cell application. The decay time in the current work decreased with increasing the Eu^{3+} ions concentration as a result of energy transfer from the sensitizer (Sm^{3+}) to activator (Eu^{3+}) ions at 403 nm excitation. Based on the aforementioned investigations including impressive down-converting features with notable chromaticity coordinates of thermally stable NCMS: Sm^{3+}/Eu^{3+} phosphors are highly preferred for solar cell and other optoelectronic applications.

Author contributions

D: Conceptualization, Formal analysis, Investigation, Methodology, Software, Writing—original draft; MJ: Supervision, Formal analysis, Methodology, Visualization, Validation, Writing—review and editing. All authors read and approved the final manuscript.

Funding

One of the authors, Ms. Deepali, is grateful to Delhi Technological University, Delhi, for providing the financial support.

Data availability

The raw/processed data required to reproduce these findings cannot be shared at this time as the data also forms part of an ongoing study.

Declarations

Conflict of interest The authors have no conflict of interest to declare.

References

1. S. Razzaq, A. Asghar, C. Lou, H. Diao, S. Huang, Y. Yin, J. Alloys Compd. **907**, 164512 (2022)
2. Y. Tan, Y. Yan, H. Du, X. Hu, G. Li, Y. Kuang, M. Li, D. Guo, Opt. Mater. (Amst). **85**, 538–544 (2018)
3. M. Yu, H. Xu, Y. Li, Q. Dai, G. Wang, W. Qin, J. Colloid Interface Sci. **559**, 162–168 (2020)
4. J. Sahariya, P. Kumar, A. Soni, Mater. Chem. Phys. **199**, 257–264 (2017)
5. W.J. Ho, S.K. Feng, J.J. Liu, Y.C. Yang, C.H. Ho, Appl. Surf. Sci. **439**, 868–875 (2018)
6. W. Bin Hung, T.M. Chen, Sol. Energy Mater. Sol. Cells. **133**, 39–47 (2015)
7. D.C. Yu, S. Ye, M.Y. Peng, Q.Y. Zhang, J.R. Qiu, J. Wang, L. Wondraczek, Sol. Energy Mater. Sol. Cells. **95**, 1590–1593 (2011)

8. F. Zhang, Y. Wang, Y. Tao, Mater. Res. Bull. **48**, 1952–1956 (2013)
9. W.J. Ho, Y.T. Shen, Y.J. Deng, C.W. Yeh, R.S. Sue, Thin Solid Films. **618**, 141–145 (2016)
10. P. Kumar, S. Kanika, R. Singh, A. Lahon, G. Gundimeda, B.K. Gupta, Gupta, J. Lumin. **196**, 207–213 (2018)
11. W.J. Ho, W. Bin Bai, J.J. Liu, H.P. Shiao, Thin Solid Films. **660**, 651–656 (2018)
12. A.R. Kadam, S.J. Dhoble, J. Alloys Compd. **884**, 161138 (2021)
13. P.K. Tawalare, P.D. Belsare, S.V. Moharil, Opt. Mater. **100**, 109668 (2020)
14. R.S. Yadav, S.B. Monika, S.J. Rai, Dhoble, Prog. Solid State Chem. **57**, 100267 (2020)
15. K. Jha, M. Jayasimhadri, J. Lumin. **194**, 102–107 (2018)
16. R. Yantake, A. Sidike, T. Yusufu, J. Rare Earths. **40**, 390–397 (2022)
17. R. Sagheer, M. Khalil, V. Abbas, Z.N. Kayani, U. Tariq, F. Ashraf, Optik. **200**, 163428 (2020)
18. K. Rawat, A.K. Vishwakarma, K. Jha, Mater. Res. Bull. **124**, 110750 (2020)
19. T. Tatarchuk, M. Bououdina, W. Macyk, O. Shyichuk, N. Paliychuk, I. Yaremiy, B. Al-Najar, M. Pacia, Nanoscale Res. Lett. **12**, 141 (2017)
20. M. Deepali, Jayasimhadri, Luminescence. **37**, 1465–1474 (2022)
21. J. Zheng, X. Wu, Q. Ren, W. Bai, Y. Ren, M. Wang, O. Hai, Opt. Laser Technol. **122**, 105857 (2020)
22. Y. Yang, P. Gong, W. Ma, R. Hao, X. Fang, Chin. Phys. B **30**, 067803 (2021)
23. D. Jiang, L. Geng, S. Zhou, Y. Wang, Inorg. Chem. Commun. **142**, 109668 (2022)
24. Z. Xie, X. Liu, W. Zhao, J. Mater. Sci. Mater. Electron. **31**, 7114–7122 (2020)
25. Z. Wang, P. Li, Z. Yang, Q. Guo, J. Lumin. **151**, 170–175 (2014)
26. F. Zhang, W. Zhang, Z. Zhang, Y. Huang, Y. Tao, J. Lumin. **152**, 160–164 (2014)
27. R.N. Perumal, A.X. Lopez, G. Subalakshmi, Optik. **170**, 125–131 (2018)
28. J. Jeong, V.R. Bandi, B.K. Grandhe, K. Jang, H.S. Lee, S.S. Yi, J.H. Jeong, J. Korean Phys. Soc. **58**, 306–310 (2011)
29. T. Wang, Y. Hu, L. Chen, X. Wang, J. Mater. Sci. Mater. Electron. **26**, 5360–5367 (2015)
30. S. Kaur, A.S. Rao, M. Jayasimhadri, Ceram. Int. **43**, 7401–7407 (2017)
31. H. Kaur, M. Jayasimhadri, M.K. Sahu, P.K. Rao, N.S. Reddy, Ceram. Int. **46**, 26434–26439 (2020)
32. G. Ramakrishna, H. Nagabhushana, S.C. Prashantha, S.C. Sharma, B.M. Nagabhushana, Spectrochim. Acta - Part A Mol. Biomol. Spectrosc. **136**, 356–365 (2015)
33. M.K. Sahu, M. Jayasimhadri, K. Jha, B. Sivaiah, A.S. Rao, D. Haranath, J. Lumin. **202**, 475–483 (2018)
34. K.B. Morebodi, L. Reddy, A. Balakrishna, L.J.B. Erasmus, H.C. Swart, P.L. Masiteng, Solid State Sci. **134**, 107059 (2022)
35. H.X. Jiang, S.C. Lü, Mater. Res. Bull. **111**, 140–145 (2019)
36. D.L. Dexter, J.H. Schulman, J. Chem. Phys. **22**, 1063 (1954)
37. W.J. Yang, L. Luo, T.M. Chen, N.S. Wang, Chem. Mater. **17**, 3883–3888 (2005)
38. Q. Chen, X. Yang, G. Zhang, Q. Ma, S. Han, B. Ma, Opt. Mater. (Amst). **111**, 110585 (2021)
39. M. Fhoula, T. Koubaa, M. Dammak, Opt. Laser Technol. **130**, 106352 (2020)
40. N.P. Rajesh, G. Subalakshmi, C.K. Jayasankar, Opt. Mater. (Amst). **85**, 464–473 (2018)
41. X. Min, Z. Huang, M. Fang, Y. Liu, C. Tang, X. Wu, Inorg. Chem. **53**, 6060–6065 (2014)
42. S. Lou, P. Zhang, Y. Chen, Q. Xiong, K. Qiu, Ceram. Int. **47**, 10174–10184 (2021)
43. K. Li, R.V. Deun, Dyes Pigm. **155**, 258–264 (2018)
44. S. Song, J. Si, J. Zhang, G. Cai, RSC Adv. **9**, 1029–1035 (2019)
45. K. Saidi, M. Dammak, RSC Adv. **10**, 21867–21875 (2020)
46. C. Huang, T. Kuo, T. Chen, Opt. Express **19**, 238–241 (2011)
47. R. Datt, S. Bishnoi, D. Hughes, P. Mahajan, A. Singh, R. Gupta, S. Arya, V. Gupta, W.C. Tsoi, Sol. RRL. **6**, 2200266 (2022)
48. H. Guo, Q. Shi, K.V. Ivanovskikh, L. Wang, C. Cui, P. Huang, Mater. Res. Bull. **126**, 110836 (2020)
49. S.J. Dai, D. Zhao, R.J. Zhang, L. Jia, Q.X. Yao, J. Alloys Compd. **891**, 161973 (2022)
50. S. Yi, X. Hu, B. Liang, G. Hu, W. Zhao, Y. Wang, J. Lumin. **207**, 105–113 (2019)
51. J. Wu, C. Liu, H. Jia, Y. Qi, Z. Liu, Y. Hu, F. Feng, J. Lumin. **245**, 118791 (2022)
52. Y. Ding, N. Guo, M. Zhu, W. Lv, R. Ouyang, Y. Miao, B. Shao, Mater. Res. Bull. **129**, 110869 (2020)

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

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

Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-17236-2

Efficient image encryption using the Tinkerbell map in conjunction with linear feedback shift registers

Puneet Kumar Pal[1] · **Dhirendra Kumar**[1] · **Varun Agarwal**[1]

Received: 19 October 2022 / Revised: 25 July 2023 / Accepted: 25 September 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract
Images are considered crucial for conveying information due to their visualisation properties and capacity to hold a large amount of data. To safeguard image data from potential risks of information leakage, numerous image encryption algorithms have been developed. These algorithms often make use of chaotic maps, known for their high unpredictability, ergodicity, and sensitivity to parameters and initial values. This paper introduces a new chaos-based digital image encryption algorithm to protect digital images from unauthorised access or attacks. The algorithm heavily relies on permutation and diffusion processes. Specifically, it employs two rounds of pixel permutation using a Tinkerbell chaotic map sequence on the source image and two rounds of pixel diffusion using the Linear Feedback Shift Register and Logistic Map sequences on the permuted image. The encryption algorithm's effectiveness is thoroughly examined through various cryptanalysis techniques, including key space analysis, information entropy, correlation coefficient, differential attack, key sensitivity, histogram analysis, occlusion attack, noise attack, and encryption execution time. The experimental results are then compared with the most recent literature to demonstrate the algorithm's reliability and its ability to withstand diverse attacks. Notably, the proposed algorithm stands out as a secure and viable encryption solution, requiring less computational time than previous studies, thus making it a more practical option for real-world applications.

Keywords Image encryption · Chaos based cryptography · Logistic map · Tinkerbell map · Linear feedback shift register

✉ Dhirendra Kumar
dhirendrakumar@dtu.ac.in

Puneet Kumar Pal
puneetkumarpal_2k21phdam12@dtu.ac.in

Varun Agarwal
agarwalvarun2000@gmail.com

[1] Department of Applied Mathematics, Delhi Technological University, Delhi, India

Published online: 19 October 2023

1 Introduction

Due to the advancement in information and communication technology, most multimedia data, such as video, images, and audio, is exposed to a shared and accessible network. As shared and accessible networks are prone to attacks, this necessitates security against potential information leakage hazards. Among multimedia data, images are one of the most widely used forms of information that is shared over insecure networks, e.g., territorial images for security surveillance and medical images. These images may contain personal and confidential information that can be misused by attackers. These challenges led to the development of several techniques to provide secure communication of sensitive images over insecure communication mediums. Several methods such as encryption [5, 13, 18, 21, 51], watermarking [7, 16, 47], and data hiding [33, 53] have evolved to provide secure communication for sensitive information in the form of images. Image encryption is one of the popular techniques that aims to encrypt the pixel intensity values for secure transmission. The encryption algorithms based on conventional symmetric keys such as DES [43] and AES [48] cannot be used in image encryption due to the extensive data capacity and data redundancy of images. Furthermore, the presence of significant correlation among neighbouring pixels in plain images cannot be removed by traditional encryption algorithms. By analyzing correlation, adversaries can extract information from the cipher image and thus obtain the plain image. Hence, encryption techniques should be designed to reduce the correlation among the pixels of the cipher image. For this, the encryption algorithm should encompass permutation to modify the location of image pixels and diffusion to modify the pixel intensity values. In literature, it has been shown that the permutation and diffusion of pixels for encryption can be achieved effectively using chaotic maps [3, 22, 25, 56, 58].

Chaos emerges from nonlinear dynamical systems that are neither convergent nor periodic. Due to the unpredictable behaviour of chaotic maps and their dependence on initial seed values, these maps are often employed in image cryptography. Since the inception of the application of chaos in cryptography [32], the study of the chaotic system has attracted the attention of researchers working on multimedia security. This leads to the development of several chaotic maps and their utilisation for chaos-based image encryption. Some of the chaotic maps extensively used in image encryption systems are Logistic Map [39], Baker Map [30], Cat Map [34], and Chen Map [52]. Recently, the research work [9] proposed a novel 2D infinite collapse map based on two 1D infinite collapse maps with the modulation operation for image encryption. Another research work [14] by Fridrich discussed the application of a 2D chaotic map for permutation in the encryption algorithm. In the research work [50], the authors designed a 2D chaotic map based on an existing Hénon map for image cryptography. In research work [19], authors suggested a fractional-order 2D discrete chaotic map for image encryption. Recently, authors of the research work [26] utilized the Salomon function [40] and proposed a 2D Salomon map for developing an image encryption algorithm. The research work [36] suggested a 2D Logistic Coupling Cubic Chaotic map (2D-LCCCM), which combines Logistic map and Cubic map. The research work [41] introduces the 2D Hénon Sine map based on the Hénon map and Sine map. In the research work [44], a 2D-SCMCI map is constructed using one linear function and two 1D chaotic maps. The authors in the research work [27] used a circuit to implement and test the dynamical behaviour of chaos under various beginning circumstances. Regardless of its advantages and disadvantages, the effectiveness of chaotic maps is often restricted in terms of chaos indicators. Image encryption applications, of course, necessitate high-performance chaotic systems.

Recently, an algorithm [42] based on the Chameleon chaotic map and DNA coding rules was proposed. The work is resistant to differential attack for the three cases, i.e., line equilibrium, self-excited equilibrium, and no equilibrium. For stable equilibrium, unified averaged changed intensity (UACI) values are not satisfactory. The research work [52] uses Chen's chaotic system in combination with a semi-tensor product and Hilbert curve to encrypt the image. In this research work, permutation and diffusion are performed at the same time. The research work [38] used an intertwining Logistic map and Brownian motion to permute and diffuse the pixels of the plain image, respectively. The authors added random data to the surroundings to increase the complexity of the cipher image, even for the same keys. An encryption technique [15] based on the improved Sine-Tent map, using the concept of hiding the least significant bits, is proposed. After hiding the bits, the pixel information changes in an unpredictable way, leading to a robust encryption method. Then, the image is encrypted using the sequences generated by the improved Sine-Tent map. In research work, [34] authors proposed masking of the image before encryption. The generalised Arnold's cat map is combined with DNA rules and a 2D-Logistic-Sine-coupling map. This is implemented to provide a robust and efficient image encryption algorithm. The authors in the research work [11] used the Logistic map, piece-wise linear chaotic map, and DNA encoding to propose an encryption algorithm. A new cryptosystem, [61], based on a random Hamiltonian path and an adjusted Bernoulli map, used the Hamiltonian path to permute pixels' and the adjusted Bernoulli map to diffuse the pixels' of the plain image. A multi-chaotic image encryption algorithm [62] based on 2D Logistic-adjusted-Sine mapping (2D-LASM) and Logistic-Sine system (LSS) in combination with random DNA coding is proposed. The random sequences obtained from 2D-LASM are used to get column and row permuted matrices. The elements of the second column and row permuted matrices are used as row and column coordinates in the permutation phase. This was applied to permuted DNA-encoded images. While in diffusion, a DNA XOR operation is performed on the centre point and horizontal line, and the encrypted image is subsequently diffused from the central line to the bottom and top directions of the matrix. The research work [45] proposed a synchronous permutation-diffusion encryption scheme based on an intertwining Logistic map and DNA. In research work [59], the authors used double chaotic maps, SHA-3 and an auto-updating system to propose an encryption scheme. The authors of the research work [60] used a 2D hyperchaotic system to generate the chaotic sequences. The generated sequences are used for the permutation and diffusion of image pixels. The last pixel and the current sequence term are XOR-ed with the current permuted pixel to get the current diffused pixel. In research work [6], a Chaotic Economic Model is used to encrypt the images. The algorithm is less sensitive to the keys since it is efficient only up to the error 10^{-14}. The algorithm is unable to resist differential attacks. In the research work [24], a modified image encryption algorithm is described that connects a Sine map and a chaotic Tent map with a Hill cipher to improve security and create a modified algorithm. In terms of security and complexity, this chaos-based image encryption technique is exceptionally efficient. The adoption of the Sine map shortens the time required for a chaotic system to iterate. The temporal complexity of the encryption procedure grows as iteration time increases. The authors of the research work [12] developed a bit-reversion operation and used it in the diffusion of the image pixels. The confusion and diffusion operations are dependent on the newly developed 2D$-e\pi$ map. In the research work [28], the plain images are encrypted based on chaotic maps and cellular automata, and SHA256. The research work [46] is based on 2D-CLSS hyperchaotic map. The encryption process uses simultaneous confusion and diffusion processes to encrypt the image. The authors in the research work [55] used the fractal sorting matrix and chaotic Chen system in confusion and diffusion processes.

Another frequently explored field in cryptography is cryptanalysis which aims to break the security of the existing cryptography algorithms. The concept of 'cryptanalysis' refers to unauthorized decryption without having knowledge of the secret key [29]. Several such pairs of research publications reporting cryptography algorithms and corresponding cryptanalysis algorithms can be noticed [20, 29, 54] and [10, 37, 49], respectively. Hence, there is always a motivation for developing newer image encryption algorithms.

The following are the most significant and major weaknesses identified in the literature:

1. The majority of classic algorithms are built on key streams in a unique and limited way.
2. The algorithms proposed are not tested on an extensive image dataset.
3. Most standard methods have failed to obtain meaningful results in terms of security quality indicators.
4. Many algorithms require a high computational time which might make them infeasible for practical use.
5. Most of the image encryption methods have not conducted a deeper statistical study of the encryption/decryption system against various types of attacks.
6. The majority of recent research has not conducted detailed and wide-ranging comparisons with relevant research works.

In this work, two chaotic maps, namely the Tinkerbell Map (TM) and the Logistic Map (LM), are explored to develop an image encryption algorithm. By introducing two chaotic maps, TM and LM, into the image cryptography system, we achieve a greater number of control parameters. However, this augmentation in control parameters also amplifies the system's susceptibility to variations in the initial values during encryption. Also, the simplicity of these chaotic maps offers rapid processing speed which makes the image cryptography system time efficient. As the chaotic maps have a large key space and are sensitive to control parameters and initial values, the encryption algorithm developed using these maps is resistant to brute-force attacks, chosen plaintext attacks, and ciphertext attacks. Further, Linear Feedback Shift Register (LFSR) is used since it generates uncorrelated sequences. The sequences generated are used to diffuse the image pixels and hence hide the statistical relationship between neighbouring pixels, thus resisting statistical attacks. The algorithm developed has a fast computational speed in comparison to other related methods. Several performance measures such as information entropy, histogram analysis, correlation coefficient, cross-correlation coefficient, CPU processing time, occlusion, and noise attack are evaluated for different image data sets to validate the image encryption algorithm. Experimental results indicate that the proposed image encryption algorithm is immune to several attacks such as brute-force attacks, chosen plaintext and ciphertext attacks, correlation attacks, statistical attacks, differential attacks, occlusion attacks, and noise attacks. The main contributions of the proposed work can be summarized as follows:

1. The proposed image encryption algorithm has the advantage of having two discrete non-linear chaotic maps. Since both chaotic maps are sensitive to the initial values and control parameters, the proposed image encryption algorithm is key-sensitive and has a large keyspace.
2. The encryption algorithm uses the Linear Feedback Shift Register (LFSR) to generate the sequences. The generated sequences are used in the diffusion process. Since LFSR generates uncorrelated sequences, there is a low correlation between the image pixels of the cipher image. Thus, the algorithm is resistant to statistical attacks.
3. The research work presents an analysis of a variety of images for encryption tasks.
4. The computational time for the proposed approach is very low. It can be used for various real-life applications.

The remainder of this paper is arranged as follows: Section 2 discusses the preliminaries on the proposed approach. Section 3 discusses the proposed encryption and decryption algorithms in detail. Section 4 includes the experimental results. Finally, Section 5 summarises the research outcome of the study and future directions.

2 Preliminaries

This section discusses the two chaotic maps and the Linear Feedback Shift Register. A brief introduction to the Tinkerbell map and Logistic map is provided. Further, the phase and bifurcation diagrams of the Tinkerbell map are displayed along with the bifurcation diagram and Lyapunov spectrum of the Logistic map.

2.1 Tinkerbell map (TM)

TM is a chaotic discrete map with four control parameters (a, b, c, and d). The TM is defined as [17]:

$$(z_{n+1}, w_{n+1}) = (z_n^2 - w_n^2 + az_n + bw_n, 2z_n w_n + cz_n + dw_n) \tag{1}$$

The above equation shows chaotic behaviour for parameter values $a = 0.9$, $b = -0.6013$, $c = 2.0$, $d = 0.50$. For the initial values of the involved variable $z_0 = -0.72$ and $w_0 = -0.64$, Fig. 1a and b shows the plot of TM map and bifurcation diagram respectively. A detailed description of TM can be found in [17].

2.2 Logistic map (LM)

LM is the 1-D simple dynamic non-linear equation, it shows complex chaotic characters [59]. The mathematical expression for the LM is given below:

$$x_{n+1} = \mu x_n (1 - x_n) \tag{2}$$

Where $\mu \in (0, 4]$ is the control parameter. The x_n is a chaotic sequence generated with the initial seed x_0. The produced chaotic sequence is pseudo-random, non-periodic, and unexpected for the proper selection of bifurcation parameter μ value. The bifurcation diagram and Lyapunov Exponent corresponding to LM are shown in Fig. 1c and d, respectively. From the bifurcation diagram, it can be observed that the optimal value of μ lies in the range of 3.6 to 4. Since the Lyapunov exponent of the LM is positive in this range, we can say that LM is chaotic in nature.

2.3 Linear feedback shift register (LFSR)

In computing, an LFSR is a shift register whose input bit is a linear combination of its prior state. If a good feedback function is used, the LFSR generates a random bit stream with a lengthy and complicated cycle. An n-stage LFSR is distinguished by a feedback polynomial of degree n over $GF(2)$. If the input polynomial is primitive, the resulting sequence of states is periodic and has a period $(2^n - 1)$. Here, we have considered an initial input in the form of polynomial $x^8 + x^7 + x^6 + x^5 + x^2 + x + 1$ (11110011) with degree $n = 8$. There are 255 possible initial states. Figure 2 exhibits the working of LFSR used in this

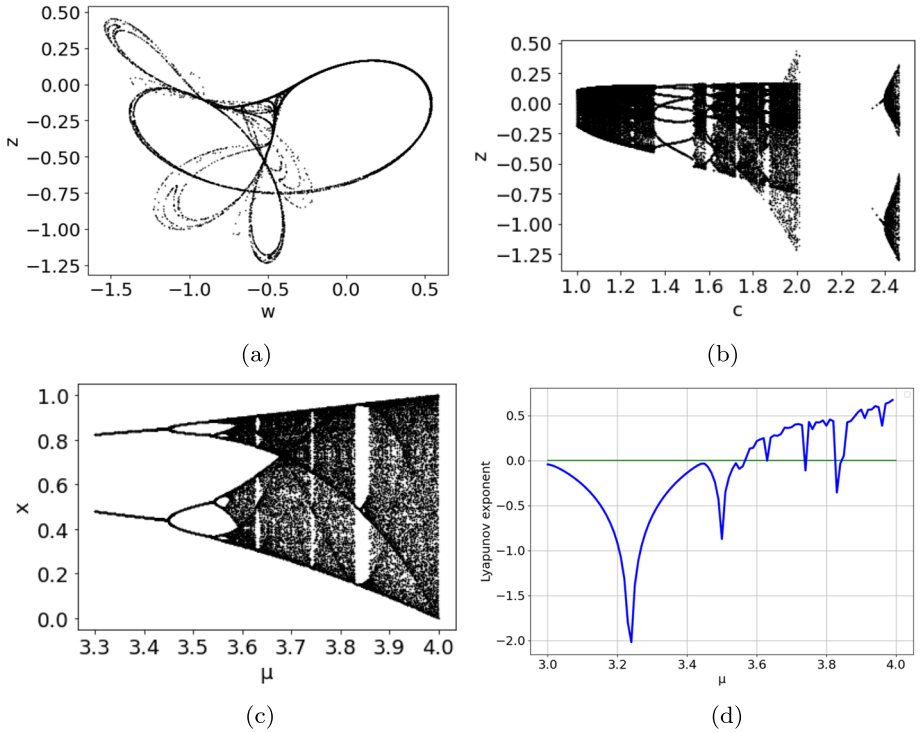


Fig. 1 (a) Phase diagram of Tinkerbell map; (b) Bifurcation diagram of Tinkerbell map; (c) Bifurcation diagram of Logistic map; (d) Lyapunov exponent of Logistic map

work. The bits E_0, E_2, E_5, E_7 are the bits to be XORed to get the output bit sequence $E_1, E_2, E_3, E_4, E_5, E_6, E_7, E_8$ and feedback to the rightmost bit $E_8 = E_0 \oplus E_2 \oplus E_5 \oplus E_7$.

3 Proposed methodology

In this section, we have discussed the proposed encryption methodology that utilizes LFSR, TM, and LM chaotic sequences. These chaotic sequences provide a good amount of permutation and diffusion capability in the image encryption process. Further, the image decryption process is described.

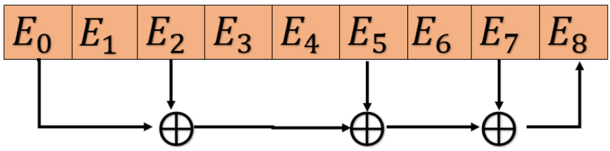


Fig. 2 Schematic representation of LFSR

3.1 Encryption algorithm

The proposed image encryption algorithm consists of four steps: key generation for TM, sequence generation, two-step permutation, and two-step diffusion process. All of these steps are performed to modify the image property, which results in an encrypted image. The Fig. 3 shows the flow diagram of the encryption process that involves a two-step permutation process and a two-step diffusion process. The permutation process in the encryption algorithm makes the task of decryption difficult, and if a single bit in a secret key is changed, the calculation of most of the bits in the ciphertext is highly affected. The goal of the confusion process is to hide the relationship between the plaintext and the key. Diffusion asserts that if one bit of the plaintext is altered, half the number of bits in the ciphertext should change, and similarly if a single bit of the ciphertext is changed, about half of the plaintext bits should change. The goal of the Diffusion process is to conceal the statistical relationship between the ciphertext and the plaintext. For encrypting an image P of size $M \times N$, the first step is to obtain the two initial seeds for the TM map. The initial seed for the TM map is generated using the following (3).

$$\begin{aligned} x0 &= \frac{\sum_{i=1}^{M} \sum_{j=1}^{\frac{N}{2}} P[i, j]}{M \times \frac{N}{2} \times 256}, \\ y0 &= \frac{\sum_{i=1}^{M} \sum_{j=\frac{N}{2}}^{N} P[i, j]}{M \times \frac{N}{2} \times 256} \end{aligned} \tag{3}$$

Using the initial seeds x_0 and y_0 obtained from (3) and the parameters described in Subsection 2.1, the TM is used to generate a 2-D discrete chaotic sequence of length $4 \times \max(M, N)$. Using the first dimension of obtained TM map, two sequences X_1 and X_2 of size M are obtained after ignoring the initial few terms so that both sequences X_1 and X_2 are non-overlapping. Similarly, another two sequences Y_1 and Y_2 of size N are extracted using the second dimension of the TM map. Further, the sequences X_1, X_2, Y_1, and Y_2 are sorted and their indices are stored in X_1^{ind}, Y_1^{ind}, X_2^{ind}, and Y_2^{ind} respectively, for implementing the two-step permutation processes described by (4) and (5). Here, the first level permutation of the image pixels is implemented using the index sequence X_1^{ind}, Y_1^{ind} and C_1 is obtained using

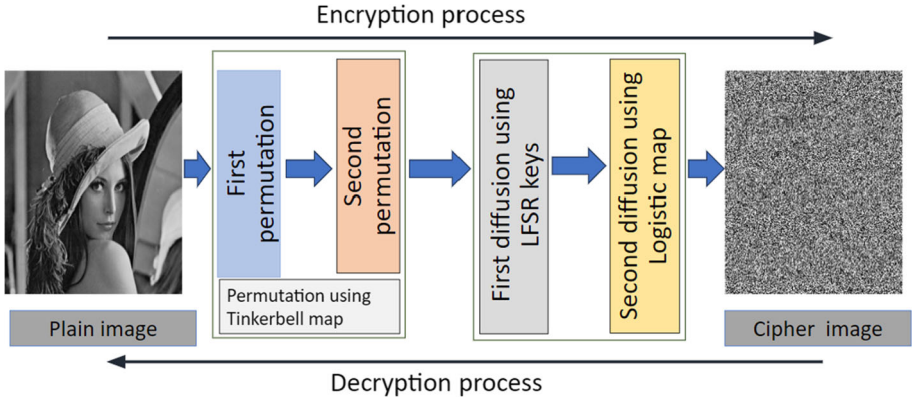


Fig. 3 Encryption and decryption process of the proposed approach

the plain image P as described by the following equation.

$$C_1[X_1^{ind}[i], Y_1^{ind}[j]] = P[i, j] \tag{4}$$

In continuation, the second step of the permutation process is performed using the other index sequences (X_2^{ind}, Y_2^{ind}) and C_2 is obtained using the C_1 as described by the following equation.

$$C_2[X_2^{ind}[i], Y_2^{ind}[j]] = C_1[i, j] \tag{5}$$

The C_2 is the permutated image obtained after a two-step permutation process. Now, the pixels of C_2 are to be diffused. A two-step diffusion task is performed using the sequences of LFSR, and LM. In the first step of the diffusion process, LFSR is utilized to obtain the chaotic sequence of length $2 \times M \times N$ using the parameters described in Subsection 2.3 and sequence K having elements equivalent to the size of C_2 is extracted after ignoring initial few terms. The element-wise XOR operation is performed between the sequence K and C_2 and first-level diffused image C_3 is obtained as described in (6).

$$C_3[i, j] = C_2[i, j] \oplus K[i, j] \tag{6}$$

Further, for the second level of the diffusion process, the LM map is utilized. A chaotic sequence of length $2 \times M \times N$ is obtained using the parameters described in Subsection 2.2 and sequence L having elements equivalent to the size of C_3 is extracted after ignoring the initial few terms. The element-wise XOR operation is performed between the sequence L and C_3 and the second-level diffused image C_4 is obtained as described in the following equation.

$$C_4[i, j] = C_3[i, j] \oplus L[i, j] \tag{7}$$

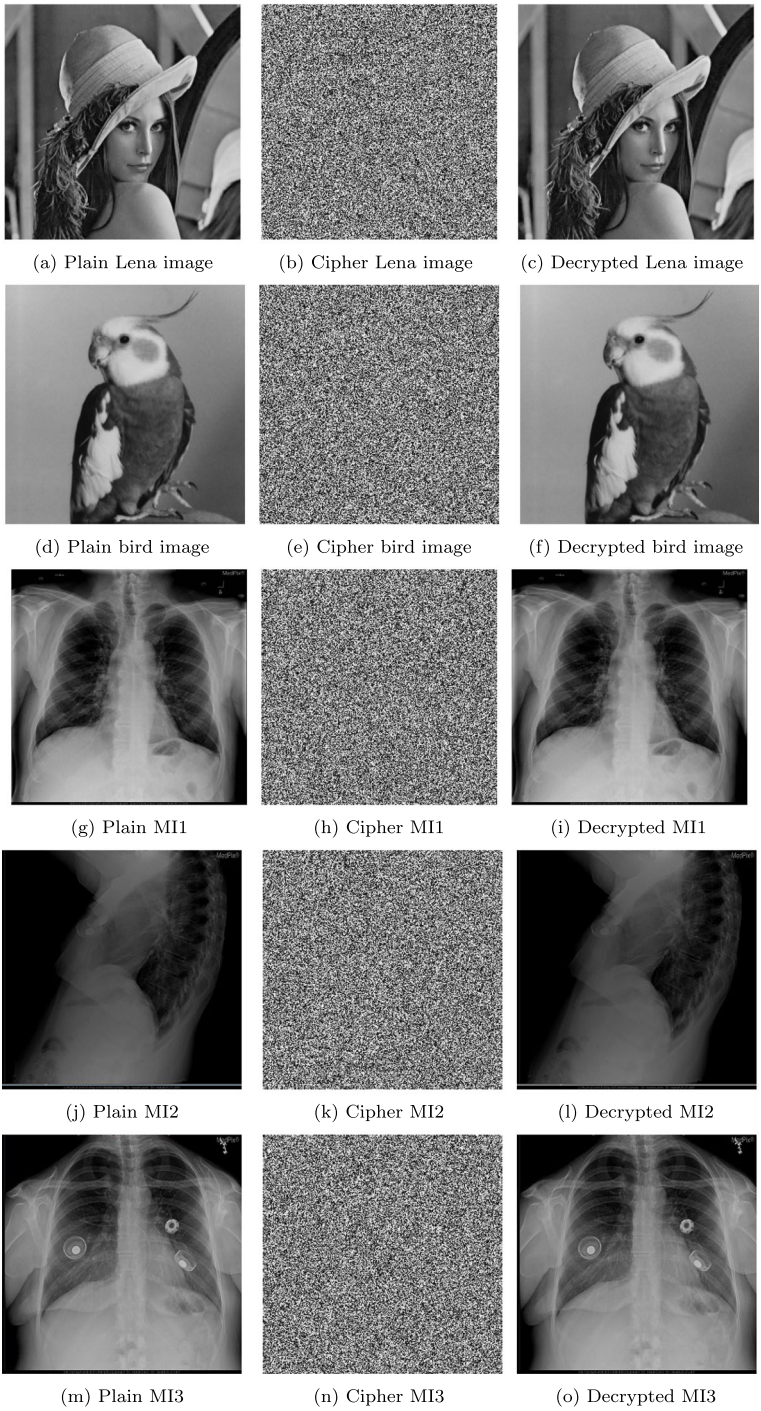
The obtained final image C_4 is reshaped into the size of $M \times N$, and hence the encrypted image is obtained.

3.2 Decryption algorithm

Since the encryption and decryption processes are symmetric, the decryption process involves the inverse steps of the encryption process. First, inverse diffusion is applied using the sequences of LM. In the second step, LFSR sequences are used to diffuse the pixels. At last, the inverse permutation is applied using the TM. The decryption process is exhibited in Fig. 3.

4 Experimental results and discussion

The efficiency, robustness, and execution time analysis of the proposed method is observed by comparing the results of the proposed method with those of various other methods available in the literature. The algorithm was implemented on the device with Windows 10, 2.00 Ghz Intel core processor, and 4 GB of RAM consisting of the Python 3 version. To perform the experiment, eight images of size 256×256 are collected. Five gray-scale images [1] (Lena, Goldhill, Bridge, Camera, and Bird images) and three medical images [2] are utilized for evaluating and comparing the proposed image encryption method. Figure 4 displays the plain, encrypted and corresponding decrypted images. Further, the security analysis based on several performance measures suggests that the proposed algorithm is secure and resistant to several attacks. The seed value for TM is initialized as the mean of image pixel intensity. The initial value of the LM is 0.564789352 and $\mu = 3.98555678$. The initial seed for the LFSR is [11110011].



(a) Plain Lena image (b) Cipher Lena image (c) Decrypted Lena image

(d) Plain bird image (e) Cipher bird image (f) Decrypted bird image

(g) Plain MI1 (h) Cipher MI1 (i) Decrypted MI1

(j) Plain MI2 (k) Cipher MI2 (l) Decrypted MI2

(m) Plain MI3 (n) Cipher MI3 (o) Decrypted MI3

Fig. 4 Plain, Cipher, and Decrypted images obtained using the proposed encryption and decryption algorithm

4.1 Key space analysis

A robust image encryption technique requires a broad key space. The dimension of the key space directly affects the security of an image encryption technique. If the key space is small, ciphered images can be illegally decrypted by brute-force attacks. The key space of a reliable encryption algorithm must be at least 2^{100} [4].

The keys used in the encryption process are $p, q, r, s, x_0, y_0, \mu, z_0$. The values are $p = 0.9$, $q = -0.6013$, $r = 2.0$, $s = 0.50$, $x_0 = 0.5501400232315063$, $y_0 = 0.4294922351837158$, $\mu = 3.98555678$, $z_0 = 0.564789352$. On combining, the key space is of size $10 \times 10^4 \times 10^2 \times 10^{16} \times 10^{16} \times 10^8 \times 10^9 = 10^{56} \approx 2^{186} >> 2^{100}$. Since, the key space is greater than 2^{100}, and hence the proposed image encryption approach is capable of resisting brute-force attacks.

4.2 Information entropy analysis

Information entropy measures the randomness introduced by the image encryption algorithm in the cipher image. The ideal value of the information entropy for a given cipher image should be 8. A closer value of the information entropy to 8 indicates a higher level of randomness in the cipher image. The mathematical formula to calculate information entropy is [35]:

$$H(S) = -\sum_{i=0}^{255} p_i \log_2 p_i \tag{8}$$

where $S = \{p_0, p_1, \ldots, p_{255}\}$ and p_i is the probability of occurrence of gray level i. Table 1 shows the information entropy for the cipher images obtained using the proposed image encryption algorithm along with state-of-the-art methods. Further, the mean values of the information entropy for the different image encryption algorithms are quoted in Table 1. From Table 1, it can be noted that the mean value of the information entropy for the proposed image encryption algorithm is 7.9973 which is very close to other mean values. Hence, the cipher image is more random. The best performance achieved by a method compared to other methods is shown in boldface in all the tables.

4.3 Histogram analysis

A histogram is a common tool for evaluating the uniformity of pixel intensity distribution in encrypted images. The distribution of pixel intensity values corresponding to plain and cipher images are shown by the histogram in Fig. 5. From Fig. 5, it can be observed that the histogram of cipher images is uniformly distributed. It can be concluded that the images are well-diffused and the algorithm can resist statistical analysis attacks.

4.4 Adjacent correlation coefficient analysis

This technique is used to analyse the relationship between the adjacent pixels of the image. The correlation coefficient is calculated using (9).

$$r(x, y) = \frac{Cov(x, y)}{\sqrt{\sigma_x}\sqrt{\sigma_y}} \tag{9}$$

Table 1 Comparison of information entropy for the proposed image encryption algorithm along with related research works

Research works	Lena	Goldhill	Bridge	Camera	Bird	MI1	MI2	MI3	Mean
Proposed Method	7.9974	7.9971	7.9972	7.9973	7.9972	7.9971	7.9973	7.9967	**7.9973**
Mishra et al. [34]	7.9943	7.9127	7.9840	7.9944	7.9959	7.9922	7.9749	7.9818	7.9788
Raghuvanshi et al. [38]	7.9975	7.9975	7.9965	7.9974	**7.9974**	7.9972	7.9973	7.9974	**7.9973**
Gao et al. [15]	**7.9976**	7.9972	7.9970	7.9973	7.9963	7.9974	7.9968	7.9972	7.9971
El-Shafai et al. [11]	7.9970	**7.9976**	7.9974	7.9971	7.9974	7.9971	7.9973	7.9973	**7.9973**
Zhang et al. [61]	7.9973	7.9974	7.9970	7.9974	7.9967	7.9968	7.9973	7.9973	7.9972
Zheng et al. [62]	7.9973	7.9969	7.9970	7.9971	7.9970	7.9971	7.9973	7.9974	7.9971
Suri and Vijay [45]	7.9969	7.9972	7.9975	7.9970	7.9971	7.9969	7.9973	7.9969	7.9971
Ye and Huang [59]	7.9734	7.9734	7.9734	7.9734	7.9734	7.9491	7.4165	7.9079	7.8926
Yuan et al. [60]	7.9972	7.9972	7.9971	7.9970	7.9974	**7.9975**	7.9973	7.9975	**7.9973**
Askar et al. [6]	7.5289	7.5243	7.5293	7.5330	7.5227	7.5315	7.5110	7.5264	7.5259
Teng et al. [46]	7.9971	7.9968	7.9969	7.9972	7.9954	**7.9975**	7.9972	7.9975	7.9970
Li et al. [28]	7.9951	7.9973	7.9959	**7.9977**	7.9901	7.9969	7.9947	7.9966	7.9955
Erkan et al. [12]	7.9973	7.9973	7.9965	7.9972	7.9974	7.9968	7.9971	7.9974	7.9971
Xian and Wang [55]	**7.9976**	7.9974	7.9972	7.9974	7.9967	7.9972	**7.9974**	**7.9976**	**7.9973**
Lai et al. [26]	**7.9976**	7.9968	**7.9977**	7.9972	7.9970	7.9971	**7.9974**	7.9975	**7.9973**

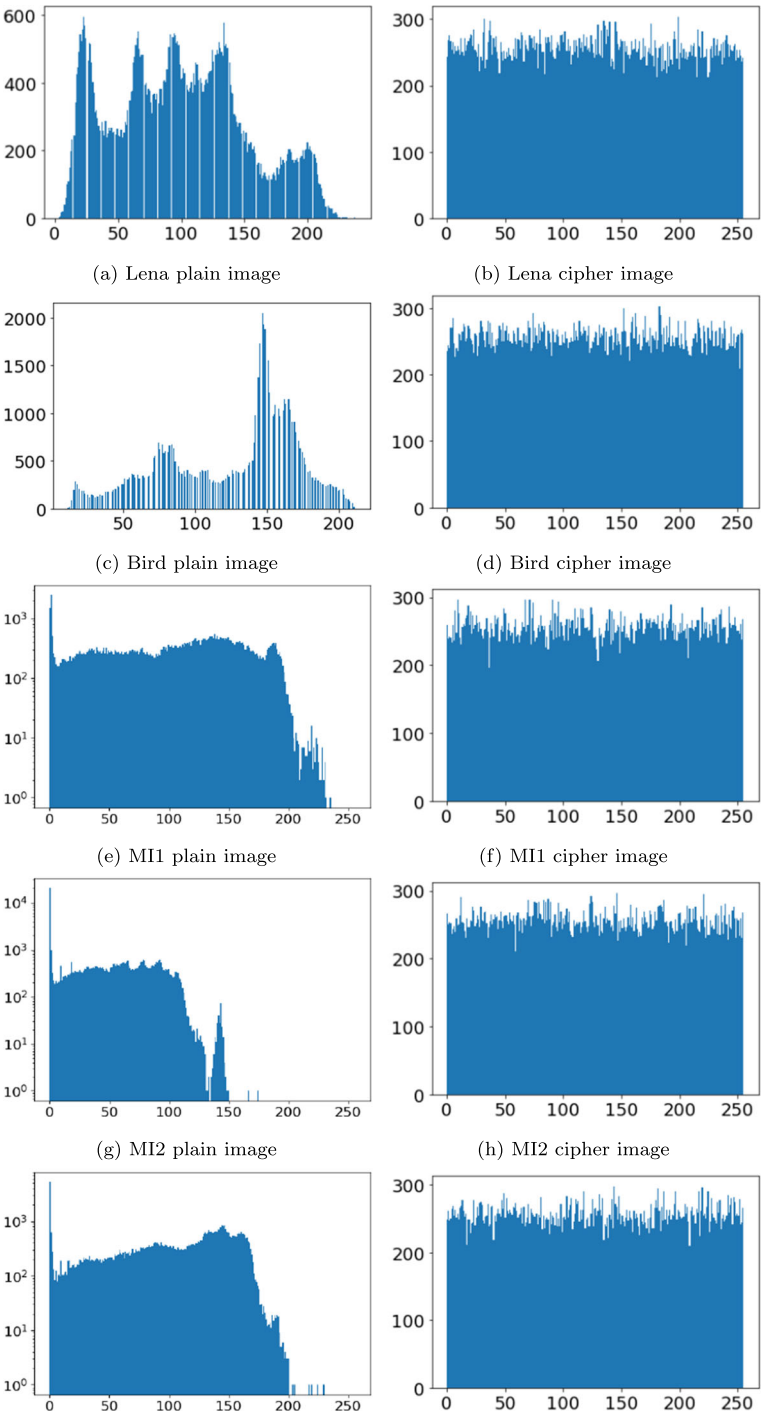
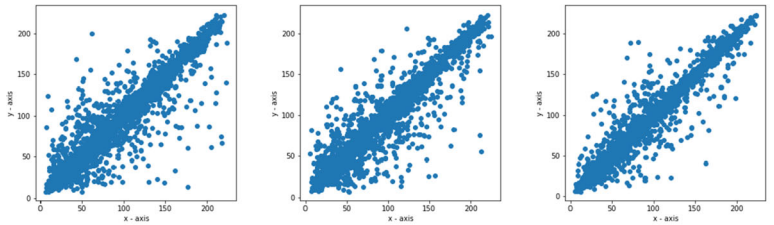
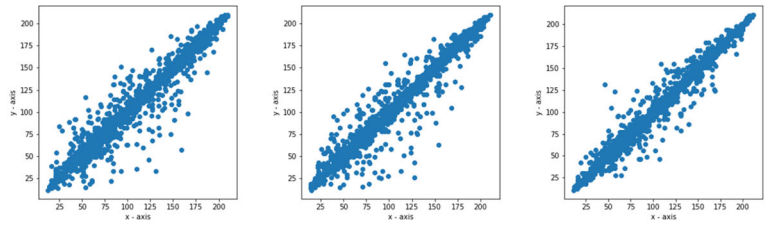


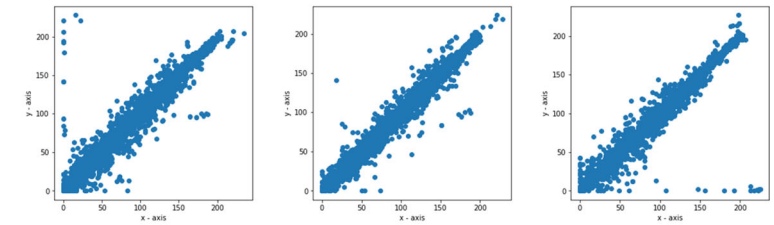
Fig. 5 Histogram of plain and cipher images obtained using proposed encryption algorithm



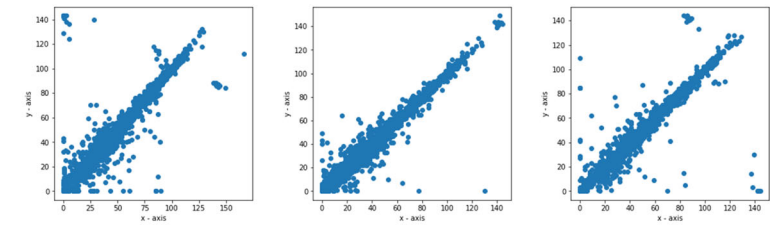
(a) Lena DD (b) Lena HD (c) Lena VD



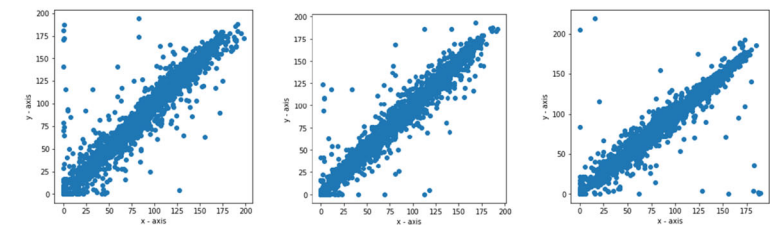
(d) Bird DD (e) Bird HD (f) Bird VD



(g) MI1 DD (h) MI1 HD (i) MI1 VD

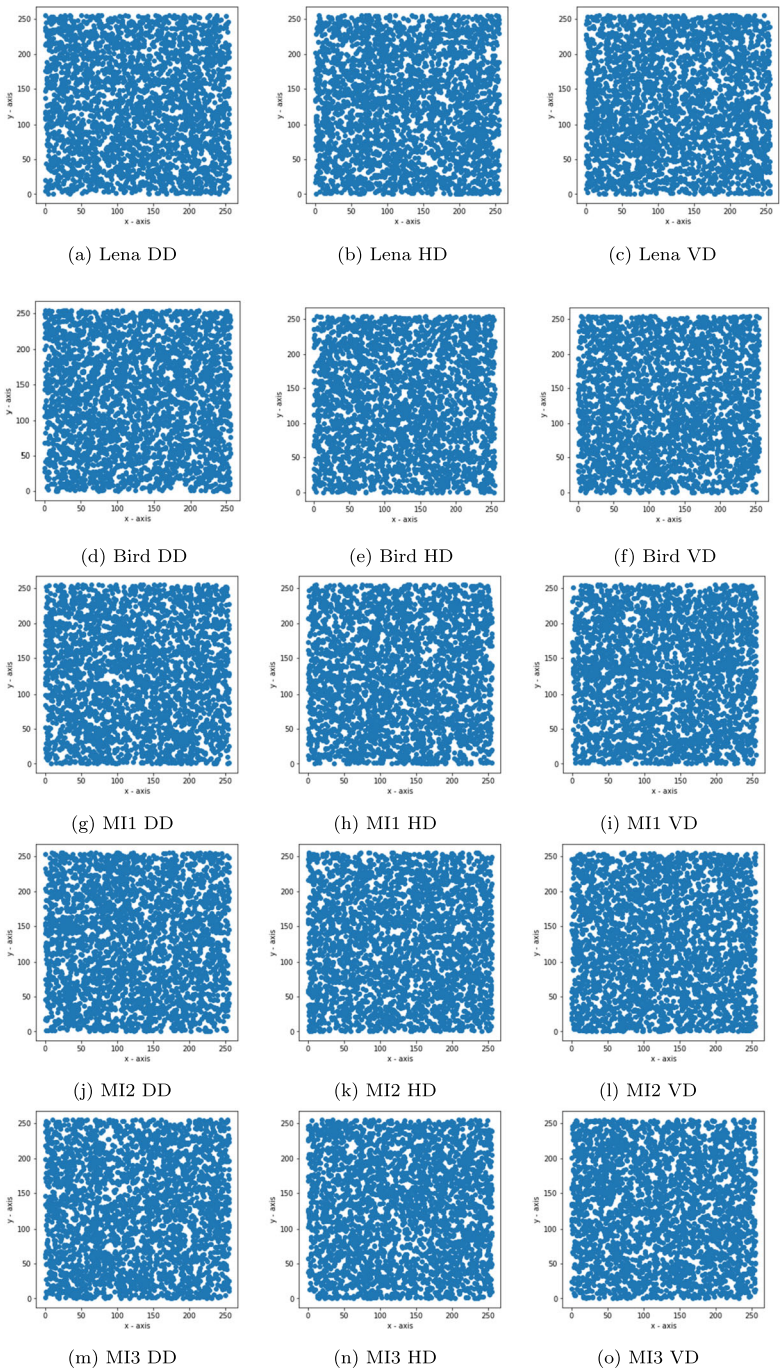


(j) MI2 DD (k) MI2 HD (l) MI2 VD



(m) MI3 DD (n) MI3 HD (o) MI3 VD

Fig. 6 Correlation distribution of 3000 random adjacent pixel pairs of various plain images



(a) Lena DD (b) Lena HD (c) Lena VD

(d) Bird DD (e) Bird HD (f) Bird VD

(g) MI1 DD (h) MI1 HD (i) MI1 VD

(j) MI2 DD (k) MI2 HD (l) MI2 VD

(m) MI3 DD (n) MI3 HD (o) MI3 VD

Fig. 7 Correlation distribution of 3000 random adjacent pixel pairs of cipher images obtained using the proposed encryption algorithm

Table 2 Correlation coefficient of plain and cipher images obtained using encryption algorithm

Image name	Horizontal direction		Vertical direction		Diagonal direction	
	Plain image	Cipher image	Plain image	Cipher image	Plain image	Cipher image
MI1	0.9868	0.0236	0.9649	-0.0062	0.9657	0.0034
MI2	0.9893	0.0057	0.9655	-0.0017	0.9514	0.0145
MI3	0.9806	-0.0006	0.9711	-0.0298	0.9678	0.0286
Bird	0.9804	0.0114	0.9887	-0.0034	0.9779	0.0168
Bridge	0.9344	-0.0157	0.9111	-0.0323	0.8678	0.0164
Camera	0.9339	0.0317	0.9584	0.0368	0.9234	-0.0116
Goldhill	0.9457	-0.0103	0.9425	0.0417	0.9104	-0.0472
Lena	0.9445	-0.0167	0.9672	0.0026	0.9324	-0.0223

Table 3 Comparison of correlation coefficient of cipher Lena image obtained using proposed encryption algorithm along with related research works

Correlation coefficient of cipher Lena image			
Research work	Horizontal	Vertical	Diagonal
Proposed Algorithm	-0.0167	0.0026	-0.0223
Mishra et al. [34]	-0.0144	0.0157	0.0216
Raghuvanshi et al. [38]	0.0168	0.0164	0.0069
Gao et al. [15]	0.0108	0.0270	0.0240
Zheng et al. [62]	-0.0400	-0.0273	-0.0143
Zhang et al. [61]	0.0143	-0.0061	0.0069
Ye and Huang [59]	0.0281	0.0245	0.0458
Askar et al. [6]	0.0428	-0.0225	0.0265
Teng et al. [46]	-0.0004	0.0105	0.0006
Li et al. [28]	0.0045	0.0018	0.0010
Erkan et al. [12]	0.0208	0.0107	0.0001
Xian and Wang [55]	0.0167	0.0245	0.0021
Lai et al. [26]	0.0001	0.0045	0.0008

where $Cov(x, y)$ is co-variance of x and y. σ_x and σ_y represent variances of x and y, respectively. The correlation coefficient value lies in the range $[-1, 1]$. A correlation value near zero indicates no correlation, while a value close to 1 or -1 indicates a significant correlation. The image encryption algorithm is said to be effective if it introduces a low correlation (close to zero) between adjacent pixels in the cipher image. To validate the proposed approach, 3000 pixels are randomly chosen from the cipher image, and the correlation coefficient is calculated between the randomly chosen pixels and the adjacent pixels in all three directions, i.e., Horizontal directions (HD), Vertical directions (VD), and Diagonal directions (DD). Figures 6 and 7 show the plot of 3000 randomly chosen pixel values and corresponding adjacent pixel values for different plain and cipher images. From Fig. 6, it can be noted that the plain images have a high correlation. From Fig. 7 it can be noted that the encryption process using the proposed approach effectively removed the adjacent pixel correlation. Further, the adjacent pixel correlation coefficient values of the plain and cipher images in all three directions are given in Table 2 for different images. It can be noted from Table 2 that the proposed encryption algorithm has significantly reduced the adjacent pixel correlation coefficient. Further, a comparison table of the adjacent pixel correlation coefficient values on the Lena image for the proposed image encryption algorithm along with the state-of-the-art methods is given in Table 3. It can be noted from Table 3 that the proposed image encryption algorithm successfully removed adjacent image pixel correlation.

4.5 Cross-correlation analysis

The confusion and diffusion operations of the encryption technique should reduce cross-correlation between the plain image and the cipher image. A low cross-correlation between the plain and cipher images guarantees that a differential attack can be prevented. The cross-correlation between plain and cipher images is calculated using (9). To calculate the cross-correlation value, three thousand pixel locations are selected randomly, and the pixel values are extracted from both the plain and the cipher image. The cross-correlation coefficient of the selected pixel values for different plain images and corresponding cipher images for the proposed image encryption algorithm along with other related methods is shown in Table 4.

Table 4 Comparison of cross-correlation coefficient obtained using the proposed encryption algorithm along with related research works

Image	Bird	Bridge	Camera	Goldhill	Lena	MI1	MI2	MI3	Mean
Proposed Algorithm	-0.0079	-0.0104	-0.0049	0.0074	0.0032	-0.0173	-0.0237	-0.0143	-0.0679
Mishra et al. [34]	-0.0044	0.0117	0.0216	-0.0091	-0.0104	-0.0049	0.0074	0.0061	0.0180
Raghuvanshi et al. [38]	-0.0004	0.0157	-0.0006	0.0105	0.0017	0.0107	0.0345	0.0075	0.0796
Gao et al. [15]	0.0041	0.0011	0.0010	0.0005	0.0008	-0.0004	-0.0018	0.0006	0.0059
Zheng et al. [62]	0.0208	0.0107	0.0023	0.0043	0.0001	-0.0104	-0.0089	0.0055	0.0244
Zhang et al. [61]	0.0167	0.0145	-0.0021	0.0120	0.0040	0.0018	0.0010	-0.0005	0.0474
Ye and Huang [59]	0.0051	0.0045	-0.0008	0.0024	-0.0079	-0.0105	0.0015	-0.0025	-0.0082
Askar et al. [6]	0.0578	0.0583	0.0664	0.0567	0.0474	0.0597	0.0382	0.0505	0.0544
Teng et al. [46]	-0.0001	0.0167	0.0245	0.0121	0.0008	0.0045	0.0018	0.0019	0.0622
Li et al. [28]	-0.0020	-0.0104	-0.0049	0.0074	0.0061	-0.0173	-0.0237	-0.0143	-0.0591
Erkan et al. [12]	-0.0044	0.0570	0.0316	-0.0014	-0.0104	-0.0049	0.0074	0.0010	0.0759
Xian and Wang [55]	-0.0004	0.0105	-0.0046	0.0150	0.0018	0.0107	0.0245	0.0045	0.0620
Lai et al. [26]	0.0045	0.0018	0.0010	0.0002	-0.0082	0.0026	0.0115	0.0013	0.0147

It shows that the plain images and encrypted images are not correlated at all, thus supporting the efficiency of the algorithm.

4.6 Occlusion and noise attack

Occlusion and noise attacks involve losing or corrupting some parts of the encrypted data (ciphertext), leading to potential image loss or distortion [23]. To evaluate the encryption scheme's effectiveness, we assess its ability to recover the original image even in the presence of such occlusion or noise attacks. In occlusion attack, a significant chunk of the encrypted image is lost during transmission. To show the robustness of the proposed approach toward the occlusion attack, we have cropped some part of the encrypted image shown in Fig. 8 (a), (b), (c), and (d). Further, we have decrypted the modified encrypted images and the decrypted images are shown in Fig. 8 (e), (f), (g), and (h). Figure 8 indicates that the suggested decryption process can successfully retrieve a substantial amount of the image's visual information, providing enough data to comprehend the visual content of the original image.

Further, To show the robustness of the proposed approach toward the noise attack, we have corrupted the encrypted images with salt and pepper noise (Fig. 9 (a), (b), (c), (d), and (e)). Whenever this noisy encrypted image is decrypted, this retains the plain image's visual content. Figure 9 shows that the proposed decryption process successfully recovers the significant image's visual information, which is sufficient to understand the visual content of the plain image. Hence, the encryption algorithm is resistant to occlusion and noise attacks.

4.7 Differential attack

Adversaries may change a few pixels of the plain image to obtain a relationship between the original encrypted image and the new encrypted image leading to a potential security vulnerability [57].

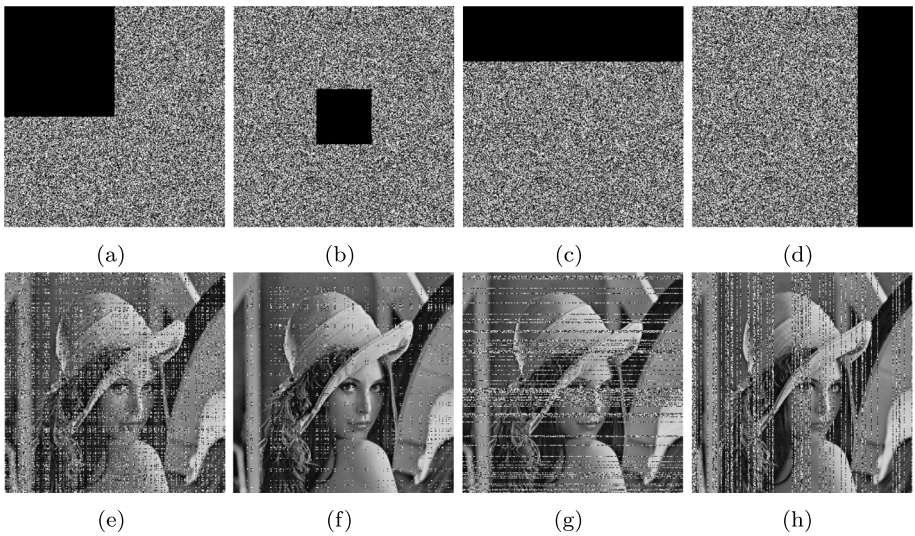


Fig. 8 Occlusion attack on cipher image (a)-(d) Occluded image; (e)-(h) Decrypted images of corresponding occluded cipher images

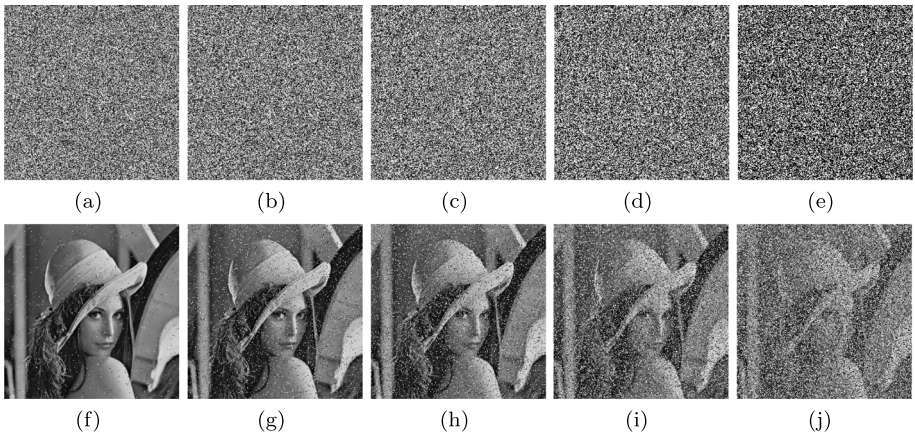


Fig. 9 Salt and pepper noise attack on Lena cipher image: (a) Cipher image with 1% noise, (b) Cipher image with 5% noise, (c) Cipher image with 10% noise, (d) Cipher image with 25% noise, (e) Cipher image with 50% noise, (f)-(j) Decrypted images of corresponding cipher noise images

In general Number of Pixels Change Rate ($NPCR$) and Unified Averaged Changed Intensity ($UACI$) are utilized to test the encryption algorithm against differential attack. The $NPCR$ and $UACI$ values can be calculated using the following formula:

$$NPCR = \frac{\sum_{i,j} D(i,j)}{M \times N} \times 100\% \tag{10}$$

$$D(i,j) = \begin{cases} 1, \text{ if } R_1(i,j) \neq R_2(i,j); \\ 0, \text{ if } R_1(i,j) = R_2(i,j); \end{cases}$$

$$UACI = \frac{1}{M \times N}\left[\sum_{i,j} \frac{|R_1(i,j) - R_2(i,j)|}{255}\right] \times 100\% \tag{11}$$

Where M is the width, and N is the height of the input image. R_1 is the encrypted image obtained from the plain input image, and R_2 is the encrypted image obtained after modifying the plain input image. It has been reported in the literature that the higher the values of $NPCR$, the better the algorithm is at resisting differential attacks. The theoretically expected value of $NPCR$ and $UACI$ are 99.6094% and 33.4635%, respectively. The values closer to the theoretically expected values suggest that the encryption algorithm is sensitive to changes in pixel intensity values of the plain images. The values of $NPCR$ and $UACI$ for the proposed image encryption method are calculated for a modified plain image. Tables 5 and 6 show the $NPCR$ and $UACI$ values for the proposed image encryption method along with other related research works. It can be noted from Tables 5 and 6 that the mean values of $NPCR$ and $UACI$ are 99.6004 and 33.5320 corresponding to the proposed image encryption method which is very close to the theoretical expected values. It can be concluded that the algorithm is sensitive to changes in pixel intensity values of the image and can resist differential attacks.

4.8 Peak signal to noise ratio (PSNR)

$PSNR$ is used to assess the resilience of the proposed image cryptosystem in terms of randomness introduced in the encrypted image with respect to the plain image. The mathematical

Table 5 Comparison of NPCR values obtained using the proposed encryption algorithm along with related research works

Research work	Lena	Goldhill	Bridge	Camera	Bird	MI1	MI2	MI3	Mean
Proposed Algorithm	99.5895	99.6323	99.6399	99.5987	99.5865	99.6185	99.5987	99.5392	99.6004
Mishra et al. [34]	99.6094	99.5605	99.5255	99.5941	99.6628	**99.6964**	99.6506	99.6429	99.6178
Raghuvanshi et al. [38]	99.6184	99.5613	99.6079	99.5748	99.6169	99.6199	99.5824	99.6184	99.6000
Gao et al. [15]	99.6287	99.6473	99.6146	**99.6542**	99.6304	99.6048	99.6292	99.6368	99.6308
El-Shafai et al. [11]	99.6002	99.6552	99.5911	99.6033	99.6140	99.6307	99.6231	99.5834	99.6126
Zhang et al. [61]	**99.6458**	99.6235	99.6015	99.6145	99.6422	99.6353	99.6429	**99.6521**	**99.6322**
Zheng et al. [62]	99.6033	99.6048	99.5911	99.6094	99.5987	99.6155	99.6048	99.5880	99.6020
Suri and Vijay [45]	99.5926	99.5850	99.6063	99.6063	99.5880	99.6017	99.5941	99.6475	99.6027
Ye and Huang [59]	99.6030	99.6125	99.6424	99.5987	99.5987	98.9136	96.2524	98.0713	98.9116
Yuan et al. [60]	99.6031	**99.6875**	99.6145	99.6458	99.6125	99.6146	99.6452	99.6054	99.6286
Askar et al. [6]	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Teng et al. [46]	99.5926	99.5911	99.6033	99.5972	99.6458	99.6124	**99.6658**	99.6353	99.6179
Li et al. [28]	99.6307	99.6140	99.6277	99.5850	99.6353	99.6017	99.5834	99.6201	99.6122
Erkan et al. [12]	99.6155	99.6109	99.6445	99.6353	99.6429	99.6017	99.5987	99.6307	99.6225
Xian and Wang [55]	99.6017	99.5941	99.6201	99.6277	**99.6521**	99.6063	99.5880	99.6307	99.6151
Lai et al. [26]	99.6216	99.5590	**99.6475**	99.5972	99.6507	99.5392	99.5697	99.5911	99.5970

Table 6 Comparison of UACI values obtained using the proposed encryption algorithm along with related research works

Research work	Lena	Goldhill	Bridge	Camera	Bird	MI1	MI2	MI3	Mean
Proposed Algorithm	33.3225	33.5119	33.5098	33.5342	33.3426	33.5098	**33.9835**	33.5416	**33.5320**
Mishra et al. [34]	**33.6005**	33.2272	33.4668	**33.5990**	33.5490	33.2352	33.4897	**33.7651**	33.4916
Raghuvanshi et al. [38]	33.3755	33.5111	33.2766	33.4134	33.3275	33.4276	33.5246	33.4053	33.4077
Gao et al. [15]	33.4711	33.3773	33.4909	33.4808	33.3611	33.3818	33.4436	33.3781	33.4231
El-Shafai et al. [11]	33.4002	33.4168	33.3954	33.4550	33.4324	33.3477	33.3477	33.5486	33.4180
Zhang et al. [61]	33.5545	33.4461	33.5217	33.5496	33.5432	33.4120	33.3014	33.5953	33.4905
Zheng et al. [62]	33.4711	**33.6560**	33.4519	33.4155	33.2196	33.4910	33.5021	33.4298	33.4546
Suri and Vijay [45]	33.3773	33.4741	33.4588	33.4815	33.0182	33.4370	33.5783	33.5432	33.4211
Ye and Huang [59]	33.3773	33.4711	30.0976	33.4829	33.5202	33.4568	33.5314	33.3477	33.0356
Yuan et al. [60]	33.5202	33.4568	33.5314	33.5432	33.3477	33.4568	33.4909	33.5432	33.4863
Askar et al. [6]	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Teng et al. [46]	33.4441	33.4775	33.4964	33.5427	32.9914	**33.5875**	33.5877	33.6223	33.4687
Li et al. [28]	33.0881	33.1238	32.9774	33.1021	28.4884	33.1394	33.0152	32.9003	32.4793
Erkan et al. [12]	33.4608	33.5202	33.4568	33.5314	**33.5529**	33.3999	33.5072	30.7786	33.1510
Xian and Wang [55]	33.5825	33.3882	**33.5488**	33.3742	33.4674	33.5445	33.4829	33.4464	33.4794
Lai et al. [26]	33.5709	33.3882	33.5463	33.4150	33.4909	33.3720	33.4320	33.4909	33.4633

Table 7 Comparison of PSNR values between plain and cipher images for the proposed approach along with related research works

Images	Bird	Bridge	Lena	Camera	Goldhill	MI1	MI2	MI3	Mean
Proposed algorithm	9.3387	8.8645	8.3636	9.0332	8.5432	8.3298	6.5362	8.5492	8.4448
Mishra et al. [34]	9.3367	8.8366	8.3649	9.0799	8.5414	8.3455	6.5132	8.5082	8.4408
Raghuvanshi et al. [38]	9.3367	8.8366	8.3649	9.0799	8.5414	8.3455	6.5132	8.5082	8.4408
Gao et al. [15]	9.3809	8.8721	8.3713	9.0315	8.6121	8.3530	6.5292	8.5582	8.4635
El-Shafai et al. [11]	9.3399	8.8657	8.3819	9.0355	8.6004	8.3609	6.5364	8.4958	8.4521
Zhang et al. [61]	9.3421	8.8595	8.3811	9.1072	8.6104	8.3536	6.5652	8.5374	8.4696
Zheng et al. [62]	9.3534	8.8300	8.3872	9.0362	8.5495	8.3267	6.5109	8.5338	8.4410
Suri and Vijay [45]	9.3399	8.8146	8.3970	9.0276	8.6048	8.3397	6.5033	8.5384	8.4457
Ye and Huang [59]	9.3399	8.8366	8.3872	9.0362	8.6048	8.0263	5.6404	8.3526	8.2780
Yuan et al. [60]	9.3421	8.8595	8.3811	9.1072	8.6104	8.3268	6.5174	8.5255	8.4587
Askar et al. [6]	10.7968	10.6888	10.9702	9.8704	11.0437	10.5721	9.9517	10.9561	10.6062
Teng et al. [46]	9.2971	8.8543	8.3671	9.0609	8.5638	8.3457	6.5357	8.5559	8.4476
Li et al. [28]	9.4210	8.9904	8.4547	9.1829	8.7993	8.5612	6.9657	8.7824	8.6447
Erkan et al. [12]	9.3037	8.8506	8.3883	9.0416	8.5947	8.3197	6.5064	8.5600	8.4456
Xian and Wang [55]	9.2896	8.8355	8.3974	9.0240	8.5598	8.3291	6.4731	8.5260	8.4293
Lai et al. [26]	8.8099	9.0529	8.5382	8.5537	8.5947	8.3457	6.5357	8.7800	8.4014

formula to compute $PSNR$ is shown in (12).

$$PSNR = 10\log_{10}\left(\frac{(2^{n-1})^2}{MSE}\right) \tag{12}$$

MSE denotes the mean square error and is calculated as follows:

$$MSE = \frac{1}{M \times N}\sum_{x=1}^{M}\sum_{y=1}^{N}\left[P(x,y) - E(x,y)\right]^2$$

where M and N are the width and height of the image, and P and E are plain image and the encrypted image, respectively. It is suggested that plain and encrypted images have low $PSNR$ values whereas plain and decrypted images have high $PSNR$ values [11]. Table 7 exhibits the obtained $PSNR$ values for plain and cipher images for the proposed image encryption method along with other related methods. From Table 7, it is observed that $PSNR$ values are small which shows the efficacy of the proposed image encryption method.

4.9 Key sensitivity analysis

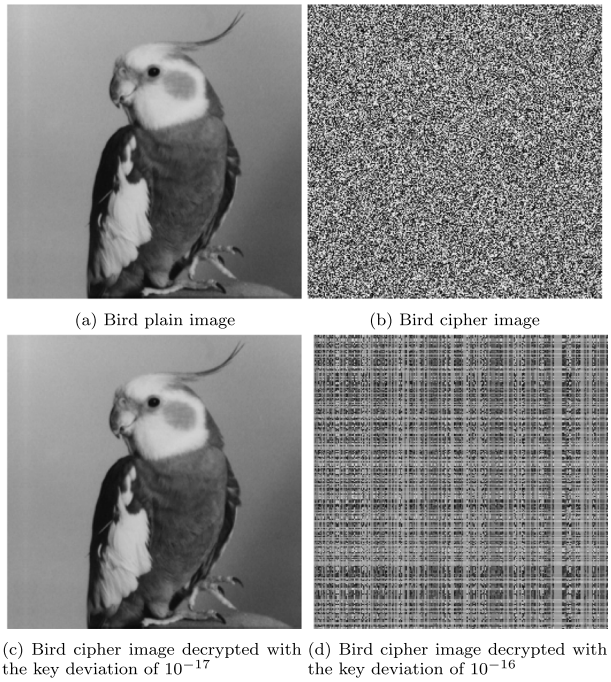
The image encryption algorithm is said to be sensitive to the keys if a small change in the keys does not lead to successful decryption of the cipher image [31]. Hence, the algorithm can resist chosen plaintext or ciphertext attacks. The higher the key sensitivity, the better the algorithm is against chosen plaintext or ciphertext attacks. Initially, the encrypted image was successfully decrypted using the original encryption keys. The decrypted images are shown in Fig. 4. Further, a key deviation of 10^{-16} was introduced in the decryption keys, and the modified keys were used to decrypt the image. It was observed that the decryption process failed for the modified keys. The decrypted image is shown in Fig. 10(d). Therefore, we can conclude that the proposed scheme has high key sensitivity.

4.10 Randomness test

An ideal cipher image must possess a good amount of randomness in the distribution of the pixel intensity values, which is closely related to the algorithm's ability to survive statistical attacks. The National Institute of Standards and Technology (NIST) SP800-22 is a developed tool for evaluating the randomness of a given sequence by running fifteen distinct types of randomization trials [8]. The test uses a p-value at a significance level β. The sequence passes the test whenever $p - value \geq \beta$ is considered random [8]. Table 8 shows the $p - value$ obtained corresponding to 15 tests for the Lena cipher image for $\beta = 0.01$. It can be noted from Table 8 that the proposed image encryption algorithm successfully passed all the tests.

4.11 Structural similarity index measure(SSIM)

This measure calculates the degree of similarity between plain and encrypted images [11]. The mathematical formula to calculate $SSIM$ is described in (13). The $SSIM$ should lie in the range $[-1, 1]$. We have calculated $SSIM$ values for the proposed image encryption algorithm along with other related methods. The $SSIM$ values obtained between plain and encrypted images are quoted in Table 9. The average value of $SSIM$ is 0.0090, which is very close to zero. Hence, it can be concluded that the encrypted images have no similarity with



(a) Bird plain image (b) Bird cipher image

(c) Bird cipher image decrypted with the key deviation of 10^{-17} (d) Bird cipher image decrypted with the key deviation of 10^{-16}

Fig. 10 Key sensitivity analysis results of the proposed encryption algorithm

plain images.

$$SSIM = \frac{(2\mu_I\mu_D + C_1)(2\sigma_{ID} + C_D)}{(\mu_I^2 + \mu_D^2 + C_1)(\sigma_I^2 + \sigma_D^2 + C_2)} \tag{13}$$

Table 8 NIST SP 800-22 test results

Test Name	p-value	result
Frequency Test	0.9230	pass
Block Frequency Test	0.7769	pass
Run Test	0.6585	pass
Run Test (Longest Run of Ones)	0.8050	pass
Binary Matrix Rank Test	0.5222	pass
Discrete Fourier Transform (Spectral) Test	0.2199	pass
Non-overlapping Template Matching Test	0.0751	pass
Overlapping Template Matching Test	0.2742	pass
Universal Statistical Test	0.1950	pass
Linear Complexity Test	0.7416	pass
Serial Test	0.5674	pass
Approximate Entropy Test	0.7604	pass
Cumulative Sums	0.9401	pass
Random Excursion Test	0.8769	pass
Random Excursion Variant Test	0.9599	pass

Table 9 Comparison of SSIM values between plain and cipher images for the proposed approach along with related research works

Images	Bird	Bridge	Lena	Camera	Goldhill	MI1	MI2	MI3	Mean
Proposed algorithm	0.0099	0.0113	0.0077	0.0102	0.0098	0.0088	0.0055	0.0087	0.0090
Mishra et al. [34]	0.0107	0.0072	0.0083	0.0074	0.0075	0.0096	0.0056	0.0084	0.0081
Raghuvanshi et al. [38]	0.0107	0.0072	0.0083	0.0074	0.0075	0.0096	0.0056	0.0084	0.0081
Gao et al. [15]	0.0100	0.0107	0.0093	0.0066	0.0093	0.0082	0.0052	0.0080	0.0084
El-Shafai et al. [11]	0.0095	0.0116	0.0101	0.0052	0.0085	0.0084	0.0050	0.0079	0.0083
Zhang et al. [61]	0.0098	0.0086	0.0087	0.0079	0.0122	0.0081	0.0048	0.0096	0.0087
Zheng et al. [62]	0.0107	0.0056	0.0079	0.0070	0.0109	0.0077	0.0056	0.0071	0.0078
Suri and Vijay [45]	0.0095	0.0080	0.0105	0.0082	0.0112	0.0085	0.0053	0.0089	0.0088
Ye and Huang [59]	0.0086	0.0082	0.0052	0.0080	0.0093	0.0076	0.0044	0.0082	0.0074
Yuan et al. [60]	0.0098	0.0072	0.0087	0.0074	0.0122	0.0087	0.0048	0.0083	0.0084
Askar et al. [6]	0.0255	0.0350	0.0273	0.0245	0.0262	0.0212	0.0130	0.0218	0.0243
Teng et al. [46]	0.0098	0.0100	0.0064	0.0054	0.0109	0.0077	0.0045	0.0087	0.0079
Li et al. [28]	0.0100	0.0064	0.0096	0.0095	0.0088	0.0090	0.0055	0.0096	0.0086
Erkan et al. [12]	0.0089	0.0084	0.0098	0.0096	0.0098	0.0074	0.0049	0.0084	0.0084
Xian and Wang [55]	0.0082	0.0067	0.0072	0.0094	0.0074	0.0079	0.0054	0.0079	0.0075
Lai et al. [26]	0.0069	0.0082	0.0066	0.0103	0.0072	0.0048	0.0084	0.0108	0.0079

4.12 Resistance to classic attacks

According to the Kerckhoff principle, the procedure for encryption and decryption processes is usually known. Due to this, the overall focus should be on the cryptanalysis of these methods. The attacker always tries to obtain the secret key used in the encryption and decryption processes. If the privacy of the key is breached, then the confidentiality of the data is compromised. Typically, there are four types of attacks: known-plaintext attack, ciphertext-only attack, chosen-plaintext attack, and chosen-ciphertext attack. Since the initial values for the TM are obtained from the plain image, a subtle change in the plain image gives a completely different TM sequence and hence a different cipher image. Thus, the algorithm can resist the chosen-plaintext attack which is the strongest among the four mentioned attacks. Hence, it can resist all classical attacks.

4.13 Time complexity analysis

The algorithm's time complexity is examined. Due to the symmetric design of the method, the steps for encryption and decryption are mutually inverse, thus we analyze the complexity of the encryption process to establish the overall complexity. The proposed encryption method is separated into two stages: confusion, and diffusion. We may deduce from the preceding process analysis in the algorithm that the time complexity of pixel confusion and diffusion is $O(M \times N)$. As a result, the complexity of the encryption technique is $O(M \times N)$, where M and N are the image width and height, respectively. Further, the processing time for the proposed encryption algorithm and other methods is calculated. For this experiment, the plain Lena image is utilised and the encryption time of the Lena image is calculated which is shown in Table 10. The table indicates that the technique is efficient in terms of time consumption and can quickly turn the original image into an unrecognisable one. Further, the proposed encryption algorithm is suitable for real-world applications.

Table 10 Comparison of processing time for the proposed approach along with related research works

Research work	Time(in sec.)
Proposed Algorithm	0.4062
Ye and Huang [59]	0.5781
Zhang et al. [61]	8.7813
Raghuvanshi et al. [38]	3.6875
Yuan et al. [60]	0.6718
Askar et al. [6]	0.5156
Gao et al. [15]	2.9062
Zheng et al. [62]	5.5625
Suri and Vijay [45]	6.3906
Mishra et al. [34]	5.6875
El-Shafai et al. [11]	4.3751
Teng et al. [46]	0.4375
Li et al. [28]	2.5313
Erkan et al. [12]	0.2813
Xian and Wang [55]	0.0781
Lai et al. [26]	0.6562

5 Conclusion

In this paper, Tinkerbell chaotic map (TM), Linear Feedback Shift Register (LFSR), and Logistic Map (LM) are used to propose an image encryption algorithm for the secure transmission of the images. The initial values for the TM are derived from the input image. A two-level permutation of the image pixels is performed by the indices of the sequences obtained from TM. Two-level permutation has lowered statistical information among pixels, allowing them to resist statistical attacks. In the second step, LFSR is used to diffuse the pixels of the image. In the other step of diffusion, LM sequences are used. Two-level diffusion has made the algorithm sensitive to changes in plaintext. This minimises plaintext attacks and ciphertext attacks. To validate the performance of the proposed image encryption approach, several performance measures are calculated on a variety of image datasets, including medical images. Also, the proposed method is compared with state-of-the-art methods in terms of several quantitative and qualitative measures. Key space and key sensitivity analysis are carried out, which indicates the proposed approach utilises a huge key space, which rules out the chances of brute force attack. It is resistant to several statistical attacks, differential attacks, plaintext attacks, ciphertext attacks, and brute-force attacks. Furthermore, it is highly sensitive to the input images. A small change in the intensity value of the input image will result in a completely different cipher image. The cipher images have a low adjacent pixel correlation. The proposed image encryption algorithm consumes less CPU processing time in comparison to other related research work that validates the applicability of our proposed approach for real-time applications. Additionally, it suggests that, given the task of encrypting a huge volume of images, the proposed algorithm has the capability of encrypting all the images at a much faster rate as compared to past research. The proposed algorithm is secure and efficient, and it can resist many attacks, which makes it a suitable choice for encryption purposes. The efficiency of the proposed algorithm may decline when dealing with very large images, as the computational complexity could substantially increase. Moreover, the algorithm's security heavily depends on the quality of the chaotic maps used. In the future, we plan to explore a wide range of chaotic maps with diverse features to improve security. Additionally, we aim to develop an encryption algorithm that can effectively handle a considerable amount of data.

Author Contributions There are equal contributions to this research from all the authors of this article.

Funding details The first author received financial support from CSIR, India, with sanction order no. (08/0133(13253)/2022-EMR-I) for conducting this research work.

Data availability statements Image data utilised for analysis of the proposed approach, along with related methods, is publicly available.

Declarations

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent Informed consent was obtained from all individual participants included in the study.

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

References

1. (2023) Fractal coding and analysis group. https://links.uwaterloo.ca/Repository.html
2. (1999) Medical image samples. https://barre.dev/medical/samples/
3. Alexan W, Chen YL, Por LY, Gabr M (2023) Hyperchaotic maps and the single neuron model: a novel framework for chaos-based image encryption. Symmetry 15(5):1081
4. Alvarez G, Li S (2006) Some basic cryptographic requirements for chaos-based cryptosystems. Int J Bifurc Chaos 16(08):2129–2151
5. Amina Y, Bekkouche T, Daachi MEH, Diffellah N (2023) A novel trigonometric 3D chaotic map and its application in a double permutation-diffusion image encryption. Multimed Tools Appl 1–24
6. Askar S, Karawia A, Alshamrani A (2015) Image encryption algorithm based on chaotic economic model. Math Probl Eng 2015. https://doi.org/10.1155/2015/341729
7. Balasamy K, Shamia D (2023) Feature extraction-based medical image watermarking using fuzzy-based median filter. IETE J Res 69(1):83–91
8. Cao LC, Luo YL, Qiu SH, Liu JX (2015) A perturbation method to the tent map based on lyapunov exponent and its application. Chin Phys B 24(10):100501
9. Cao W, Mao Y, Zhou Y (2020) Designing a 2D infinite collapse map for image encryption. Sig Process 171:107457
10. Chen Y, Tang C, Ye R (2020) Cryptanalysis and improvement of medical image encryption using high-speed scrambling and pixel adaptive diffusion. Sig Process 167:107286
11. El-Shafai W, Khallaf F, El-Rabaie ESM, El-Samie FEA (2021) Robust medical image encryption based on DNA-chaos cryptosystem for secure telemedicine and healthcare applications. J Ambient Intell Humaniz Comput 12(10):9007–9035. https://doi.org/10.1007/s12652-020-02597-5
12. Erkan U, Toktas A, Toktas F, Alenezi F (2022) 2D eπ-map for image encryption. Inf Sci 589:770–789. https://doi.org/10.1016/j.ins.2021.12.126
13. Fan S, Chen K, Tian J (2023) A novel image encryption algorithm based on coupled map lattices model. Multimed Tools Appl 1–16
14. Fridrich J (1998) Symmetric ciphers based on two-dimensional chaotic maps. Int J Bifurc Chaos 8(06):1259–1284. https://doi.org/10.1142/S021812749800098X
15. Gao Z, Liu Z, Wang L (2021) An image encryption algorithm based on the improved sine-tent map. Discret Dyn Nat Soc 2021:1–16
16. Ge S, Xia Z, Fei J, Tong Y, Weng J, Li M (2023) A robust document image watermarking scheme using deep neural network. Multimed Tools Appl 1–24
17. Goldsztejn A, Hayes W, Collins P (2011) Tinkerbell is chaotic. SIAM J Appl Dyn Syst 10(4):1480–1501
18. Guan ZH, Huang F, Guan W (2005) Chaos-based image encryption algorithm. Phys Lett A 346(1–3):153–157
19. Han X, Mou J, Liu T, Cao Y (2021) A new fractional-order 2D discrete chaotic map and its DSP implement. Eur Phys J Spec Top 230(21):3913–3925. https://doi.org/10.1140/epjs/s11734-021-00331-6
20. Hua Z, Yi S, Zhou Y (2018) Medical image encryption using high-speed scrambling and pixel adaptive diffusion. Sig Process 144:134–144
21. Huang S, Huang L, Cai S, Xiong X, Liu Y (2022) Novel and secure plaintext-related image encryption algorithm based on compressive sensing and tent-sine system. IET Image Process. https://doi.org/10.1049/ipr2.12429
22. Huang Y, Zhou L (2023) A hyper-chaos-based image encryption scheme with double parity alternate scrambling. Multimed Tools Appl 1–15
23. Joshi AB, Kumar D, Mishra D, Guleria V (2020) Colour-image encryption based on 2D discrete wavelet transform and 3D logistic chaotic map. J Mod Opt 67(10):933–949
24. Kanwal S, Inam S, Hajjej F, Cheikhrouhou O, Nawaz Z, Waqar A, Khan M et al (2022) A new image encryption technique based on sine map, chaotic tent map, and circulant matrices. Secur Commun Netw 2022
25. Kumar CM, Vidhya R, Brindha M (2022) An efficient chaos based image encryption algorithm using enhanced thorp shuffle and chaotic convolution function. Appl Intell 52(3):2556–2585
26. Lai Q, Hu G, Erkan U, Toktas A (2023) A novel pixel-split image encryption scheme based on 2D salomon map. Expert Syst Appl 213(118):845
27. Lai Q, Wan Z, Kamdem Kuate PD (2020) Modelling and circuit realisation of a new no-equilibrium chaotic system with hidden attractor and coexisting attractors. Electron Lett 56(20):1044–1046
28. Li L, Luo Y, Qiu S, Ouyang X, Cao L, Tang S (2022) Image encryption using chaotic map and cellular automata. Multimed Tools Appl pp. 1–19. https://doi.org/10.1007/s11042-022-12621-9
29. Li M, Xu M, Luo J, Fan H (2019) Cryptanalysis of an image encryption using 2d henon-sine map and DNA approach. IEEE Access 7:63336–63345

30. Liu L, Miao S (2017) An image encryption algorithm based on Baker map with varying parameter. Multimed Tools Appl 76(15):16511–16527. https://doi.org/10.1007/s11042-016-3925-x
31. Man Z, Li J, Di X, Sheng Y, Liu Z (2021) Double image encryption algorithm based on neural network and chaos. Chaos, Solitons Fractals 152:111318
32. Matthews R (1989) On the derivation of a "chaotic" encryption algorithm. Cryptologia 13(1):29–42. https://doi.org/10.1080/0161-118991863745
33. Melman A, Evsutin O (2023) Image data hiding schemes based on metaheuristic optimization: a review. Artif Intell Rev 1–73
34. Mishra P, Bhaya C, Pal AK, Singh AK (2021) A medical image cryptosystem using bit-level diffusion with DNA coding. J Ambient Intell Humaniz Comput 1–22. https://doi.org/10.1007/s12652-021-03410-7
35. Moumen A (2018) Medical and biological image analysis. In: Medical and biological image analysis. IntechOpen
36. Sx Nan, Feng Xf Wu, Yf Zhang H (2022) Remote sensing image compression and encryption based on block compressive sensing and 2D-LCCCM. Nonlinear Dyn 108(3):2705–2729
37. Pak C, Huang L (2017) A new color image encryption using combination of the 1D chaotic map. Sig Process 138:129–137
38. Raghuvanshi KK, Kumar S, Kumar S, Kumar S (2021) Development of new encryption system using Brownian motion based diffusion. Multimed Tools Appl 80(14):21011–21040. https://doi.org/10.1007/s11042-021-10665-x
39. Rohith S, Bhat KNH, Sharma AN (2014) Image encryption and decryption using chaotic key sequence generated by sequence of logistic map and sequence of states of linear feedback shift register. In: 2014 International conference on advances in electronics computers and communications, pp 1–6
40. Salomon R (1996) Re-evaluating genetic algorithm performance under coordinate rotation of benchmark functions: a survey of some theoretical and practical aspects of genetic algorithms. BioSystems 39(3):263–278
41. Sharma M, Ranjan RK, Bharti V (2023) An image encryption algorithm based on a novel hyperchaotic Henon sine map. Multimed Tools Appl 82(8):11949–11972
42. Signing VF, Tegue GG, Kountchou M, Njitacke Z, Tsafack N, Nkapkop J, Etoundi CL, Kengne J (2022) A cryptosystem based on a chameleon chaotic system and dynamic DNA coding. Chaos, Solitons Fractals 155(111):777. https://doi.org/10.1016/j.chaos.2021.111777
43. Smid M, Branstad D (1988) Data encryption standard: past and future. Proc of the IEEE 76(5):550–559. https://doi.org/10.1109/5.4441
44. Sun J (2021) 2D-SCMCI hyperchaotic map for image encryption algorithm. IEEE Access 9:59313–59327
45. Suri S, Vijay R (2019) A synchronous intertwining logistic map-DNA approach for color image encryption. J Ambient Intell Humaniz Comput 10(6):2277–2290. https://doi.org/10.1007/s12652-018-0825-0
46. Teng L, Wang X, Xian Y (2022) Image encryption algorithm based on a 2D-CLSS hyperchaotic map using simultaneous permutation and diffusion. Inf Sci 605:71–85. https://doi.org/10.1016/j.ins.2022.05.032
47. Vaidya SP (2022) Fingerprint-based robust medical image watermarking in hybrid transform. Vis Comput 1–16. https://doi.org/10.1007/s00371-022-02406-4
48. Vincent R, Joan D (2001) Advanced encryption standard. In: Proceedings of federal information processing standards publications, national institute of standards and technology, pp 19–22
49. Wang H, Xiao D, Chen X, Huang H (2018) Cryptanalysis and enhancements of image encryption using combination of the 1D chaotic map. Sig Process 144:444–452
50. Wang M, Wang X, Zhao T, Zhang C, Xia Z, Yao N (2021) Spatiotemporal chaos in improved cross coupled map lattice and its application in a bit-level image encryption scheme. Inf Sci 544:1–24. https://doi.org/10.1016/j.ins.2020.07.051
51. Wang X, Feng L, Zhao H (2019) Fast image encryption algorithm based on parallel computing system. Inf Sci 486:340–358
52. Wang X, Gao S (2021) A chaotic image encryption algorithm based on a counting system and the semi-tensor product. Multimed Tools Appl 80(7):10301–10322. https://doi.org/10.1007/s11042-020-10101-6
53. Wang Z, Feng G, Wu H, Zhang X (2023) Data hiding during image processing using capsule networks. Neurocomputing 537:49–60
54. Wu J, Liao X, Yang B (2018) Image encryption using 2d hénon-Sine map and DNA approach. Sig Process 153:11–23
55. Xian Y, Wang X (2021) Fractal sorting matrix and its application on chaotic image encryption. Inf Sci 547:1154–1169. https://doi.org/10.1016/j.ins.2020.09.055

56. Xiao D, Liao X, Wei P (2009) Analysis and improvement of a chaos-based image encryption algorithm. Chaos, Solitons Fractals 40(5):2191–2199
57. Yang YG, Wang BP, Yang YL, Zhou YH, Shi WM, Liao X (2021) Visually meaningful image encryption based on universal embedding model. Inf Sci 562:304–324. https://doi.org/10.1016/j.ins.2021.01.041
58. Yavuz E, Yazıcı R, Kasapbaşı MC, Yamaç E (2016) A chaos-based image encryption algorithm with simple logical functions. Comput Electr Eng 54:471–483
59. Ye G, Huang X (2016) A secure image encryption algorithm based on chaotic maps and SHA-3. Secur Commun Netw 9(13):2015–2023. https://doi.org/10.1002/sec.1458
60. Yuan HM, Liu Y, Gong LH, Wang J (2017) A new image cryptosystem based on 2D hyper-chaotic system. Multimed Tools Appl 76(6):8087–8108. https://doi.org/10.1007/s11042-016-3454-7
61. Zhang W, Wang S, Han W, Yu H, Zhu Z (2020) An image encryption algorithm based on random hamiltonian path. Entropy 22(1):73. https://doi.org/10.3390/e22010073
62. Zheng J, Luo Z, Zeng Q (2020) An efficient image encryption algorithm based on multi chaotic system and random DAN coding. Multimed Tools Appl 79(39):29901–29921. https://doi.org/10.1007/s11042-020-09454-9

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

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

Cluster Computing (2023) 26:3473–3501
https://doi.org/10.1007/s10586-023-04144-1

Examining deep learning's capability to spot code smells: a systematic literature review

Ruchika Malhotra[1] · Bhawna Jain[1,2] · Marouane Kessentini[3]

Received: 20 July 2023 / Revised: 26 August 2023 / Accepted: 1 September 2023 / Published online: 3 October 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract
Code smells violate software development principles that make the software more prone to errors and changes. Researchers have developed code smell detectors using manual and semi-automatic methods to identify these issues. However, three key challenges have limited the practical use of these detectors: developers' subjective perceptions of code smells, lack of consensus in the detection process, and difficulty in setting appropriate detection thresholds. While code smell detection using machine learning has progressed significantly, there still appears to be a gap in understanding the effective utilization of deep learning (DL) approaches. This paper aims to review and identify current methods for code smell detection using DL techniques. A systematic literature review is conducted on 35 primary studies from a collection of 8739 publications between 2013 and the present. The analysis reveals that common code smells detected include Feature Envy, God Classes, Long Methods, Complex Classes, and Large Classes. The most popular DL algorithms used are Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN), often combined with other techniques for better results. Algorithms that train models on large datasets with fewer independent variables demonstrate exemplary performance. The paper also highlights open issues and provides guidelines for future metric identification and selection research.

Keywords Antipatterns · Code smells · Deep learning · Software smells · Smell detection

1 Introduction

With the advent of technological advancements in the industry, software maintenance is becoming a crucial phase of the software development life cycle. Software systems are changed to fix bugs, enhance base features or add new requirements [1]. But due to time constraints and early deliverables, these changes are frequently performed carelessly, resulting in the introduction of "technical debt" in the code and degrading the system's original design. Technical debt refers to the presence of "antipatterns or smells" in the software. Generally, smells causes problems in two sections of development- either at the code part or the design part. The former refers to a problem with code writing and formulation [2], commonly known as "code smells", while the latter refers to issues with the designing of the software itself and is termed as model smells or design flaws [3]. Amongst the two, *Code Smells* is considered more severe, as they not only worsen developers' comprehension of the source code [4], but their emergence also affects the object classes making them more fault-prone [5, 6]. In 1998, Brown et al. [7] also explained the defects caused due to antipatterns in software maintenance. Table 1 defines a short description of such code smells. Although these defects don't always yield the program to act wrong or crash, they make it more challenging to comprehend and maintain [8]. As a result, they increase the expense and effort for maintainability [9, 10]. According to earlier research, the costs of the maintenance stage range from about 50% to 80% of the entire cost. If technical debts are not managed correctly, they can even lead to the failure of a software project by creating interest that eats up the money intended for new features. As a result, the research

✉ Bhawna Jain
bhawnajain@igdtuw.ac.in

[1] Delhi Technological University, Delhi, India

[2] Indira Gandhi Delhi Technical University for Women, Delhi, India

[3] University of Michigan-Flint College of Innovation & Technology, Michigan, USA

Table 1 Types of Code Smells Detected

Code Smell	**Description**
Alternative Classes with Different Interfaces	A situation where a class might operate with distinct classes but their interfaces differ
Blob	A class or module in software that is too complicated and huge
Brain Class	A class that takes on too many duties, making it difficult to manage and unnecessarily complicates the code
Brain Method	A method that has become overly complicated has too many obligations and needs to be easier to read and manage
Comments	Unnecessary misuse of words to cover up bad code structure
Class Hierarchy Problem	A design with too deep and complicated class hierarchies makes it easier to maintain and rigid
Data Class	Classes that only contain data and no logic
Data Clumps	Items of data that frequently occur together
Divergent Change	Altering one class forces to modify another class
Duplicate Code	Code that performs the same function as another code
Feature Envy	A method that prioritizes attributes and properties from other classes above those from its own
Functional Decomposition	A function is subdivided into small classes with one method each, which completely circumvents the OOP aspects
God Class	A massive class that focuses a lot of duties, manages and supervises a wide variety of objects, and successfully performs every function in the program
Global data	There is no way to identify the specific code that changed it. Therefore, it can be altered from any part of the code
Inappropriate Intimacy	When there is a close coupling between two classes
Incomplete Library Class	When the software uses a faulty library
Insider Trading	Issues with coupling brought on by trading data amongst modules
Large Class	A class with several instance variables and methods that try to perform many things
Lazy Class	An under performing class that has to be eliminated
Lazy Element	Software components that are intended to expand but do not follow the rules of software evolution
Long Method	A method that is lengthy and difficult to comprehend, modify, or extend
Long Parameter List	A lengthy, difficult-to-represent parameter list
Message Chains	A series of requests sent between objects without any additional behavior being added
Middle man	When one class heavily influences another class's behavior
Mutable Data	Data changes can result in unforeseen implications and challenging issues
Mysterious Name	Names that are unimportant and do not accurately describe the software's components
Parallel Inheritance Hierarchies	A circumstance in which two connected and concurrent class hierarchies exist
Primitive Obsession	It represents the use of primitives rather than smelly classes, reducing their value and reusing abilities
Refused Bequest	A child class that does not purely work and supports the implementation of its parent class
Shotgun Surgery	When altering a class necessitates modifying all other classes simultaneously
Spaghetti Code	Overly complex and poorly organized source code with a lack of modularization, insufficient documentation, and an unsteady control flow
Speculative Generality	When extra code is added to the software in anticipation of future developments
Switch Statements	Polymorphism is replaced by the use of type codes or run-time class type detection
Temporary Fields	A variable in the class is utilized only in certain circumstances

community has been particularly active in developing various techniques to detect such smells.

With the growing need to detect code smell, research is being conducted in this domain in different parts of the world. Researchers across the globe work extensively in this field, with the United States being at the top. These researchers mentioned a series of solutions for detecting code smells. The contribution from the top 9 active countries in this domain is shown in Fig. 1. Early solutions to detection problems included manual detection and removal, which was time-consuming and resource-draining. Hence, new approaches have been developed to resolve this issue effectively.

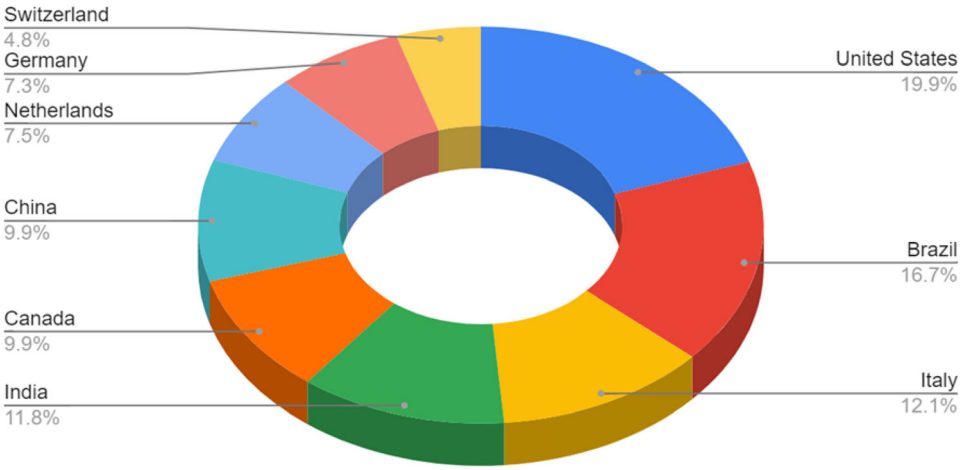


Fig. 1 Contribution of top 9 active countries in code smell research

The current boom in the acceptance and usage of deep learning (DL) techniques has aided in the advancement of software development. However, attempts to use it to enhance software maintainability, a crucial quality trait for large-scale projects, have yet to show conflicting outcomes. One of the problems found, for instance, indicates that each study proposes a new model from scratch without building upon or using the current models [11, 12]. Likewise, the complexity of software has grown, expanding the amount of data that can be evaluated, but evaluation measures still need to be put in place for them.

Taking into consideration the amount of research that has been done in code smells detection techniques, few attempts have been made to review the trends of the confluence between DL approaches along with code smell and the methods used to create models to predict code smell, to the extent that it has been explored over the past ten years [13–15]. Considering the abovementioned gaps, this paper aims to conduct a Systematic Literature Review (SLR) on how DL techniques can effectively detect code smells. This SLR targets articles published in the past ten years (Jan, 2013 - July, 2023). The goal of this SLR is to: a) identify the types of code smells researchers use, b) assess the training methods employed to develop and assess DL approaches, c) acknowledge the metrics that impact the evaluation of the approach, and d) performance analysis of the efficiency of the current DL models. This study will thus make it easier for the community to make research-based judgments by taking into account any potential gaps or areas with a plethora of studies that might be combined to produce certainties.

Major contributions of this SLR are as follows:

1. *Latest Primary Studies on Deep Learning*: A collection of 35 primary studies introducing DL models for predicting code smells has been identified, serving as a foundation for expanding knowledge in this field.
2. *DL Methods and Models*: *Study Highlights and Details*: Thorough consolidation of the primary studies has been presented, covering four key aspects: (i) types of code smells addressed, (ii) DL approaches utilized in the experiments, (iii) evaluation strategies employed in the studies, and (iv) performance analysis of the models proposed.
3. *Metric Representation and Future Benchmark*: This paper summarizes metric representation for detecting code smells. It describes the minimal software metrics necessary for identifying particular code smells and their threshold values. This can serve as a benchmark of rules for detection techniques to be developed in the future.
4. *Guidelines and Recommendations for Future Research*: Based on the results, guidelines and recommendations have been offered to facilitate future research endeavors in this domain.

The rest of the paper is organized as follows: Sect. 2 explains the related work in this domain and the research gaps that have not been considered to date. Adopted research methodology, including search strategy, Research Question (RQ) finalization, and other essential criteria required for conducting this review, is described in Sect. 3; Analysis and Results of the research questions are described in Sect. 4; Sect. 5 presents the discussions and the research gaps identified along with potential future works. Sect. 6 describes the open issues in this domain and opens new horizons for researchers to work in this domain. Section 7 provides the threats to validity for this SLR, followed by Sect. 8 concluding the paper.

2 Related work

Code smells are frequent issues in software engineering, resulting in worse code quality, more expensive maintenance, and slower program performance. In the late 1990 s, Beck et al. [16], the inventor of extreme programming, referred to the importance of design quality while creating software and popularised the phrase "Code Smells" to characterize the implementation and design defects found in source code. These code smells were described more in detail by Fowler as described in Table 1 and established "refactoring" techniques to solve such smells. Mantyla [17] categorized the code smells identified by Fowler into seven broad categories as shown in Fig. 2.

Kessentini et al., in their research [18], explained how code smells in design have emerged. Although the number and frequency of code smell in design are pretty low compared to the code smell level in source code, it still threatens the development and maintenance part of the design if code smells are not addressed.

Liu et al. in [19] examined potential issues in code from the idea of cleaning heuristics, or what is required in terms of etiquette, best practices, and discipline to maintain clean code. The list of code smells also included testing, the building environment, error handling, formatting, and element names.

Additionally, Table 2 describes the new code smells given by Wake et al. [20] in 2003. These code smells have been found in principle with more mathematical computation.

Recent research has highlighted the widespread adoption of copy-and-paste programming, leading to problematic code duplication. To combat this, Pascarella et al. [21] innovatively introduced a Siamese architecture employing identical Convolutional Neural Networks (CNNs) specifically designed to identify code clones. This involves representing code fragments as Abstract Syntax Trees (ASTs) and using dedicated CNN-based subnetworks within the architecture to extract distinctive features from paired ASTs meticulously. The final output layer adeptly gauges the similarity or dissimilarity between these code segments. Remarkably, experimental results strongly underscore the efficacy of this CNN-driven feature extraction approach in robustly detecting code clones, spanning both source code and bytecode representations.

From the articles studied, it is believed that there have been very few SLR to comprehend and summarize the research on DL models predicting code smells. Nevertheless, it's worth noting that certain secondary studies have proposed insights into code smells and tools for detecting them. Table 3 summarises the SLRs conducted in this domain and provides a vision of how this SLR differs from other existing reviews.

Zhang et al., in his SLR [25], covered code smells state-of-the-art from 2000 to 2009 that mainly aimed at (i) code smell that is of utmost attention in the field of research, (ii) design methods used to study and review the algorithms presented, and (iii) goal of the selected research studies (e.g., any empirical observation found or new techniques developed). They also investigated whether any published research provided evidence to back up the adverse effects of code smells.

Alsolai et al. [36] examined DL methods for software maintenance prediction. To thoroughly summarise, analyze, and debate diverse points of view on measurements

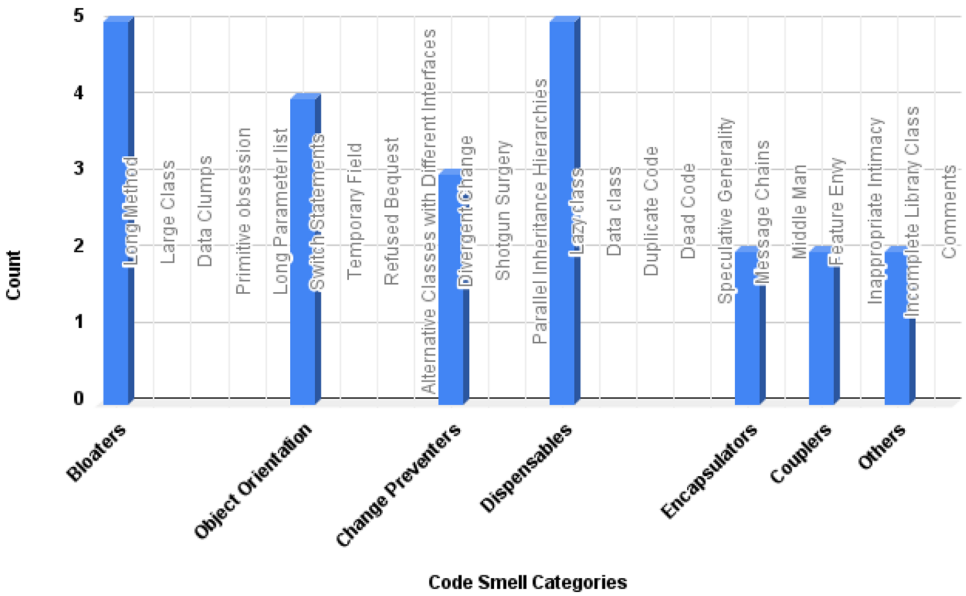


Fig. 2 Code Smell Categorization given by Mantyla

Table 2 Code smells identified by Wake in 2003

Code Smell	Description
Complicated Boolean Expressions	Snippets of code having complex conditions involving operators like NOT, OR, and AND
Dead Code	Having a variable, attribute, or code not utilized elsewhere. It typically happens after a code update with a poor cleanup
Inconsistent Names	The same term is used in several contexts and for various reasons
Null Check	Repeatedly occurring events, checking the object's null values
Magic Numbers	Numbers that are purposefully used in the code and whose values never change
Special Case	Lots of complex conditional statements are used
Type Embedded in Name	This includes commonly used names with redundant definitions like "course.addSubject(subject)" naturally instead of course.add(subject). The usage of Hungarian notation, whose types are opposite to their functions or purposes, was also included in this category
Uncommunicative Names	Names given to software items, often local variables and attributes, need to be more descriptive of their names or intended purposes, such as y or value 0. When applied to methods and classes, it becomes much more crucial

Table 3 List of SLRs Conducted in this Domain

ID	Authors	Research timeline	Target Domain	# of Studies	Ref
ES1	Sobrinho et al.	1990-2017	Knowledge and understanding of code smells	351	[22]
ES2	Arshad and Rasool	1999-2015	Techniques for detecting code smells	46	[23]
ES3	Sabir et al.	2000-2017	Identification of smells in various object-oriented and service-oriented software paradigms	78	[24]
ES4	Zhang et al.	2000-2009	Recent trends in Code smells	39	[25]
ES5	Alshayeb & Misbhauddin	2001-2012	UML Model Refactoring	94	[26]
ES6	Al Dallal	Till 2013	Refactoring code smells	47	[27]
ES7	Azeem et al.	2007-2017	ML Approaches for code smell detection	15	[28]
E8	Caram et al.	1999-2016	Genetic Learning Approaches along with ML for code smell detection	26	[29]
ES9	Abuhassan et al.	Till 2018	Gather and evaluate any methods and techniques that have been utilized to identify both design and code smells	145	[30]
ES10	Singh and Kaur	Till 2018	Refactoring in object Oriented systems	238	[31]
ES11	Agnihotri and Chug	2001-2019	Effect of code smell and Refactoring on systems	68	[32]
ES12	Lewowski and Madeyski	1999-2020	Reproducibility of existing ML techniques	46	[33]
ES13	Alazba et al.	Till Oct 2021	Evaluating DL techniques to detect both - code and design flaws	67	[34]
ES14	Lei et al.	2016-2020	Code Clone Detection using DL	21	[35]
ES15	This study	2013-July 2023	Deep Learning approaches for code smell detection	35	-

required for software maintenance, datasets, result metrics, evaluation measurements, unique algorithms and models, and ensemble models, this study selected and looked into several research problems. The 56 chosen research (reviewed up to 2018) showed relatively limited activity in software maintainability prediction compared to other software quality criteria. The majority of research employed k-fold cross-validation and concentrated on regression issues. Most studies used individual prediction models, while ensemble models were quite a few. In conclusion, ensemble models proved useful for forecasting software maintainability and showed enhanced accuracy in prediction compared to individual models. To increase the accuracy and consistency of the findings, it is necessary to

apply these and other models to a wide range of datasets, even though their use is expected.

In another study, Caram et al. [29] examined DL-based approaches and strategies for finding "code smells" in a software development environment. Genetic algorithms (GAs) were the most frequently utilized methodology, appearing in 22.22% of the publications. There was a significant amount of redundancy in the code smells that each approach handled since several different algorithms covered those smells. The decision tree and then the random forest offered the best average in terms of performance. Five of the twenty-five analyzed smells were handled by machine learning methods. The majority of them concentrated on several code smells, and aside from a few particular smells, there needed to be an approach that performed better overall.

A similar methodology was examined by Azeem et al. in [28]. Even though the research community investigated the process of designing a code smell detection approach using the heuristic method, a need for more understanding is still present about the adoption of DL approaches and how DL is used to identify code smells. In this paper, the authors reviewed the report on parameters of (i) code smell being detected; (ii) the method applied for detection; (iii) parameter creation for review, and (iv) analysis of the results obtained. The investigations conducted as a consequence revealed that the literature heavily considered *God Class, Functional Decomposition, Long Method, and Spaghetti Code* smells. The analysis also identified several problems and difficulties that the scientific community should concentrate on in the future. This analysis is a base for a more concise methodology in this paper.

In their SLR on code smells detection methods, Fernandes et al. [37] focused on publications published between 2000 and 2016. The study aimed to list all code smell detection technologies, their distinguishing characteristics, and the various code smells they can detect. In addition, a comparison of the recall, accuracy, usefulness, and agreement of the four most often used tools was made. Arshad and Rasool [23] also carried out an SLR on the methods and equipment employed for detecting antipatterns from the source code. Modern tools and procedures were categorized based on detection methodologies, and the outcomes were examined. They made some suggestions for researchers and developers working on detecting code smells or antipatterns on the tools to use.

Lei et al. [35] recently conducted a review delving into the application of DL for detecting Code Clones. The scope of their analysis encompassed 21 papers released from 2016 to 2020; however, their focus was exclusively on clones categorized as type 3 and 4. The authors discussed the restrictions, difficulties, and suggestions associated with employing DL techniques in Code Clone detection.

Alazba et al. [34] recently published a study that summarised and synthesized works that employed deep learning (DL) approach for bad smell detection. Their primary studies were selected up to October 2021 and thoroughly reviewed different models used for code smell detection. They examined 67 studies and analyzed quantitative and qualitative data to arrive at their findings.

The studies mentioned above show that very few research studies dealt directly with DL techniques for identifying code smells. Fernandes et al. primarily focused on tools rather than techniques for code smell detection, as no tool was used in implementing a DL model. As a result, they did not evaluate DL-based techniques. Rasool and Arshad focused on detecting approaches and did not consider DL approaches. Alazba et al. analyzed the DL models up to 2021 and summarized each outcome approach for future guidelines.

This research shows the promise of DL methods for software engineering's code smell detection. The most recent study indicates that DL strategies efficiently detect "code" smells, enhance code quality, and lessen the time needed for manual code inspection. However, there is still space for improvement. This paper shows that focused analysis should be maintained to advance the understanding of the topic and identify any unresolved issues that call for further research.

3 Research methodology

The following section contains the research strategy to analyze the paper selected for this SLR. It is a structured and rigorous process for locating, evaluating, and examining published literature to examine a particular research issue or phenomenon. In 2007, Kitchenham [38] wrote a paper describing the guidelines for performing systematic literature reviews in software engineering. On the policies of Kitchenhem rules, this section is elaborated for planning, conducting, and reporting the study in a recurring order. All the phases and corresponding steps that should be taken for a systematic literature review according to Kitchenham principles are shown in Fig. 3.

3.1 Planning

In the first stage, known as the planning stage, the search databases were found, Research questions were created, and the pertinent articles on the research topic were retrieved.

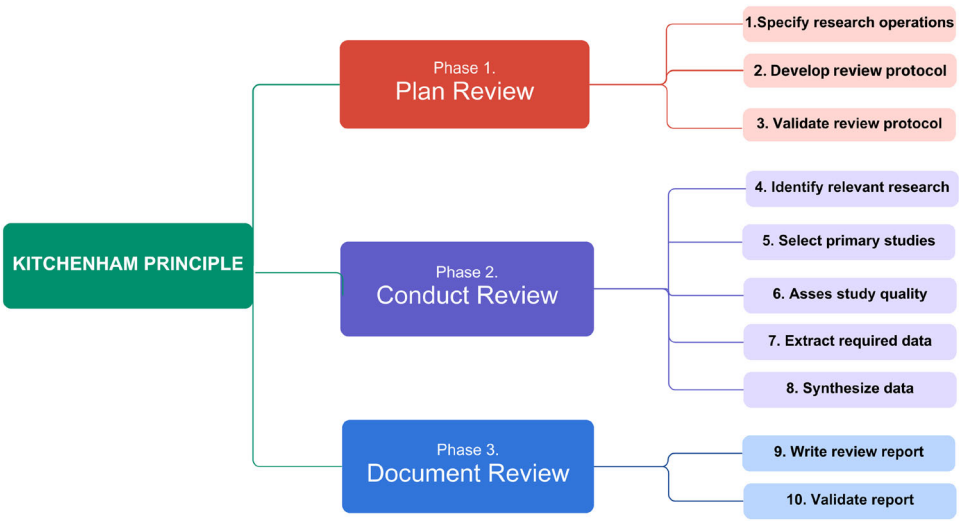


Fig. 3 Kitchenham Strategy for SLR

3.2 Research questions

The primary goal of this comprehensive literature review is to highlight the range of code smells in software systems, their identification, and various code smell detection approaches. Table 4 describes the research questions formulated for this systematic review.

Table 4 Research questions imposed on this systematic literature review

Research Questions	Motivation
RQ 1: Code smells considered	
RQ 1.1: What are the code smells currently detected by DL algorithms?	To investigate the most recent advances in DL-based code smell detection by identifying the code smells which have been considered to date
RQ 2: Deep Learning Approaches	
RQ 2.1: Which independent variables give the best result in predicting code smells?	This will help to determine how many variables are necessary for the detection process
RQ 2.2: Which dependent variable is used to get the results and analysis of code smells?	What data type will be the model output, and how can it be enhanced to get more accurate results?
RQ 2.3: Which DL algorithms are utilized in code smell detection?	This will help developers and researchers choose the most efficient DL algorithm for code smell detection
RQ 2.4: What detecting tools are most commonly used in research for code smells?	In this strategy, many tools are used to utilize time properly. Investigating these tools can help to innovate more tools
RQ 2.5: What strategies have been adopted for training the models?	Training in the best way possible will give the best results
RQ 3: Evaluation Methods	
RQ 3.1: What datasets are used to detect code smells?	This will help to identify and create an accurately labeled dataset
RQ 3.2: Which result metric is used to evaluate the models?	This will help researchers to compare the results of each tool or algorithm to determine if their metrics are similar
RQ 4: Meta-analysis of Performance	
RQ 4.1: How do independent variables impact performance?	To know if the result varies with the count of variables, making it easier for tools to analyze the dataset
RQ 4.2: Is the performance of code smell prediction models affected by DL algorithms?	How many layer(s) gives the best output and help the model to perform better by trying new combinations and learning from past researchers
RQ 4.3: What is the impact of Training Strategies?	Does within-project training and cross-project training show a significant difference? This will help to know how many datasets are needed to be compiled

3.3 Search strategy

A search method was developed to gather all the published material pertinent to this issue. The search approach included selecting search terms, criteria for selecting articles for the research, a search method, and the relevant sources for articles.

3.3.1 Search term used

To determine the appropriate search terms, a strategy was followed:-

1. Firstly, by identifying the population, the intervention, and the result, a list of key phrases was derived from the study questions. *Result* Regarding the initial stage, the focus was on identifying and structuring the critical terms for the population, intervention, and outcome. The population encompasses code smells and anti-patterns that manifest within the code of software systems. The interventions involve a range of DL techniques, algorithms, tools, and datasets designed for detecting, recognizing, and prognosing code smells. The outcomes encompass a variety of identified code smells through the application of DL algorithms, detection tools, and the metrics employed within the detection process.
2. Alternate spellings and synonyms for all significant terms were looked upon. *Result*
   - *Code Smell* "code smells" OR "code smell" OR "code bad smells" OR "bad code smells" OR "bad smell" OR "anomalies" OR "antipatterns" OR "antipattern" OR "design defect" OR "design-smells" OR "code defect" OR "code flaw" OR "smelly code" OR "design flaw"
   - *Deep Learning* "Deep Learning" OR "Neural Networks" OR "Artificial Neural Networks" OR "Deep Models" OR "Layered Neural Networks" OR "Multi-Layer Perceptron" OR "Deep Architectures" OR " Convolution Neural Network" OR "Hybrid Model" OR "CNN" OR "DL" OR "ANN" OR "AST" OR "RNN"
   - *Prediction*("prediction" OR "detection" OR "identification" OR "prediction model" OR "model"
   - *Software*"Software Maintenance" OR "Software Engineering"
3. Keywords were checked in pertinent papers, and keywords unrelated to code smells were eliminated. *Result* eliminated - "anomalies" OR "antipatterns" OR "design defect" OR "design-smells" OR "design flaw" OR "code bad smells" OR "bad code smells"
4. For the conjunction of two or more cases, AND and OR boolean operators were used, wherein OR operator was used for the concatenation of alternative spellings and synonyms. In contrast, the AND operator was used to concatenate major terms.
5. Lastly, the search string was integrated into the summarized form. The final Search term is mentioned below.

(Software Engineering OR software maintenance) AND (antipattern OR code smell OR bad smell OR code defect OR code flaw OR smelly code OR anti-pattern) AND (deep learning OR CNN OR DL OR ANN OR AST OR RNN) AND (identification OR detection) AND (tool OR method OR technique OR approach)

3.3.2 Databases to be searched

A significant part of an SLR is the choice of appropriate resources to use while looking for relevant material. To find all the material pertinent to this study's concerns, this paper used the following resources: Springer, Google Scholar, Wiley, IEEE Xplore, Elsevier Science Direct, ACM, and Engineering Village.

The desire to compile as many papers as possible to adequately carry out this systematic literature review motivated us to choose these databases. Since these sites have a vast volume of literature - conference proceedings, books, journal articles, etc. related to the paper's objectives, the sources analyzed are regarded as the most representative of code smell detection and are utilized in different SLRs.

3.4 Article selection process

This section will discuss the criteria for narrowing down the number of research articles examined. The method used in this study to select primary studies is shown in Fig. 4. The selection procedure is covered in depth in the following subsections:-

3.4.1 Databases and search terms used

The paper followed four main steps to examine and eliminate papers irrelevant to this study and select which are useful to the research questions:

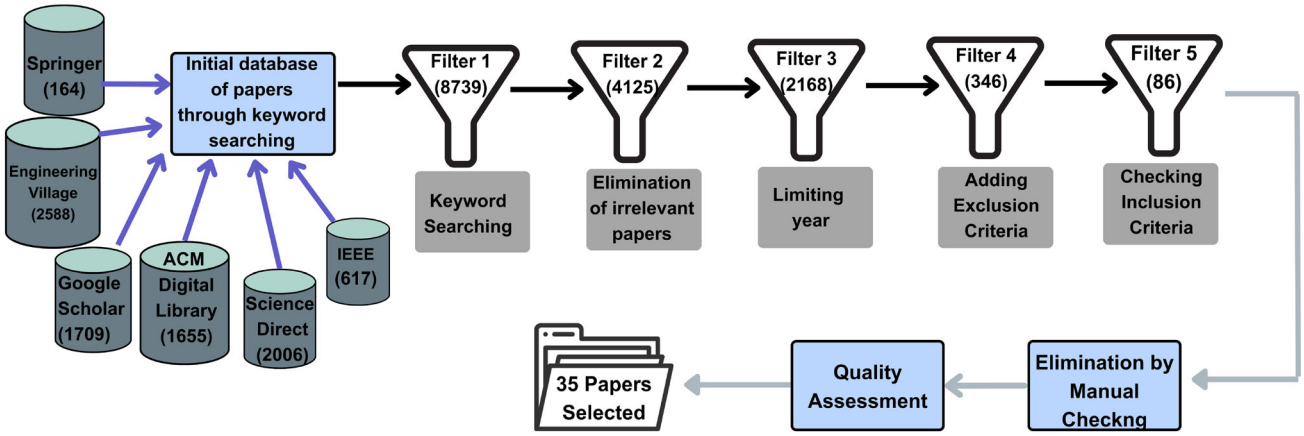


Fig. 4 Article selection process

1. Using the search terms listed in section 3.3.1 and shown in Fig. 5, this paper gathered primary research from the online libraries described in section 3.3.2. No time limit was set for the search procedure so that enough relevant materials could be gathered. Table 5 displays the search results from the digital libraries.
2. In this search criteria, keyword searching played a vital role involving search engines or databases to identify articles, books, or other materials that match specified keywords or phrases after the keyword research stage. Second, this paper used citation searching, or, more precisely, snowballing, to review these papers. The "snowball" strategy is used for searching for potentially lost papers. In particular, "snowballing" involves finding new sources by consulting a paper's references or citations. Both forward and backward snowballing were utilized in this context; the initial one has all the papers cited in the first chosen articles, while the latter has all papers that cited the initially selected papers. These various search methods gave a raw database of 8739 publications, which satisfied the first filter.
3. In the second step, the paper eliminated all the irrelevant papers by checking their relevancy to the topic and the depth of their research. The third layer involved selecting all the papers which were published after 2013. Figure 6 gives the number of papers year-wise written from 2013 till date.
4. For the fourth filter, an exclusion criterion was explicitly designed to match the relevant research questions. The criteria firstly remove all the papers irrelevant to the title of this research. This eliminated unrelated publications from the whole list of retrieved sources using the exclusion criteria (explained in 3.4.2). This stage was carried out by the second author of this research, who examined the 2,168 publications that were located, applied the exclusion criteria, and read their titles, abstracts, and keywords. The articles that did not satisfy these requirements were eliminated, and the remaining papers were picked for the inclusion criteria, which comprised the fifth screening step.
5. Similar to how the exclusion criteria were used, the inclusion criteria were used as the next filter. The databases that met this requirement were preserved, while the others were deleted. Each paper was evaluated against the inclusion criteria by the second author of this work, after which these papers underwent additional filtering.
6. After the automatic detection and removal of irrelevant papers, the authors eventually turned to manual validation and verification to eliminate duplicates, irrelevant papers, or articles that didn't satisfy the paper's requirements. For this elimination phase, all three writers worked to assess the relevant portions of all the publications. From this method, the final search string selection comprised 86 articles (or 0.6 percent of the papers discovered in the previous stage).

Fig. 5 Related Terms for the Review

Table 5 Layer-wise article selection process

Sources	Filter 1	Filter 2	Filter 3	Filter 4	Filter 5	Final Filter
IEEE	617	232	156	20	8	4
Elsevier	164	159	68	21	9	4
Science Direct	2006	732	475	84	13	6
Engineering Village	2588	1311	477	85	26	11
Google Scholar	1709	707	451	89	19	7
ACM	1655	984	541	47	11	3
Total	8739	4125	2168	346	86	35

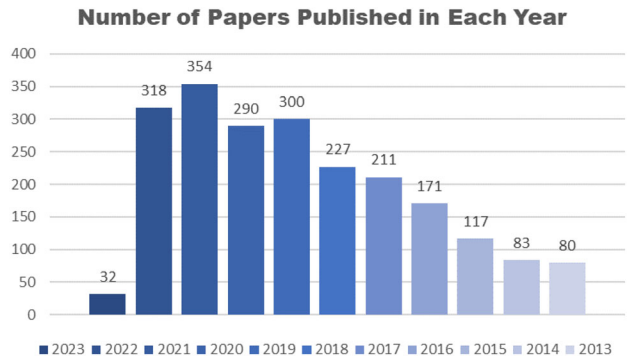


Fig. 6 Year-wise publication in last ten years

7. The 54 pertinent software engineering conferences and 32 journals were all searched manually to ensure all relevant literature was included in the ending filter and added to the final list of chosen sources. More specifically, depending on the exclusion/inclusion criteria, the paper's first author reviewed each of the paper sources and their relevancy that were discovered before deciding whether or not to include them in the final selection or not. To achieve this, she started by looking at the title. If it was obviously outside the scope of the study (for example, if it mentioned automatic test case generation), it was decided to skip the paper. In contrast, if a publication was seen as a potential paper, the second author read the abstract, introduction, and conclusion to confirm its applicability. Finally, 35 studies were chosen from this filter stage.
8. With the final set of papers, the authors did a last recheck, i.e., the quality assessment phase, to double-check if they have relevance to the topic. Table 6 describes the quality assessment criteria formulated for this research. This step resulted in the final set of papers examined in this research. All 35 sources discovered up to this point passed the quality assessment. Section 3.4.4 describes the data extraction procedure.
9. After this long examination, including keyword searching, snowballing, and manual search, **35** papers were selected and surpassed all tests. Table Appendix B describes the list of all 35 selected studies with their titles.

3.4.2 Inclusion and exclusion criteria

A paper has to meet the following criteria to be useful in explaining the research questions.

1. *Exclusion criteria*: For all the layers, the authors put up some criteria for excluding research papers from the total found. Sources that complied with the conditions listed below were disregarded for this analysis:
   - Articles with less emphasis on DL and more on other detecting methods:- other methods like machine learning and manual detection should be removed for effective verification of DL techniques used;
   - Articles authored in languages other than English:- To ensure resource efficiency and language proficiency, a broader scope is adopted;
   - Articles for which no full text is available:- Excluding articles that lack open access, are in progress, consist of abstracts or theses, and eliminate duplicate entries for in-depth examination.
2. *Inclusion criteria*: Paper with significance to DL were to be kept safe from getting lost; hence, apart from excluding sources that complied with the requirements listed above used in this study, a list of inclusion criteria was framed as well:
   - Papers that provide publicly accessible source code and utilize openly available Java projects:- This will enhance the granularity of the dataset assessment process;
   - Articles that present novel methods for enhancing the effectiveness of DL methods currently in use for code smell detection:- Developed a novel approach and described the new method and model;

The paper included all kinds of articles (e.g., journals, proceedings, and conferences) to compile the complete

Table 6 Questions for Quality Assessment

Parameter	Question
Aim	Is the aim of the study specified and relevant to the review?
State of the art	Did the authors compare the results with the State of the art technology?
Dataset	Are the datasets of the corresponding study available publicly?
Accuracy	Can the provided approach accurately identify the code smells?
Variables	What are the precise independent and dependent variables required for the detection of particular code smell ?
Algorithm	Is the DL classifier's definition clear?
Evaluation metrics	Are the evaluation metrics and procedures explicitly disclosed?
Improvement	Are the improvements in quality attributes specified?
Limitations	Are the limitations of the papers explicitly discussed?

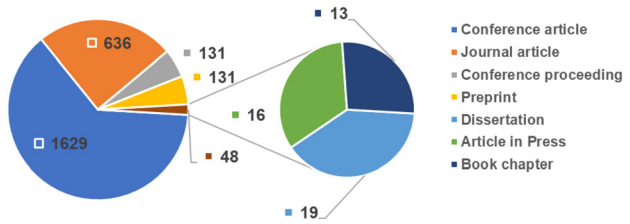


Fig. 7 Article Published in Various Categories

selection of relevant sources. Fig. 7 shows the number of articles in various categories.

3.4.3 Quality assessment

After the last phase of selection, the papers' quality was assessed. The mentioned list ensured the publications' integrity and thoroughness choices. Answers of "Yes," "Partially," or "No" were given to each of the questions mentioned in Table 6. The study was considered lacking when a research's methodological components could have been deduced from the text, even if they were not clearly stated.

All three authors of this paper carried out the approach; they evaluated each paper source following the quality assessment criteria. All three authors simultaneously gave framed answers for each paper on their level. Then, the authors had a mutual discussion to approve all the articles.

This stage was moved through in rounds. Each time a paper was chosen, the researchers' responses to the various quality questions were compared. If there was a discrepancy, the participants had to defend their positions before another round could begin. Up till an agreement was reached, this procedure was repeated.

3.4.4 Data extraction

The data required to address the research questions were extracted after choosing the final publications to be included in the SLR. Table 7 describes the data extraction form. In the data extraction, one column for research gaps was kept in addition to information regarding the specific features under consideration, such as the sorts of code smells taken into account in the chosen publications. As a result, it was easier for new research to understand the limitations of the previous ones. The second author of this work extracted the variables or attributes required to address this study's questions. However, all writers reasoned and examined each manuscript to identify potential limits before filling out the limitation section. All authors then combined the individual considerations.

4 Results

4.1 Overview

Before presenting the findings of the SLR concerning the examined research topic, this section gives a detailed description of the features of the articles that cleared the inclusion, exclusion, and quality criteria.

4.2 Demographics

Table 8 displayed the final list of primary studies examined for this paper, and the columns "Year" and "Publication Type" show how many articles were published in this domain over time in journals and conferences. As can be seen, all of the documents considered were published between 2013 and 2023 (till Date), out of which more than 65% of these primary studies were written after 2019, showing a developing trend in adopting DL approaches.

Table 7 Data extraction form

Dimension	Attribute: Description
Variety of code smells detected	Classification of various code smells being detected in the analyzed papers
Programming Language	The technique used in the paper is an application to which programming language, e.g., Java or C++, etc
DL Algorithms	Which kind and how many layers of each are employed in the study?
Type of Classification	What form of classification, such as multi-valued classification, binary classification, etc., was used to tell the code smell detection result?
Training approach	Which type of internal or external training approach was used?
Evaluation Metrics	What assessment measures, such as Precision, AUC-ROC, Recall, and F Measure, have been used to evaluate the model's accuracy?
Datasets	Which collection of data was utilized for training the model? A list of the software projects to which approaches for code smell detection have been used
Independent Variables	The indicators used to assess how likely a code component is to be impacted by a code smell, such as LCOM, LOC, etc
Dependent Variables	The indicator of a source code component's smelliness, such as the class level, method level, etc
Limitations	What research gaps are present in the current techniques?
Future Work	What are the future scopes of these research papers that can be utilized for future research?

Furthermore, it can be noted that most publications were journals; other than this, they also included conference papers and proceedings. This data indicates that most published articles have reported initial findings on how DL is used for code smell detection.

After taking a closer look at the writers of the papers, it was found that nearly 60% of these were co-authored, although the names and groups of such scientists were quite distinct. Some familiar authors, including Zhang, Liu, and Tushar Sharma, worked on different papers, but the number of papers published by the team of researchers, including them, was the most. Institutions like Athens University of Economics & Business that facilitate the studies for these researchers can be a guide to the upcoming scholars.

As a result, analysis of the primary studies determined the number of articles, the study areas researchers have covered, and the areas that still require more investigation. Young and new researchers might utilize this to get their study going in the direction the field dictates.

In Summary. It can be seen that there may still be unresolved issues with using DL algorithms for code smell detection based on the quantity and variety of research articles written by the community. Additionally, three different research teams were highlighted that are extensively working in this area of antipattern detection.

4.3 RQ 1. Code smells considered

This section examines a variety of code smells identified so far by DL models, giving insights into the most significant and least significant code smells.

4.3.1 RQ 1.1 - What are the code smells currently detected by DL algorithms?

This SLR shows that existing studies focused on more than 20 code smells. According to the data, 40% of studies focused on *Feature Envy*, which can also be stated to be the most frequently discovered code smell. With about 28% of work done on it, *God Class* is the second most popular smell. This result is partly expected because it is one of the most problematic and familiar code smells in development projects. Table 9 describes the summary of code smells detected with their occurence frequencies. Other smells that are widely considered in addition to Feature Envy and God class are the Long method (5), Large Class (4), Complex Class (4), Empty catch block (3), Data class (3), Long parameter list(2), Multifaceted abstraction (2), and Magic number (2) and Duplicate Code.

Spaghetti Code, Speculative Generality, Blobs, Misplaced Classes, and Temporary Fields are the less familiar smells in the code. It can also be highlighted that several unique code smells piqued the curiosity of scholars. It can be seen that some publications identified the two most uncommon code smells, *Brain method* and *Brain class*. But other code smells like dead code, inappropriate intimacy, etc., have not been detected yet and require more attention

Table 8 Selected studies for the review

ID	Author	Affiliation	Year	Publisher	Publication Type	Ref
S1	Ho et al.	Academia	2023	ACM	Conference	[39]
S2	Bhave and Sinha	Academia	2023	IEEE	Conference	[40]
S3	Nandini et al.	Industry	2023	ArXiv	Journal	[41]
S4	Kaur and Singh	Academia	2023	Wiley	Journal	[42]
S5	Tummalapalli et al.	Academia	2023	Springer	Journal	[43]
S6	Sepahvand et al.	Academia	2023	Elsevier	Journal	[44]
S7	Dewangan et al.	Academia	2022	MDPI	Journal	[45]
S8	Tarwani and Chug	Academia	2022	IEEE	Conference	[46]
S9	Zhang et al.	Academia	2022	Elsevier	Journal	[47]
S10	Zhang et al.	Academia	2022	IEEE	Conference	[48]
S11	Yedida and Menzies	Both	2022	IEEE	Conference	[49]
S12	Li and Zhang	Academia	2022	IEEE	Conference	[50]
S13	Virmajoki et al.	Academia	2022	IEEE	Conference	[51]
S14	Imam et al.	Academia	2022	Elsevier	Journal	[52]
S15	Lomio et al.	Both	2022	Springer	Journal	[53]
S16	Xu and Zhang	Academia	2021	SEKE	Conference	[54]
S17	Gupta et al.	Academia	2021	Springer	Conference	[55]
S18	Jo et al.	Academia	2021	MDPI	Journal	[56]
S19	Wei et al.	Academia	2021	IEEE	Journal	[57]
S20	Zhang and Dong	Academia	2021	Wiley	Journal	[58]
S21	Sharma et al.	Both	2021	Elsevier	Journal	[59]
S22	Liu et al.	Academia	2021	IEEE	Journal	[60]
S23	Wang et al.	Academia	2020	IEEE	Conference	[61]
S24	Hadj and Bouassida	Academia	2019	IEEE	Conference	[62]
S25	Barbez et al.	Industry	2019	IEEE	Conference	[63]
S26	Guo et al.	Academia	2019	ACM	Conference	[64]
S27	Das et al.	Industry	2019	IEEE	Conference	[65]
S28	Sharma et al.	Both	2019	ArXiv	Journal	[66]
S29	Liu et al.	Academia	2018	IEEE	Conference	[67]
S30	Hadj and Bouassida	Academia	2018	ENASE	Proceedings	[68]
S301	Li et al.	Academia	2017	IEEE	Conference	[69]
S32	Kim	Academia	2017	IAES	Journal	[70]
S33	White et al.	Academia	2016	IEEE	Proceedings	[71]
S34	Hall et al.	Academia	2014	ACM	Journal	[72]
S35	Kessentini et al.	Academia	2013	Springer	Journal	[18]

from the researchers as they can cause significant harm to the software in the long run.

The analysis of the primary studies notably revealed an intriguing finding: only one approach could detect nine different code smells simultaneously. At the same time, other systems were restricted to identifying a maximum of one type of smell. This implies that developers are currently limited to DL methods that can handle only a restricted range of code smells.

Moreover, most previous research on prediction models for code smell identification relied heavily on automatic detection tools. Consequently, the availability of these detectors influenced the code smell detection being investigated. As a result, the emphasis in the field of antipattern study has primarily been on identifying the most prevalent code smells, likely due to the research methodology employed. However, more recent articles have started to shift their focus towards unique code smells, thereby expanding the scope of DL models for code smell identification.

Summary for RQ 1. DL-based detection primarily targets the smell including *Long Method, God class, and Feature Envy*. This leads to the further conclusion that widely used DL algorithms focus on the most typical smells. Apart from this, it is evident that research is taking a turn toward detecting new and rare code smells. Hence, the scope and area of research in code smell are widening.

Table 9 List of Code Smells Detected in the Papers Analyzed under this SLR

S.no.	Code Smell Detected	Frequency	References
1	Blob	1	[55]
2	Brain class and Brain method	2	[47, 65]
3	Complex Class	4	[41, 50, 64, 66]
4	Complex Method	1	[50]
5	Data Class	3	[39, 45, 64]
6	Empty Catch Block	3	[46, 50, 66]
7	Feature Envy	10	[39, 45, 46, 48, 48, 60, 61, 64, 67, 68]
8	God Class	7	[45, 46, 63, 64, 68]
9	Large Class	4	[49, 52, 60, 67]
10	Long Identifier	1	[50]
11	Long Method	5	[60, 18, 45, 46, 50]
12	Long Parameter List	2	[41, 50]
13	Magic Number	2	[50, 66]
14	Misplaced Class	1	[49, 60]
15	Multifaceted Abstraction	2	[41, 66]
16	Spaghetti code	1	[53]
17	Speculative Generality	1	[53]
18	Swiss Army Knife	1	[55]
19	Temporary Field	1	[56]
20	Others	6	[39, 40, 40, 42, 46, 54, 55]

4.4 RQ 2. Deep learning approaches

This SLR focuses on the DL setup employed in the literature. The paper analyzes the primary studies on the following parameters: (i) the independent variable, (ii) the dependent variables, (iii) the classifiers employed, (iv) the tools used, and (v) the training procedures used. The following subsections provide descriptions of the findings of this analysis.

4.4.1 RQ 2.1 - Which independent variables give the best result in predicting code smells?

A DL model's independent variables, or features or predictors, are crucial for accurate predictions. Each metric plays a unique part in determining the code smells. It is categorized according to their design goals to properly analyze the different predictor types working in the primary paper. To list the independent variables used in preliminary studies, this paper precisely adhered to the product metrics classification framework. For instance, the Lack of Cohesion Between Methods (LCOM) metric was categorized as a Cohesion metric, while the Lines of Code (LOC), which measures the length of a source code file, was assigned to the Size category. Each primary study underwent this procedure.

While analyzing papers, it was noted that detecting a code smell affects the number of independent variables. Researchers developed formulae and methodologies employing an independent variable to determine whether a specific code smell exists. For instance, if LOC is greater than 48, then the source code contains a Large Class. Thus, it can be claimed that the independent variable depends on the sort of smell being identified; decisions are made based on the type and number of required metrics.

Table 10 describes the metrics condition to identify the existence of a particular code smell. Their meanings are defined in Appendix A. Lanza and Marinescu, in their book [73], gave a theoretical overview of the range of different metrics required for code smell detection. Values for this range are determined as *shallow = 1; few = 2-5; short memory cap = 7-8; and many with value > short memory cap*. Still, some conditions depend upon the size and complexity of the dataset particular to each paper. For such situations, there are some typical values that the dataset can define. To determine the values of average high and low, two statistical methods are employed:

1. *Average (AVG)*: To identify the central tendency or most usual value of the data collection.
2. *Standard deviation (STDEV)*: To determine how widely distributed the values in the data collection are.

After knowing the above values, the following keywords can be given the range as follows:

Table 10 Logical Representation of Code Smells using Metrics

Code Smell	Metrics	Metrics Category	References
Blob	LOC > high	Size	[55]
	WMC > high	Complexity	
	TCC < low	Cohesion	
Brain Method	LOC > high/2	Size	[47, 65]
	CYCLO > = high	Complexity	
	MAXNESTING > = few	Complexity	
	NOAV > many	Complexity	
Brain Class	Brain Method > = 1	-	[47, 65]
	LCOM < low	Cohesion	
	CYCLO > few	Complexity	
Complex Class	CYCLO > high	Complexity	[41, 50, 66, 74]
	DIT > high	Inheritance	
	NOM > high	Complexity	
Complex Method	CYCLO > high	Complexity	[50]
	LOC > high	Size	
Data Class	LCOM > low	Cohesion	[45, 74, 75]
	NOAM > high	Complexity	
Dispersed Coupling	MAXNESTING > shallow	Complexity	No detection in paper available
	CINT > short memory cap	Others	
	CDISP > = half	Others	
Feature Envy	ATFD > few	Coupling	[45, 48, 61, 67, 74]
	LAA < one third	Encapsulation	[1, 46, 60, 75, 76]
	FDP <= few	Complexity	
God Class	ATFD > few	Coupling	[45, 46, 74, 75, 77]
	WMC > = very high	Complexity	
	TCC < one third	Cohesion	
Intensive Coupling	MAXNESTING > shallow	Complexity	No detection in paper available
	CINT > few	Others	
	CDISP < quarter	Others	
Large Class	LOC > high	Size	[49, 52, 60, 78]
	DIT > low	Inheritance	
	CBO > high	Coupling	
Lazy Class	LOC > high	Size	No detection in paper available
	NOM <= low	Complexity	
Long Method	LOC > high	Size	No detection in paper available
	CYCLO > low	Complexity	
Long Parameter List	PAR > low	Others	[41, 50, 79]
Misplaced Class	LCOM < low	Cohesion	[49, 60]
	CBO > high	Coupling	
Multifaceted Abstraction	NOM > high	Complexity	[41, 66]
	DIT > high	Inheritance	
Refused Parent Bequest	BUR < one third	Others	No detection in paper available
	NOP > low	Others	
Shotgun Surgery	CM > short memory cap	Coupling	No detection in paper available
	CYCLO > many	Complexity	
Significant Duplication	SEC > average	Others	No detection in paper available
	SDC > = 2X(few+1)+1	Others	
	LB <=few	Others	

Table 10 (continued)

Code Smell	Metrics	Metrics Category	References
Spaghetti Code	CYCLO > high	Complexity	[74]
	LOC > high	Size	
Speculative Generality	NOM > high	Complexity	[74]
Switch Statement	NOC > low	Inheritance	[79]
Temporary Field	TCC > few	Cohesion	[78]
	LCOM > low	Cohesion	
Tradition Breaker	AMW > average	Size	No detection in paper available
	WMC > = very high/2	Complexity	
	NOM > high/2	Complexity	

- Low = AVG − STDEV.
- High = AVG + STDEV.
- Very high = (AVG+STDEV)*1.5.

These margins now clear the meaning of *low*, *high*, and *very high* in Table 10.

Most code smell-detecting models developed primarily depend on source code metrics that could detect the structural characteristics of a part of relevant code. Amongst all metrics, the Chidamber and Kemerer (CK) metric suite and Lines of Code (LOC) metrics are the most popular and utilized in most primary research. Other valuable metrics included WMC, CBO, NOM, DIT, NOC, and LCOM. This test suite was chosen primarily because it contains metrics that can measure various aspects of source code, such as cohesion, coupling, or complexity. They thus served as the foundation for developing DL methods for detecting code smells.

Numerous additional structural-based measures were also utilized in addition to the metrics used to supplement the evaluation made feasible by the CK suite. The primary research altered specific metrics to represent the structure of source code components more accurately. They specifically build alternatives, such as CK metrics, that do not consider the accessor methods like getters and setters. Instead, other metrics, such as those belonging to the encapsulation and inheritance categories, were considered predictors. This likely demonstrates the research community's openness to investigate novel metrics for enhancing the predictive capability of DL-based code smell detectors.

To broaden the discussion, the papers reviewed show that only some studies have looked at metrics other than structural ones, which is necessary for antipattern detection. Zhang et al. [47] worked on semantic data analysis. Here, it can be observed that the usage of semantic data expands the scope of correctly detecting code smells. Because these metrics strongly complement one another, DL approaches would perform significantly better if they were more widely used.

None of the studied papers used alternative metrics, such as dynamic or metrics that could distinguish between different concerns. Such metrics fit the detection of many kinds of code smells perfectly. Dynamic metrics, for example, could be effectively used to identify instances of *Message Chains*, defined as methods that perform multiple methods calls. Concurrently, due to their ability to quantify concern's scattering and tangling properties, concern separation metrics are already thought to detect code smells effectively like *Divergent Change* and *God Class*. Adding those metrics can significantly improve the effectiveness of current approaches and, more importantly, can be used to detect more code smells at various granularity levels. Many researchers have produced labeled data sets with several independent variables & labeling tools on GitHub.

The more significant point is that none of the earlier studies considered developer-related data (such as workload or experience) that has been successfully used in other software engineering contexts, such as bug or change prediction. Given the subjectivity with which developers interpret these factors in code smell detection, these factors might even be more crucial. In other words, developers can be recommended to beware of code smells that are most important for them at a given time based on their expertise or workload. Therefore, their incorporation offers researchers a significant chance to close the knowledge gap between the field and practice.

Summary of RQ 2.1. The majority of primary studies examined used product metrics. Also, studies went with the combination of different kinds of metrics for efficiency. However, it was found that only a few studies examined the utility of process and text metrics together as an alternative way of identifying code smells. Likewise,

it was observed that analyzing semantic metrics improved the effectiveness of code smell detection.

4.4.2 RQ 2.2 - Which dependent variable is used to get the results and analysis of code smells?

The dependent variables of code smell-detecting models give information about the severity of smelliness. A classification type defines the sort of output that a code smell prediction model generates. Binary categories and probability ranges are the most common ways of determining dependent variables. In the first situation, a model seeks to determine whether or not a smell afflicts a software artifact. In contrast, the latter aims to forecast the possibility of a source code entity being smelly, typically ranging between 0 and 1. The fundamental distinction between the two classification approaches is the choice of a threshold to separate those objects. From this study, it is observed that 60% of the papers used binary classification while 40% went for probability classification.

There are mainly two reasons why binary type classification is so popular: firstly, it makes the dataset categorical, and secondly, predictions in DL models become easier. Numerous code smells in a single dataset become readily identifiable. This SLR also contains datasets in which a boolean type dependent variable is employed to determine whether or not a specific code smell is present in the dataset by giving the output as TRUE/FALSE.

Studies that defined classification methods had two types of results. Either the code is (i) smelly or (ii) non-smelly. Code smells like *Temporary Field, Large Class, and Misplaced Class* had their smelliness predicted using probability classification wherein the percentage of the range of code smell was given in the project. Other code smells were identified using binary classification, which could tell if the program was smelly. Four code smells have been tested across both classification types: *God Class, Long Method, Feature Envy, and Data Class*. It can be seen that both categorization types are effective from their respective point of view because the results from both are precise and clear.

Summary for RQ 2.2. Two classification categories are identified: binary classification and probability classification. Through investigation, it was found that 40% of primary studies worked using probability and 60% using binary classification type. According to the findings, both have worked satisfactorily, and newer approaches can be formed.

4.4.3 RQ 2.3 - Which DL algorithms are utilized in code smell detection?

The algorithm of each model is the most crucial component. This review featured various articles that used some of the most common DL algorithms, including CNNs and Artificial Neural Networks (ANNs). This result is summarised in Fig. 9. Diversifying the number of CNN layers improved the prediction accuracy. While some researchers found that increasing the layer count boosted prediction accuracy, other studies found that changing the layer count made no impact. Consequently, it can be seen that the precision is independent of the quantity of CNN layers.

While using any DL model, the dataset must be augmented. Since there were limited datasets with pre-augmentation, methods required a data augmentation technique that could address the issue of imbalanced datasets. Through investigation, it was found that the Synthetic Minority Oversampling Technique (SMOTE) is one of the most popular data augmentation techniques. It is a statistical approach that enables the model to increase the number of instances in the data set in a balanced manner. The component creates new illustrations from existing minority situations specified as input.

Continuing with the analysis, some papers advanced the level of model creation by combining two or more learning layers to create a more efficient layer. The model was also fitted with additional deep-learning neural networks, including Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and bi-LSTM, extensively employed with CNN. When combined with CNN layers, they performed more accurately. In most situations, combining these models worked better than a single-layer model. This shows us that the best features of more than one layer, combined with the basic CNN layer, can vary and even increase the chances of getting above-fair results.

Dense layers and output layers were finally used to finish the models. The number of thick layers varied from paper to paper, but the real impact was caused by the specific DL algorithm used in previous layers. Hence, it was observed that there is no constraint on the number of dense and output layers to the effectiveness of the models against antipattern detection.

Summary for RQ 2.3. It is observed that CNN and RNN are the most used algorithms with a combination of different DL layers, including LSTM and Bi-LSTM, strengthening the models' precision. Amongst all, CNN and LSTM layers are the most effective DL algorithms, as 29% of the studies used them.

4.4.4 RQ 2.4 - What detecting tools are most commonly used in research for code smells?

Researchers have employed detection tools in papers for code smell identification and performing possible refactoring. Figure 8 presents the list of tools utilized in detection. Some of the most essential and frequently used tools include jDeodorant and iPlasma, which are metric-based detection tools. Both of these tools are Eclipse plug-ins. An Eclipse plug-in is an extensible framework applied to individual software that locates and fixes software design issues by implementing the proper refactoring. JDeodorant uses several cutting-edge approaches and methods to spot code smells and provide the right refactoring to fix them. The tool currently recognizes five categories of bad code smells: *Feature Envy, Type Checking, Long Method, God Class, and Duplicated Code*. Other utilities, besides the Eclipse plugin, are available as independent software. They can function like an Eclipse plugin without IDE and require a complete setup before being utilized.

Some automatic detection tools, such as BDST and jDev, are available for refractive imaging. Then, a semi-automatic detection technique called SACSEA is included in the list. Despite being semi-automatic, it is utilized to reflect the most prevalent code smells, such as *God Class and Feature Envy* and not-so-popular smells like *Magic Numbers and Global variables*. Other popular tools include JProfile, NIPAD, PMD, infusion, DECOR, refactoring assistance, and CCFinder.

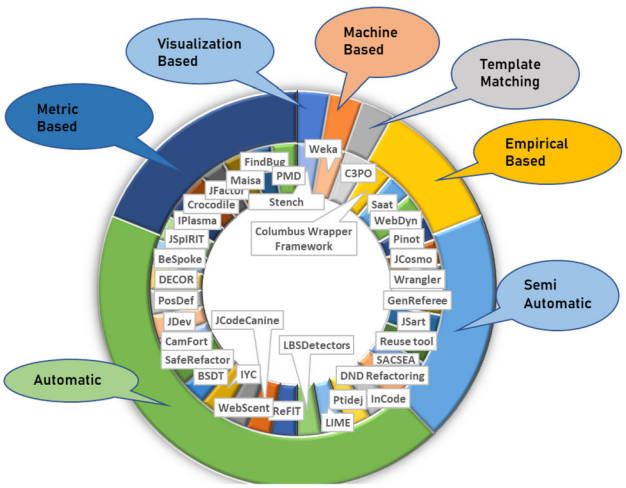


Fig. 8 Available Tools with their Detection Mechanism

Summary for RQ 2.4. The research methods make excellent use of detection tools. It is observed that metric-based tools are used most commonly. Apart from this, both semi-automatic and fully automatic tools are often employed. Both were found to be equally valuable for detecting standard code smells.

4.4.5 RQ 2.5 - What strategies have been adopted for training the models?

DL models are trained using two forms of training data: within and cross-project. Training a model using data from the same project or dataset for which the model will be used is called "within-project training." Such training methods are utilized when a model is being constructed for a particular job or issue, and the training data also represents the actual data well. The benefit of within-project training is that it enables the model to become familiar with the particular patterns and characteristics of the data used, which may improve its performance on that data.

Contrarily, cross-project training entails building a model using information from a different project or dataset than the one it will be applied to. This training is frequently utilized when the target project lacks labeled data or the model has to be trained on more varied data to enhance its generalization capabilities. Cross-project training can help the model discover more universal patterns and traits that hold across several datasets, which could improve performance on new, unexplored data.

It is important to remember that cross-project training might be more complex than within-project training since the data may have different properties or distributions from the target dataset, which can result in over-fitting or subpar generalization performance. It is crucial to carefully choose the cross-project data and utilize the proper methodologies, such as data augmentation or transfer learning, to aid the model's adaptation to the new data.

Most algorithms of prediction models developed so far have had their effectiveness assessed in intra- and inter-project contexts, resulting in the training of the same model twice, once using data from the project under study and once taking help from other project datasets. The training strategy is enhanced when a project is applied to cross-project and within-project datasets. If a given approach is evaluated across projects, it should also be employed within projects to benchmark its genuine capabilities fairly. Hence, researchers should focus more on cross-project training for efficient results.

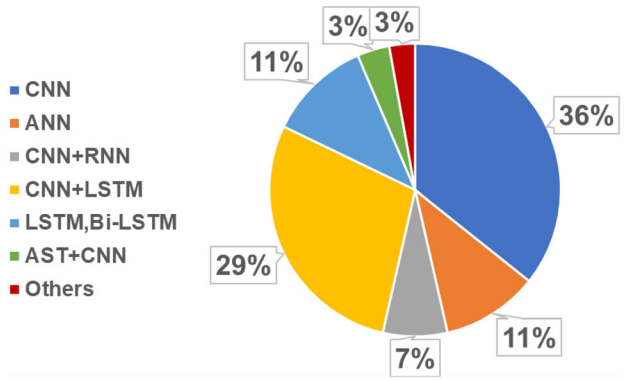


Fig. 9 Commonly used Deep Learning Algorithms

Summary for RQ 2.5. While 40% of the code smell predicting models were only considered in a cross-project situation, most were tested in both cross- and within-project settings. It was found that cross-project training gave better results.

4.5 RQ 3. - Evaluation methods

The third research question focused on the assessment methods for the code smell predicting model. It includes (i) datasets exploited in this domain and (ii) outcome metrics utilized for result summarization.

4.5.1 RQ 3.1 - What datasets are used to detect code smells?

The analysis of the datasets used in the publications is of utmost importance for comprehending the assessment setup of the primary research. To analyze papers efficiently, the authors agreed to review documents written in a particular language. Object-oriented language efforts, particularly those in Java, were the focus of the research papers and journals. This is because most code smells were discovered in Java programs, one of the most widely used object-oriented languages. However, the primary obstacle to continuing with Java is that most Java projects are private. Due to this, numerous researchers searched GitHub for various industrial and publicly accessible Java scripts. The primary source of publicly available datasets to these researchers is shown in Fig. 10. As can be seen, Apache provides the complete sets of data in Java. Xerces, JHotDraw, Eclipse, Android API, and others comprise the top part of the dataset sources. These codes were formally implemented and examined for any code smells using automatic and manual detection.

Some researchers went through both forms of detection to increase their precision. To double-check the accuracy of the computational check, manual detection was performed after the data set had been submitted for automated detection. This required significant time and money. Some researchers created private datasets for their research. This was generally done to create a dataset with unique variables specifically used for detecting a particular code smell the research focuses on. Another condition where datasets were manually created was when there was a shortage of codes containing code smells. Since this is the latest study, codes consisting of code smells are limited, and researchers need a large amount of data heuristically formulated for model training. In such situations, refactoring was applied to generate datasets according to the requirement.

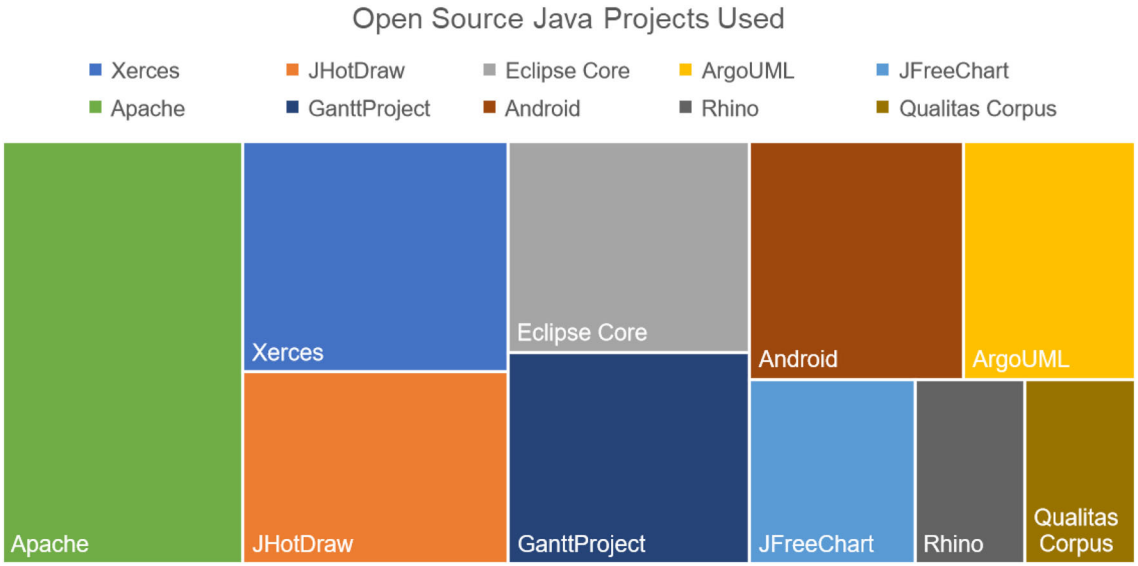


Fig. 10 Commonly used Datasets in Detecting Code Smells

Summary for RQ 3.1. This section concluded by providing the primary sources for openly available datasets to researchers working on this study. Apache is the most widely used publicly available dataset. Due to a shortage of code smell-containing datasets, few researchers generated datasets using refactoring. Publicly available datasets are open for reproduction, but manually generated datasets are private.

4.5.2 RQ 3.2 - Which result metric is used to evaluate the models?

Figure 11 shows all the different evaluation metrics used by primary studies. The most often used statistic by almost ten studies has been *F1-score*; *Accuracy* score was available for 13 papers. Some papers described their result section with a combination of three metrics mainly- F1 score, precision, and recall. The results of the three metrics complement each other and give more significant details of the result metrics.

Positive rate metric provided information about the proportion of detected code smells in the article. Authors of papers Gupta and Singh [78] and Palomba et al. [80] used this method to detect the percentage of code infected by smells. Other least frequent metrics included the *Matthews correlation coefficient* (MCC). Yu et al. [75] compared the MCC values of different algorithms applied and formulated the results. MCC is a reliable metric summarising the classifier's performance to one value, assuming positive and negative examples are given equal weight.

The *Area Under the Curve (AUC)*, a metric commonly employed in binary classification, has emerged as a prominent tool for assessing the outcomes of code smell detection. This measure was particularly prevalent in studies like [42] showcasing a range of values from 87.4% to 95.8%. Likewise, [53] leveraged AUC to achieve values around 95%. The utility of AUC curves lies in their capacity to illustrate diversity through the alterations in curve patterns. The graphical representation indicates a prevalent preference for accuracy as the primary metric employed in code smell detection research. Additionally, it illustrates consistently elevated outcomes within this metric, accompanied by reduced loss values.

Although threshold-dependent measures like F1-score, accuracy, and recall are frequently utilized, they can occasionally be biased in how well the models perform. Consequently, if they are supplemented with threshold-independent measures like MCC, the findings will be more effective.

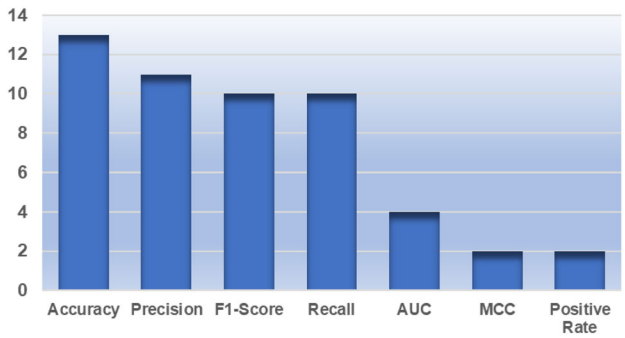


Fig. 11 Evaluation Metrics Used for Result in Primary Studies

Summary for RQ 3.2. The most commonly used metric for result evaluation is Accuracy, with 13 papers using it, followed by F1-Score, Recall, and Precision. Using threshold-independent metrics like MCC together with dependent metrics can give comparably better results.

4.6 RQ 4. - Meta-analysis of Performance

The efficiency in code smell predicting models is subject to a statistical meta-analysis presented in this final research question. For such investigations, the prediction models should be re-run on a common dataset with a similar evaluation standard to be easily compared. However, SLRs are of limited scope, so the goal is to synthesize the results from significant studies while considering the limitations of the previous studies. In this work, the focus is on how (i) independent variables, (ii) DL methodology, and (iii) training technique influenced the performance shown by these code smell prediction models.

Nearly 90% of the publications employed F-measure, accuracy, and recall as the main result analysis metrics for the performance calculation. This made it easier for us to evaluate the performance of various publications. This section will discuss how the three factors mentioned above have impacted each paper's results section. Table 11 describes the studies that used the above criteria with their results. "NR" denotes the absence of a reported metric for a particular study. Studies indicating metric improvements compared to prior algorithms without providing the precise values for their proposed method are marked with an asterisk "*" symbol.

In contrast to the inverse variance approach, this research paper has given precedence to all the chosen publications, irrespective of the dataset size or the research domain. Identifying optimal combinations of algorithms and tools for achieving the most favorable results necessitates a thorough analytical investigation.

Table 11 Evaluation metrics reported in primary studies

ID	Ref	F1-Score	Recall	Precision	Accuracy
S1	[39]	83	NR	NR	NR
S2	[40]	91.7	92.3	92	NR
S5	[43]	NR	NR	NR	97.58
S6	[44]	NR	NR	NR	75
S7	[45]	NR	NR	NR	95.75
S8	[46]	NR	NR	NR	99
S9	[47]	NR	NR	NR	4.41*
S10	[48]	43	35	40.8	1.02*
S12	[50]	74	65	86	NR
S13	[51]	NR	NR	NR	87
S14	[52]	NR	NR	NR	94.21
S17	[55]	NR	NR	NR	96.84
S18	[56]	NR	96	78	NR
S19	[57]	98	97	98	NR
S20	[58]	NR	NR	NR	2.01*
S21	[59]	22	35	16	NR
S22	[60]	NR	NR	70	80
S23	[61]	55.97	85.35	47.09	NR
S25	[63]	77	86	71	NR
S27	[65]	NR	NR	NR	97
S32	[70]	22	16	99	NR

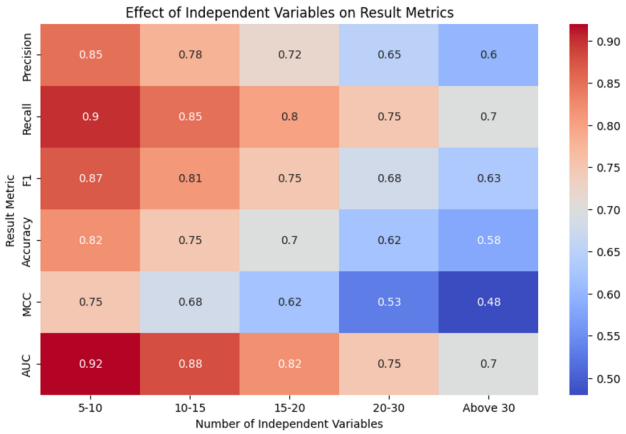


Fig. 12 Performance analysis for Independent Variables

Area Under the Curve (AUC), indicating its sensitivity to selecting appropriate variables.

Summary for RQ 4.1. Reducing the count of independent variables is imperative, as it directly impacts the model's accuracy of code smell detection. Among the various performance metrics, the one most profoundly influenced by adjustments in the independent variable count is the Area Under the Curve (AUC).

4.6.1 RQ 4.1 - How do independent variables impact performance?

According to the study, numerous independent factors are needed to identify code smells. While some studies consider structural elements, others also think of semantics. However, the argument remains that the more independent variables complicate the process, so more time is required. As a result of this, the model may end up being over-fitted. The papers also considered several intricate variable columns. Therefore, the impact it has reduces the accuracy of internal project models as well as cross-fitting models.

The heatmap depicted in Fig. 12 illustrates the influence of the number of independent variables on the performance metrics. Notably, reducing the usage of variables to focus solely on the essential ones for code smell detection leads to a heightened emphasis on critical features, consequently yielding improved outcomes. Conversely, incorporating numerous independent variables boosts the risk of diverting attention towards less pertinent features, potentially causing the model to overfit in certain instances. Among the performance metrics, the most substantial variation arising from alterations in the variable count is observed in the

4.6.2 RQ 4.2 - Is the performance of code smell prediction models affected by DL algorithms?

The DL algorithms used in the papers showed that CNN was the most used algorithm. Only using convolution layers increased precision up to an average of 90%, but it was investigated that increasing convolution layers in the model had little or no improvement in performance. The results in Fig. 13 present a comparative analysis of various DL models and their combinations based on different result metrics. Notably, the combination of CNN and LSTM consistently demonstrates the most favorable outcome across all four performance metrics. These metrics always exhibit values that fall within the higher range, indicating the attainment of remarkable results. The depicted boxplot is a valuable tool for assessing and contrasting the effectiveness of distinct deep-learning models about diverse result metrics.

The performance models are improved by adding more DL layers to the convolution layers. LSTM layers included in the CNN layers increase the results in precision- 0.0408, recall - 0.0035, and f1 - 0.043 times. In another paper, bi-LSTM layers increased the development as precision- 0.83,

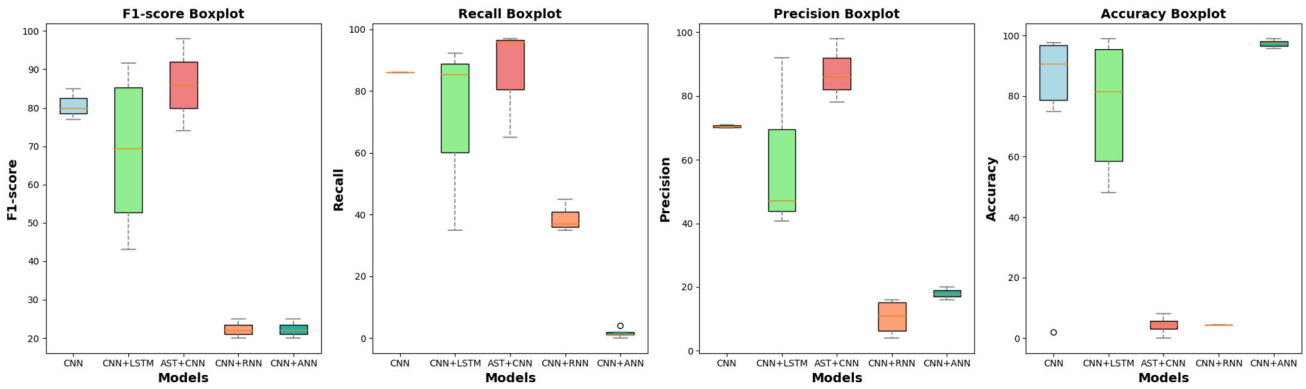


Fig. 13 Performance metrics for different DL models

recall- 0.74, and f1- 0.78 times. One of the papers used CNN and Gated Recurring Units (GRU) together. This enhanced model increased the model accuracy by 4.41%.

Summary for RQ 4.2. It is observed that using different types of DL layers increased the accuracy. Using algorithms like LSTM, bi-LSTM, and GRU gives efficient results.

4.6.3 RQ 4.3 - What is the impact of Training Strategies

As discussed earlier, the models are trained by two means, either within the project via splitting the dataset in training and testing or by cross-project. Within the project, splitting gave a way to create the algorithm using the dataset in the same format. There was no need to pre-process or make the columns similar using tools. This enhances the efficiency of the model as well. Data is first brought to a similar format as the training data in cross-project testing. However, this section analyzes them equally and fairly based on their Odds Ratio (OR). Accordingly, checking the accuracy given by both types of strategies shows that the OR of within-project models is better than cross-project ones.

Figure 14 visually presents the performance of distinct models about cross-project and within-project strategies, aiming to deliver outcome metrics. In most instances, it becomes evident that within-project training strategies yield more favorable results.

Summary for RQ 4.3. Variation in data contributes to the prediction of code smells. It is observed that within-project models perform up to 19% better than cross-project ones. Therefore, it can be said that within-project models increase the efficiency of the models.

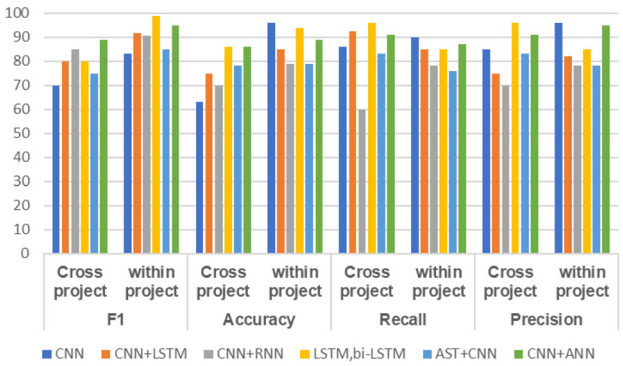


Fig. 14 Performance metrics for different Training Strategies

5 Discussion and future guidelines

Deep learning for code smell detection is a very young and expanding study area. As a result, most of the papers have employed DL techniques in their papers and identified research gaps that require more attention. Some studies emphasized these weaknesses and even indicated the areas needing repair. Future work was a significant subject of these papers. In other research, future work signified their next move towards more innovation. This section discusses the research gaps present in the primary studies and possible plans for the future.

RQ 1. - *Limited number of code smells detected using DL approaches due to limited resources:* From this research question, it can be formulated that a minimal number of code smells like *Feature Envy, God Class, Long Method, and Complex Class* have been detected using DL algorithms. However, most of the other smells given by Fowler, Brown, and Wake, like *Spaghetti Code, Refused Bequest and Functional Decomposition* mentioned in Table 1 and 2 have not yet received much attention. Hence, it becomes the most typical area for development to broaden the research to find other code smells, too. Due to the small number of projects and the limitations of available resources and technological advancements, just a tiny

percentage of code smells could be detected in a single thesis. A collection of code smells was found comparable to specific publications to be discovered simultaneously. This made it possible to improve the algorithm to meet future requirements, such as finding *Brain Class and Brain Method* simultaneously. So, efforts should be made to detect multiple code smells simultaneously to improve efficiency. Code smells, which are more frequent, may also be detected using the same method to find them.

RQ 2. - *Lack of information about metrics selected for detecting particular smell:* Selecting independent and dependent variables is one of the key methodology components of research. The papers discussed how several independent variables are required for continuing studies to identify a certain code smell. However, no paper gave concrete information about the metric rules that have been selected to detect the smell. Hence, it becomes difficult for the researchers to identify the correct set of independent variables. The basic metric rules and their selection strategy for some code smells have been given in Table 10, which can be used for strategy formulation. It is also observed that the detection strategy for some smells largely depends on the size of the dataset. Hence, a future project will also improve feature selection on such datasets for efficient time management.

RQ 2. - *Hybrid algorithms for detection and optimization:* DL method is one of the most often studied future directions. After completing their research, several academics realized that expanding the combinations of many DL layers might be a more practical approach toward code smell detection. A few researchers put out the notion of adding LSTM layers to the CNN model. One of the papers recommended utilizing Natural Language Processing (NLP) more effectively for better results. When the area of DL is acquainted with the field of code smell, researchers also want to extend this notion to **quantum computing**. The problem of code smell also exists in quantum computer-written codes. Development for expanding the idea of detecting code smell in quantum computing is still ongoing since the design and syntax of these codes differ from standard OOP languages. It is expected that the number of quantum papers for validity could have been higher than expected. **Use of genetic algorithms to optimize the performance of the models is also one of the future directions for researchers.** Utilizing genetic algorithms is essential for devising an optimized approach for metric parameters, thereby enhancing both model training efficiency and the detection process. Research into optimized techniques and evolutionary detection within the research domain holds potential.

RQ 3. - *Limited number of openly available datasets:* Due to the increased source and code formulation, it takes longer to create a great dataset. A minimal set of openly available datasets can be used for code smell detection. Existing datasets have been either exhausted to detect code smells or possess only some code smells. **Area of *Generative AI* can be explored to manually generate new structural and semantic datasets in a time-bound manner.**

RQ 4. - *Meta Analysis and Language restrictions:* The meta-analysis emphasized that various factors, namely the independent variables, DL algorithm, and training strategy, can significantly influence the performance of DL approaches. This finding is crucial for future research, as it underscores the importance of thoroughly evaluating the impact of each aspect related to the settings of DL approaches. Doing so makes it easier to interpret the performance and understand how the technique responds to different environments. Additionally, primary studies mainly concentrated on Java code but can be extended to other object-oriented languages, such as Python, C, C++, etc. Language extension should also be employed to widen the scope of available semi-automatic and automatic detection tools.

6 Open issues

This section identifies open issues or research gaps in the primary studies considered in this SLR. It is observed that DL algorithms could identify a particular group of code smells. The number of code smells detected can vary by increasing the independent variables or introducing new detection tools. This issue can be a focus of concern for further research advancements. Relying on a single benchmark dataset and a small number of code smells for experimentation cannot produce appropriate results. This can lead to negligence in adjusting model hyperparameters before training to reach a local optimum, affecting overall performance. Another study discusses their drawback as relying on a single benchmark dataset and a small number of code smells for experimentation. They also neglected to adjust the model hyperparameters before training, which led to a local optimum and impacted overall performance.

As discussed above, SMOTE is the most used data augmenting technique. It is helpful in this field, but a few papers showed no improvement in detecting God class code smell using SMOTE. Hence, a new approach other than SMOTE should be brought to use. Another area for improvement is that more optimization points lead to subpar performance. It demonstrates both the need for optimization points and the viability of their concept. Here, evolutionary algorithms can be used to achieve such optimization.

Apart from them, issues like exploring the field of quantum computing and genetic algorithms are very

critical. As discussed, quantum code smells are evolving and can threaten software maintenance in quantum computing. Chen et al. in [81] briefly explained some code smells seen exquisitely in the quantum field. It requires early detection and removal tools. However, research in this field is rare as of now. Similarly, there are steps where researchers are trying to accomplish detection by genetic algorithm or nature-inspired algorithm to detect multiple smells simultaneously.

7 Threats to validity

Listed below are the identified challenges that were encountered during the process of conducting this SLR.

7.1 Construct validity threats

The challenge of finding comprehensive and relevant search keywords can impact the construct validity. If the chosen keywords do not adequately represent the concepts of interest, the retrieved studies might not truly reflect the intended scope of the research. A thorough search strategy was created to lessen the risks in this area. After extracting the first list containing keywords related to the research questions, alternate spellings and synonyms were examined. Verifying such phrases is done to check if these phrases appeared in the pertinent publications. The search terms act as a guide to obtaining the relevant papers. It's also crucial to point out that the second author of this research double-checked the procedures used to choose the relevant studies. The articles published from 2013 to date by the best software engineering publishers were manually searched again.

7.2 Internal validity threats

The process of applying inclusion and exclusion criteria involves subjective judgment. The exclusion and inclusion criteria are used for the study selection procedure. The procedure is carried out in two steps: first, the paper's first author used the criteria to screen out irrelevant publications, while the second author cross-checked. Consequently, an excellent agreement existed between them, which validated the entire process.

In terms of the quality evaluation and data extraction process, this paper established a formal procedure that resulted in (i) the definition of a dataset that can confirm the presence of important information in the specified publications and papers and (ii) collecting data that can answer relative research questions.

7.3 External validity threats

The study's chosen timeline was confined to the period between 2013 and 2023, which means that the assessment of papers and studies was limited to this timeframe. However, it must be recognized that advancements have been beyond this timeline, particularly in emerging fields. More recent techniques have been developed, and it's crucial to acknowledge their significance and incorporate them into consideration. The outcomes could differ based on the timeline of the selected studies.

7.4 Conclusion validity threats

Additional threats stem from the meta-analysis done to compare the results of the original studies' performance. Random and fixed effect models are categorized based on the F-measure, Recall, and Precision of the methods suggested in the literature. It is vital to remember that, by its very nature, the specified technique cannot account for the defect design and differences in the original research, even if it may be viewed as a convenient way to aggregate and synthesize the findings of several investigations. This constraint is accepted to provide a broad overview of the effects of independent factors, DL methods, and training methodology on the effectiveness of code smell prediction models. For advanced research, a thorough benchmark analysis can be included that re-runs each prediction model on a shared dataset, which would unquestionably reinforce the findings of this study.

8 Conclusions

This paper presents a systematic literature review on DL techniques for code smell detection. The target papers for this review are published between 2013 and till date. From an initial collection of 8739 articles, 35 papers have been selected through a specially designed selection process.

The primary studies have been thoroughly consolidated, covering four important aspects:

1. *Prevalent Code smells*: The analysis of the paper reveals the prevalence of various severe code smells in Java projects, with notable instances being *Feature Envy, God Class,* and *Long Method*. Identifying code smells in clusters is favored, as it enables the simultaneous detection of multiple code smells through shared independent metrics.
2. *DL Approached Used*: The experiments predominantly involved utilizing DL approaches like *CNN, RNN,* and *LSTM*. The results highlighted the effectiveness of hybrid models like CNN+LSTM and bi-LSTM. The

assessment and categorization of detection tools based on their relevance were performed. Evaluating training procedures emphasized the superior efficiency of intra-project training over cross-project training. The research also investigated software metrics employed in code smell detection, focusing on the metrics crucial for detecting particular code smells.

3. *Evaluation methods employed*: An evaluation was conducted on open-source datasets to determine the effectiveness of DL models, revealing that Apache stood out as the most frequently utilized dataset for DL-based detection. Additionally, assessing commonly employed result metrics showed that accuracy was the predominant choice for model evaluation across most papers.
4. *Performance analysis of the proposed models*: Exploring independent variables and their required count for effective detection revealed that fewer variables led to improved outcomes. Analysis of model outcomes across various evaluation methods indicated that hybrid models yielded superior results. Furthermore, it was determined that the development of models is not dependent on the number of layers.

This paper presents guidelines and recommendations for future research efforts based on the results. Additionally, a summary of metric representation for detecting code smells is provided in this paper. This summary can serve as a benchmark of rules for the development of detection techniques in the future.

It can be concluded from this SLR that code smell detection is still a vast study area and that future researchers can advance DL by combining quantum computing and genetic algorithms for code smell detection.

Appendix A. List of Metrics with their Definitions used in this Study

Acronym	Definition
AMW	Afferent Method and Weighted Method
ATFD	Access to Foreign Data
BUR	Buse Readability Score
CBO	Coupling Between Objects
CDISP	Class Dependency Metric
CINT	Coupling Intensity
CM	Coupling Between Objects
CYCLO	Cyclomatic Complexity
DIT	Depth of Inheritance Tree
FDP	Foreign Data Providers
LAA	Locality of Attribute Accesses
LB	Basic Lack of Cohesion in Methods
LCOM	Lack of Cohesion in Methods
LOC	Lines Of Code
MAXNESTING	Maximum Nesting Level
NOAM	Number of Accessed Members
NOAV	Number of Accessed Variables
NOC	Number of Children
NOM	Number of Methods
NOP	Number of Parents
PAR	Number of Parameters
SDC	Size of Data Context
SEC	Signal Event Coupling
TCC	Tight Class Cohesion
WMC	Weighted Methods Count

Appendix B. List of Primary Studies Selected for this SLR

ID	Title	References
S1	Fusion of deep convolutional and LSTM recurrent neural networks for automated detection of code smells	[39]
S2	Deep Multimodal Architecture for Detection of Long Parameter List and Switch Statements using DistilBERT	[40]
S3	DACOS-A Manually Annotated Dataset of Code Smells	[41]
S4	Improving the Quality of Open Source Software	[42]
S5	Web Service Anti-patterns Detection Using CNN with Varying Sequence Padding Size	[43]
S6	Using Word Embedding and Convolution Neural Network for Bug Triaging by Considering Design Flaws	[44]
S7	Code Smell Detection Using Ensemble Machine Learning Algorithms	[45]
S8	Application of Deep Learning models for Code Smell Prediction	[46]
S9	DeleSmell: Code smell detection based on deep learning and latent semantic analysis	[47]
S10	Feature Envy Detection with Deep Learning and Snapshot Ensemble	[48]
S11	How to improve deep learning for software analytics: (a case study with code smell detection)	[49]
S12	Multi-Label Code Smell Detection with Hybrid Model based on Deep Learning	[50]

ID	Title	References
S13	Detecting Code Smells with AI: a Prototype Study	[51]
S14	The automation of the detection of large class bad smell by using genetic algorithm and deep learning	[52]
S15	A machine and deep learning analysis among SonarQube rules, product, and process metrics for fault prediction	[53]
S16	Multi-granularity code smell detection using deep learning method based on abstract syntax tree	[54]
S17	An Empirical Study on Predictability of Software Code Smell Using Deep Learning Models	[55]
S18	Two-Pass Technique for Clone Detection and Type Classification Using Tree-Based Convolution Neural Network	[56]
S19	FCCA: Hybrid Code Representation for Functional Clone Detection Using Attention Networks	[57]
S20	MARS: Detecting brain class/method code smell based on metric-attention mechanism and residual network	[58]
S21	Code smell detection by deep direct-learning and transfer-learning	[59]
S22	Deep Learning Based Code Smell Detection	[60]
S23	Feature Envy Detection based on Bi-LSTM with Self-Attention Mechanism	[61]
S24	Deep Representation Learning for Code Smells Detection using Variational Auto-Encoder	[62]
S25	Deep Learning Anti-Patterns from Code Metrics History	[63]
S26	Deep semantic-Based Feature Envy Identification	[64]
S27	Detecting code smells using deep learning	[65]
S28	On the feasibility of transfer-learning code smells using deep learning	[66]
S29	Deep learning based feature envy detection	[67]
S30	A hybrid approach to detect code smells using deep learning	[68]
S31	Cclearner: A deep learning-based clone detection approach	[69]
S32	Finding bad code smells with neural network models	[70]
S33	Deep learning code fragments for code clone detection	[71]
S34	Some code smells have a significant but small effect on faults	[72]
S35	What you like in design use to correct bad-smell	[18]

Funding The authors have not disclosed any funding.

Data availibility No data was used for the research described in the article.

Declarations

Competing interests 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. Bastías, O.A., Díaz, J., López Fenner, J.: Exploring the intersection between software maintenance and machine learning; a systematic mapping study. Appl. Sci. **13**, 1710 (2023)
2. AlOmar, E.A., Ivanov, A., Kurbatova, Z., Golubev, Y., Mkaouer, M.W., Ouni, A., Bryksin, T., Nguyen, L., Kini, A., Thakur, A.: Just-in-time code duplicates extraction. Inform. Softw. Technol. **158**, 107169 (2023)
3. Li, Z., Wang, S., Wang, W., Liang, P., Mo, R., Li, B.: Understanding bugs in multi-language deep learning frameworks, (2023). arXiv:2303.02695
4. Al Khatib, S.M., Alkharabsheh, K., Alawadi, S.: Selection of human evaluators for design smell detection using dragonfly optimization algorithm: An empirical study. Inform. Softw. Technol. **155**, 107120 (2023)
5. Kim, D.K.: A deep neural network-based approach to finding similar code segments. IEICE TRANSACTIONS Inform. Syst. **103**, 874–878 (2020)
6. Alkharabsheh, K., Alawadi, S., Kebande, V.R., Crespo, Y., Fernández-Delgado, M., Taboada, J.A.: A comparison of machine learning algorithms on design smell detection using balanced and imbalanced dataset: a study of god class. Inform. Softw. Technol. **143**, 106736 (2022)
7. Brown, W.H., Malveau, R.C., McCormick, H.W.S., Mowbray, T.J.: AntiPatterns: refactoring software, architectures, and projects in crisis. Wiley, Hoboken (1998)
8. Gesi, J., Liu, S., Li, J., Ahmed, I., Nagappan, N., Lo, D., de Almeida, E.S., Kochhar, P.S., Bao, L.: Code smells in machine learning systems, (2022). arXiv:2203.00803
9. Muralidhar, N., Muthiah, S., Butler, P., Jain, M., Yu, Y., Burne, K., Li, W., Jones, D., Arunachalam, P., McCormick, H.S., Ramakrishnan, N.: Using antipatterns to avoid mlops mistakes, (2021). arXiv:2107.00079
10. Kaur, I., Kaur, A.: A novel four-way approach designed with ensemble feature selection for code smell detection. IEEE Access **9**, 8695–8707 (2021)
11. Fowler, M.: Refactoring, Addison-Wesley Professional, (2018)
12. Spadini, D., Palomba, F., Zaidman, A., Bruntink, M., Bacchelli, A., On the relation of test smells to software code quality, in: IEEE international conference on software maintenance and evolution (ICSME). IEEE **2018**, 1–12 (2018)
13. Barbez, A.: Deep Learning Structural and Historical Features for Anti-Patterns Detection, Master's thesis, École Polytechnique de Montréal, (2018). https://publications.polymtl.ca/3724/
14. Kruchten, P., Nord, R.L., Ozkaya, I.: Technical debt: from metaphor to theory and practice. IEEE Softw. **29**, 18–21 (2012)
15. Gupta, A., Suri, B., Misra, S.: A systematic literature review: code bad smells in java source code, in: Computational Science and Its Applications–ICCSA 2017: 17th International Conference, Trieste, Italy, July 3-6, 2017, Proceedings, Part V 17, Springer, pp. 665–682, (2017)
16. Fowler, M., Beck, K., Brant, J., Opdyke, W., Roberts, D.: Refactoring: Improving the design of existing code addison-wesley professional. CA, USA, Berkeley (1999)
17. Mantyla, M.: Bad smells in software-a taxonomy and an empirical study, Ph.D. thesis, PhD thesis, Helsinki University of Technology, (2003)

18. Kessentini, M., Mahaouachi, R., Ghedira, K.: What you like in design use to correct bad-smells. Softw. Qual. J. **21**, 551–571 (2013)
19. Liu, H., Liu, Q., Niu, Z., Liu, Y.: Dynamic and automatic feedback-based threshold adaptation for code smell detection. IEEE Trans. Softw. Eng. **42**, 544–558 (2015)
20. Wake, W.C.: Refactoring workbook, Addison-Wesley Professional, (2004)
21. Pascarella, L., Palomba, F., Bacchelli, A.: Fine-grained just-in-time defect prediction. J. Syst. Softw. **150**, 22–36 (2019)
22. de Paulo Sobrinho, E.V., De Lucia, A., de Almeida Maia, M.: A systematic literature review on bad smells-5 w's: which, when, what, who, where. IEEE Trans. Softw. Eng. **47**, 17–66 (2018)
23. Rasool, G., Arshad, Z.: A review of code smell mining techniques. J. Softw.: Evol. Process **27**, 867–895 (2015)
24. Sabir, F., Palma, F., Rasool, G., Guéhéneuc, Y.-G., Moha, N.: A systematic literature review on the detection of smells and their evolution in object-oriented and service-oriented systems. Softw.: Prac. Exp. **49**, 3–39 (2019)
25. Zhang, M., Hall, T., Baddoo, N.: Code bad smells: a review of current knowledge. J. Softw. Maint. Evol.: Res. Prac. **23**, 179–202 (2011)
26. Misbhauddin, M., Alshayeb, M.: Uml model refactoring: a systematic literature review. Empir. Softw. Eng. **20**, 206–251 (2015)
27. Misbhauddin, M., Alshayeb, M.: Uml model refactoring: a systematic literature review. Empir. Softw. Eng. **20**, 206–251 (2015)
28. Azeem, M.I., Palomba, F., Shi, L., Wang, Q.: Machine learning techniques for code smell detection: A systematic literature review and meta-analysis. Inform. Softw. Technol. **108**, 115–138 (2019)
29. Caram, F.L., Rodrigues, B.R.D.O., Campanelli, A.S., Parreiras, F.S.: Machine learning techniques for code smells detection: a systematic mapping study. Int. J. Softw. Eng. Knowl. Eng. **29**, 285–316 (2019)
30. AbuHassan, A., Alshayeb, M., Ghouti, L.: Software smell detection techniques: a systematic literature review. J. Softw.: Evol. Process **33**, e2320 (2021)
31. Singh, S., Kaur, S.: A systematic literature review: Refactoring for disclosing code smells in object oriented software. Ain Shams Eng. J. **9**, 2129–2151 (2018)
32. Agnihotri, M., Chug, A.: A systematic literature survey of software metrics, code smells and refactoring techniques. J. Inform. Process. Syst. **16**, 915–934 (2020)
33. Lewowski, T., Madeyski, L.: How far are we from reproducible research on code smell detection? a systematic literature review. Inform. Softw. Technol. **144**, 106783 (2022)
34. Alazba, A., Aljamaan, H., Alshayeb, M.: Deep learning approaches for bad smell detection: a systematic literature review. Empir. Softw. Eng. **28**, 77 (2023)
35. Lei, M., Li, H., Li, J., Aundhkar, N., Kim, D.-K.: Deep learning application on code clone detection: a review of current knowledge. J. Syst. Softw. **184**, 111141 (2022)
36. Alsolai, H., Roper, M.: A systematic literature review of machine learning techniques for software maintainability prediction. Inform. Softw. Technol. **119**, 106214 (2020)
37. Fernandes, E., Oliveira, J., Vale, G., Paiva, T., Figueiredo, E.: A review-based comparative study of bad smell detection tools, in: Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering, pp. 1–12, (2016)
38. Kitchenham, B.A.: Systematic review in software engineering: where we are and where we should be going, in: Proceedings of the 2nd international workshop on Evidential assessment of software technologies, p. 1–2 (2012)
39. Ho, A., Bui, A.M.T., Nguyen, P.T., Di Salle, A.: Fusion of deep convolutional and lstm recurrent neural networks for automated detection of code smells, in: Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering, EASE '23, Association for Computing Machinery, New York, NY, USA, p. 229-234. (2023) https://doi.org/10.1145/3593434.3593476
40. Bhave, A., Sinha, R.: Deep multimodal architecture for detection of long parameter list and switch statements using distilbert, in: 2022 IEEE 22nd International Working Conference on Source Code Analysis and Manipulation (SCAM), pp. 116–120. (2022) https://doi.org/10.1109/SCAM55253.2022.00018
41. Nandani, H., Saad, M., Sharma, T.: Dacos-a manually annotated dataset of code smells, (2023). arXiv:2303.08729
42. Kaur, S., Singh, S.: Improving the quality of open source software, Agile Software Development: Trends, Challenges and Applications, 309–323 (2023)
43. Tummalapalli, S., Kumar, L., Bhanu Murthy, N.L.: Web service anti-patterns detection using cnn with varying sequence padding size, in: Mobile Application Development: Practice and Experience: 12th Industry Symposium in Conjunction with 18th ICDCIT 2022, Springer, pp. 153–165 (2023)
44. Sepahvand, R., Akbari, R., Jamasb, B., Hashemi, S., Boushehrian, O.: Using word embedding and convolution neural network for bug triaging by considering design flaws. Sci. Comput. Program. **228**, 102945 (2023)
45. Dewangan, S., Rao, R.S., Mishra, A., Gupta, M.: Code smell detection using ensemble machine learning algorithms. Appl. Sci. **12**, 10321 (2022)
46. Tarwani, S., Chug, A.: Application of deep learning models for code smell prediction, in: 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), pp. 1–5. (2022) https://doi.org/10.1109/ICRITO56286.2022.9965048
47. Zhang, Y., Ge, C., Hong, S., Tian, R., Dong, C., Liu, J.: Delesmell: code smell detection based on deep learning and latent semantic analysis. Knowl.-Based Syst. **255**, 109737 (2022)
48. Zhang, M., Jia, J.: Feature envy detection with deep learning and snapshot ensemble, in: 2022 9th International Conference on Dependable Systems and Their Applications (DSA), pp. 215–223. (2022) https://doi.org/10.1109/DSA56465.2022.00037
49. Yedida, R., Menzies, T.: How to improve deep learning for software analytics (a case study with code smell detection), in: 2022 IEEE/ACM 19th International Conference on Mining Software Repositories (MSR), pp. 156–166. (2022) https://doi.org/10.1145/3524842.3528458
50. Li, Y., Zhang, X.: Multi-label code smell detection with hybrid model based on deep learning., in: SEKE, pp. 42–47. (2022)
51. Virmajoki, J., Knutas, A., Kasurinen, J.: Detecting code smells with ai: a prototype study, in: 2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO), pp. 1393–1398. (2022) https://doi.org/10.23919/MIPRO55190.2022.9803727
52. Imam, A.T., Al-Srour, B.R., Alhroob, A.: The automation of the detection of large class bad smell by using genetic algorithm and deep learning. J. King Saud Univ.: Comput. Inform. Sci. **34**, 2621–2636 (2022)
53. Lomio, F., Moreschini, S., Lenarduzzi, V.: A machine and deep learning analysis among sonarqube rules, product, and process metrics for fault prediction. Empir. Softw. Eng. **27**, 189 (2022)
54. Xu, W., Zhang, X.: Multi-granularity code smell detection using deep learning method based on abstract syntax tree, volume 2021-July, Pittsburgh, PA, United states, pp. 503 – 509. (2021) https://doi.org/10.18293/SEKE2021-014
55. Gupta, H., Kulkarni, T.G., Kumar, L., Neti, L.B.M., Krishna, A.: An empirical study on predictability of software code smell using deep learning models, in: Advanced Information Networking and Applications, Springer International Publishing, 2021,

pp. 120–132. https://doi.org/10.1007%2F978-3-030-75075-6_10. https://doi.org/10.1007/978-3-030-75075-6_10
56. Jo, Y.-B., Lee, J., Yoo, C.-J.: Two-pass technique for clone detection and type classification using tree-based convolution neural network. Appl. Sci. **11**, 6613 (2021)
57. Hua, W., Sui, Y., Wan, Y., Liu, G., Xu, G.: Fcca: hybrid code representation for functional clone detection using attention networks. IEEE Trans. Reliab. **70**, 304–318 (2021)
58. Zhang, Y., Dong, C.: Mars: detecting brain class/method code smell based on metric-attention mechanism and residual network. J. Softw.: Evol. Process (2021). https://doi.org/10.1002/smr.2403
59. Sharma, T., Efstathiou, V., Louridas, P., Spinellis, D.: Code smell detection by deep direct-learning and transfer-learning. J. Syst. Softw. **176**, 110936 (2021)
60. Liu, H., Jin, J., Xu, Z., Zou, Y., Bu, Y., Zhang, L.: Deep learning based code smell detection. IEEE Trans. Softw. Eng. **47**, 1811–1837 (2021)
61. Wang, H., Liu, J., Kang, J., Yin, W., Sun, H., Wang, H.: Feature envy detection based on bi-lstm with self-attention mechanism, in: 2020 IEEE Intl Conf on Parallel and Distributed Processing with Applications, Big Data and Cloud Computing, Sustainable Computing and Communications, Social Computing and Networking (ISPA/BDCloud/SocialCom/SustainCom), pp. 448–457. (2020) https://doi.org/10.1109/ISPA-BDCloud-SocialCom-SustainCom51426.2020.00082
62. Hadj-Kacem, M., Bouassida, N.: Deep representation learning for code smells detection using variational auto-encoder, in. International Joint Conference on Neural Networks (IJCNN) **2019**, 1–8 (2019). https://doi.org/10.1109/IJCNN.2019.8851854
63. Barbez, A., Khomh, F., Guéhéneuc, Y.-G.: Deep learning anti-patterns from code metrics history, in. IEEE International Conference on Software Maintenance and Evolution (ICSME) **2019**, 114–124 (2019). https://doi.org/10.1109/ICSME.2019.00021
64. Guo, X., Shi, C., Jiang, H.: Deep semantic-based feature envy identification, in: Proceedings of the 11th Asia-Pacific Symposium on Internetware, Internetware '19, Association for Computing Machinery, New York, NY, USA, (2019). https://doi.org/10.1145/3361242.3361257
65. Das, A.K., Yadav, S., Dhal, S.: Detecting code smells using deep learning, in: TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), pp. 2081–2086. (2019) https://doi.org/10.1109/TENCON.2019.8929628
66. Sharma, T., Efstathiou, V., Louridas, P., Spinellis, D.: On the feasibility of transfer-learning code smells using deep learning, (2019). arXiv:1904.03031
67. Liu, H., Xu, Z., Zou, Y.: Deep learning based feature envy detection, in: 2018 33rd IEEE/ACM International Conference on Automated Software Engineering (ASE), pp. 385–396. (2018) https://doi.org/10.1145/3238147.3238166
68. Hadj-Kacem, M., Bouassida, N.: A hybrid approach to detect code smells using deep learning., in: ENASE, pp. 137–146 (2018)
69. Li, L., Feng, H., Zhuang, W., Meng, N., Ryder, B.: Cclearner: A deep learning-based clone detection approach, in: 2017 IEEE International Conference on Software Maintenance and Evolution (ICSME), IEEE, pp. 249–260 (2017)
70. Kim, D.K.: Finding bad code smells with neural network models, International. J. Electr. Comput. Eng. **7**, 3613 (2017)
71. White, M., Tufano, M., Vendome, C., Poshyvanyk, D.: Deep learning code fragments for code clone detection, in: Proceedings of the 31st IEEE/ACM International Conference on Automated Software Engineering, ASE '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 87-98. https://doi.org/10.1145/2970276.2970326
72. Hall, T., Zhang, M., Bowes, D., Sun, Y.: Some code smells have a significant but small effect on faults. ACM Trans. Softw. Eng. Methodol. **23**, 1–39 (2014)
73. Lanza, M., Marinescu, R.: Object-Oriented Metrics in Practice. Evaluate, and Improve the Design of Object-oriented Systems, Springer Verlag, Using Software Metrics to Characterize (2010)
74. Saheb-Nassagh, R., Ashtiani, M., Minaei-Bidgoli, B.: A probabilistic-based approach for automatic identification and refactoring of software code smells. Appl. Soft Comput. **130**, 109658 (2022)
75. Li, F., Zou, K., Keung, J.W., Yu, X., Feng, S., Xiao, Y.: On the relative value of imbalanced learning for code smell detection. Prac. Exp. Softw. (2023). https://doi.org/10.1002/spe.3235
76. Ali, Y.M.B.: Adversarial attacks on deep learning networks in image classification based on smell bees optimization algorithm. Futur. Gener. Comput. Syst. **140**, 185–195 (2023)
77. Alkharabsheh, K., Alawadi, S., Kebande, V.R., Crespo, Y., Fernández-Delgado, M., Taboada, J.A.: A comparison of machine learning algorithms on design smell detection using balanced and imbalanced dataset: a study of god class. Inform. Softw. Technol. **143**, 106736 (2022)
78. Gupta, R., Kumar Singh, S.: A novel metric based detection of temporary field code smell and its empirical analysis. J. King Saud Univ.: Comput. Inform. Sci. **34**, 9478–9500 (2022)
79. Shoenberger, I., Mkaouer, M.W., Kessentini, M.: On the use of smelly examples to detect code smells in javascript. In: Squillero, G., Sim, K. (eds.) Appl. Evol. Comput., pp. 20–34. Springer International Publishing, Cham (2017)
80. Palomba, F., Bavota, G., Di Penta, M., Fasano, F., Oliveto, R., De Lucia, A.: On the diffuseness and the impact on maintainability of code smells: a large scale empirical investigation, in: Proceedings of the 40th International Conference on Software Engineering, pp. 482–482 (2018)
81. Chen, Q., Câmara, R., Campos, J., Souto, A., Ahmed, I.: The smelly eight: An empirical study on the prevalence of code smells in quantum computing, in: 2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE), IEEE, pp. 358–370 (2023)

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

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

Dr. Ruchika Malhotra is a Professor and Head of Department at the Department of Software Engineering, Delhi Technological University, Delhi, India. She was awarded the prestigious Raman Fellowship for pursuing Postdoctoral research at Indiana University Purdue University Indianapolis, USA. She received the IBM Faculty Award in 2013. Her h-index is 36, as reported by Google Scholar. She is the author of a book titled "Empirical Research in Software Engineering" published by CRC Press and co-author of a book on "Object-Oriented Software Engineering" published by PHI Learning. Her research interests are software testing, improving software quality, statistical and adaptive prediction models, software metrics, and the definition and validation of software metrics. She has

published more than 220 research papers in international journals and conferences and is amongst the world's top 2% scientists as per reports from Stanford University.

Ms. Bhawna Jain is pursuing Ph.D. in Computer Science from Delhi Technological University (DTU)-Oakland University, USA. She is an Assistant Professor in the Department of Computer Science and Engineering at Indira Gandhi Delhi Technical University for Women (IGDTUW). She has qualified for NET-JRF and GATE twice in 2017 and 2020. She has over three years of experience working as a Software Developer in the industry. She has won awards like All Rounder Performer and ACE Excellence award for being Growth Catalyst and received appreciation for her remarkable efforts in the R &D department during her industry experience. She also contributed to the Guinness Book of World Records for the "Largest online video chain of people handing over an object". She has research publications with IEEE and Springer. Her areas of interest are Artificial Intelligence, Deep Learning, Software Maintenance, and Evolutionary Computing.

Dr. Marouane Kessentini is working as Winegarden Professor and Associate Dean for Research and Graduate Studies at University of Michigan-Flint College of Innovation and Technology, Michigan, USA. He is the recipient of the prestigious 2018 President of Tunisia Distinguished Research Award, the University Distinguished Teaching Award, the University Distinguished Digital Education Award, the College of Engineering and Computer Science Distinguished Research Award, four Best Paper awards, including an IEEE 10-Year Most Influential Paper award (2011-2021), and his AI-based software refactoring invention, licensed and deployed by several Fortune 500 companies, and selected as one of the Top 8 inventions at the University of Michigan for 2018 (including the three campuses) by the UM Technology Transfer Office. He received grants from industry and federal agencies and published over 175 papers in top journals and conferences. His h-index is 46, as reported by Google Scholar.

ISSN 1392-1207. MECHANIKA. Vol. 29, No. 5, 2023: 432–437

Experimental Evaluation of CNG Substitution Ratio on Exhaust Gas Emissions of Diesel/CNG Dual Fuel Combustion

Neeraj KUMAR, Bharat Bhushan ARORA, Sagar MAJI
Mechanical Engineering Department, Delhi Technological University, New Delhi 110042, India,
E-mail: neev37@gmail.com

https://doi.org/10.5755/j02.mech.33638

Abbreviations

CO – Carbon monoxide; CR – Compression Ratio; CNG – Compressed Natural Gas; EGR – Exhaust Gas Recirculation; BTE – Break Thermal Efficiency; BSFC – Break Specific Fuel Consumption; COPD – Chronic Obstructive Pulmonary Disease; CN – Cetane Number; MN – Methane Number; RON – Research Octane Number; DF – Dual Fuel; SI – Spark Ignition; CI – Compression Ignition; CO_2 – Carbon Dioxide; PM – Particulate Matter; HC – Hydrocarbon; IC – Internal Combustion; NO_X – Nitrogen Oxide; CRDI – Common Rail Direct Injection; VCR – Variable Compression Ratio.

1. Introduction

Diesel engines are extensively used in the transportation, agricultural, and industrial sectors worldwide due to their superior energy efficiency, reliability, adaptability, and cost-effectiveness than gasoline equivalents [1]. However, the emissions from diesel engines contribute to climate change and environmental deterioration. Ignition of diesel engines produced pollutants such as HC, CO, NOx, CO_2 and smoke, and acoustic sound [2, 3]. These pollutants increase the greenhouse effect and negatively impact the health of people and the environment in many ways [4]. Many studies have shown that diesel engine emissions may cause COPD, coughing, breathing, wheezing, asthma, respiratory illnesses and premature death in humans [5]. So, to achieve the lowest possible emissions of pollutants, emission control policies worldwide reduce PM and NOx emissions from diesel engines [6]. Further, the transportation energy demand is increasing rapidly, but the availability of fossil fuels is limited. So, despite the depletion of petroleum supplies, it is necessary to find a compromise between reducing emissions from IC engines by more strict emission standards and expanding energy consumption in all modes of transportation. Therefore, it is necessary to use alternative fuels in the existing CI engines, partially or completely replacing conventional fuels. Compressed natural gas is one of the best alternative fuels to reduce exhaust gas pollutants compared to diesel and gasoline combustion [7]. Compressed natural gas has additional benefits over fossil fuels. Since it is less expensive, has more reserves, has high octane number, and has a high calorific value and can be used in both SI and CI engines in dual fuel mode [8-10]. Regarding fossil fuel depletion and environmental pollution, the DF technique is becoming more interesting for researchers. Its combustion technology is efficient and environmentally friendly, and cost-efficient. Many experimental studies have been performed to describe the combustion characteristics and exhaust emission of dual fuel CI engines with CNG and diesel varying various operational parameters such as engine load, engine speed, compression ratio, injection strategy, diesel fuel quantity, EGR and others [11-18]. Increased in CNG substitution ratio, decreased brake thermal efficiency, CO_2 and NO_X while increasing CO and HC emissions at all tested loads [19]. Increasing the CNG substitution ratio from 0 to 45 per cent, combustion pressure and HRR increased, and HC and NOx emissions were reduced. Beyond 45 to 95 percent CNG substitution ratio increases the NOx emissions and decreases CO emissions [20]. While investigating the effect of a dual fuel (Diesel-CNG) operating model to analyse the performance and emissions at moderate and relatively high loads. It is found that a reduction in specific fuel consumption, NOx, and soot while increasing carbon CO and HC emissions [21]. The performance and emission characteristics of a diesel engine operating in dual-fuel mode with CNG and diesel, with diesel proportions varying from 10% to 100% and engine loads ranging from 1.1 KW to 2.8 KW. And the findings indicate that BTE decreases and increases BSFC in DF mode than pure diesel fuel mode. At 1.1 KW load, a maximum decrease in efficiency and increased growth of 68% of specific fuel consumption were noted in dual-fuel combustion mode. The exhaust emission of CO_2 and soot was lower and increased CO in DF combustion [22]. Under DF combustion at various load conditions. The performance of engine increases while emissions were decreases [23]. The performance and exhaust emissions are improved in diesel CNG dual fuel combustion. The energy shares of CNG 80% recorded a 2.7% increment in BTE and a 22.6% reduction in CO_2 emission than 100% diesel fuel combustion [24]. This work's interest is to exploit the dual fuel (diesel/CNG) engine combustion with minimal exhaust emission for power production. The exclusiveness of the present work lies in the fact that scant investigation is reported on diesel CNG dual-fuel engine combustion by the researcher's varying CNG substitution ratios and compression ratios. The study's objective was to investigate the impact of CNG substitution at different compression ratios to analyse the gaseous exhaust emissions of stationary single-cylinder, four strokes, water-cooled, CRDI VCR, DF (diesel/CNG) engines. The CR from 13:1 to 15:1 and CNG substitution of 40%, 60% and 80% were used to take the benefit of a higher octane number of CNG. The test set-up is regulated at an engine speed of 1500 rpm with a constant engine load of 10 Kg.

2. Experiment methodology

2.1. Experiment set up

Fig. 1 illustrates the schematic diagram of the experimental set-up consisting of a single-cylinder, four-stroke, CRDI VCR engine. The test engine configuration is shown in Table 1. The test engine was coupled to a SAJ make water-cooled eddy current dynamometer, Model AG10, to measure the engine load. A K-type thermocouple was used to measure the temperatures of exhaust gases and intake air**.** The CNG cylinder kit is coupled to the inlet manifold through a rubber hose. A Programmable electronic control unit (ECU), Model Nira i7r, measures and controls the CNG flow rate. AVL Di Gas 444N gas analyzer and AVL 437 smoke meter measured the exhaust gas emissions and smoke. The physicochemical properties of diesel and CNG fuels are listed in Table 2. The technical measurement of the exhaust gas analyser are given in Table 3.

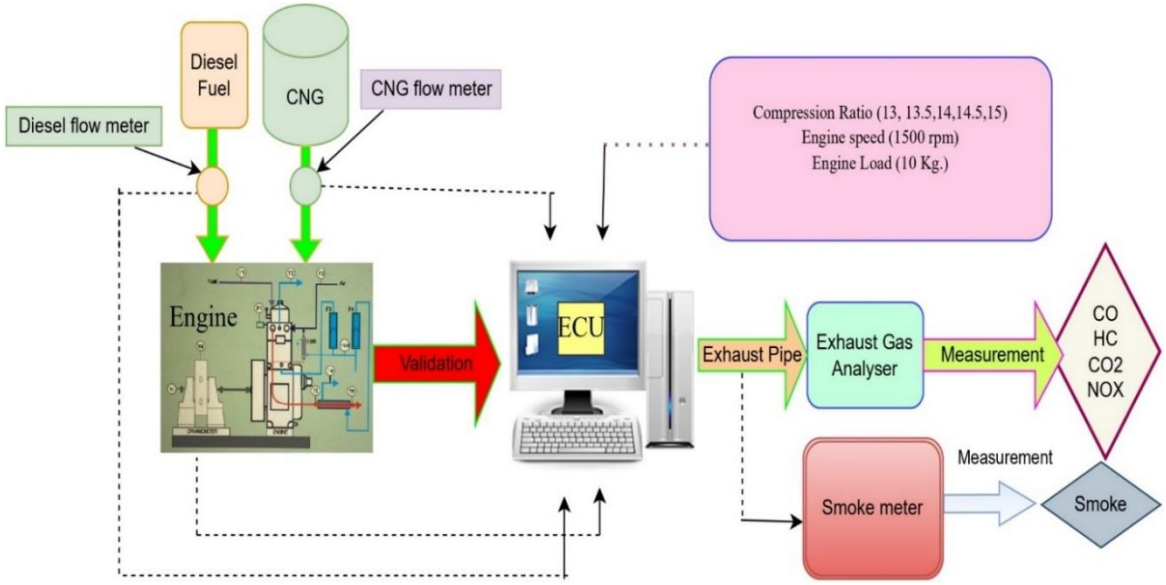


Fig. 1 Schematic diagram of experiment setup

Table 1

Technical data of the experimental engine setup

Parameters	Description
Engine	Make Kirloskar, Single cylinder,4-stroke,CRDI VCR engine
Bore, mm Stroke, mm	87.5 110
Type of cooling	Water-cooled
Displacement, cc	661
Rated power, kW Rated speed, rpm	3.5kW 1500
Compression ratio used	13, 13.5, 14, 14.5 and 15
Variable compression ratio range	12:1 - 18:1
Overall dimension, mm	W 2000 × D 2500 × H 1500
Fuel used	Diesel, Diesel/CNG dual fuel

Table 2

Fuel properties

Properties	Diesel	CNG
CN	52	0
MN	0	82
RON	20	130
Density, kg/m^3	830	0.72
Fuel calorific value, KJ/kg	42000	50000
Stoichiometric A/F ratio	14.5	17.05

Table 3

Technical data of exhaust gas analyser

Emission	Measurement	Resolution
CO	0-15% Vol.	0.001% Vol.
HC	0-20000 ppm Vol.	1 ppm/10 ppm
CO_2	0-20% Vol.	0.1% Vol.
NO_X	0-5000 ppm Vol.	1 ppm Vol.

2.2. Methodology

An experiment using D100% fuel and diesel/CNG DF combustion modes were conducted to examine the engine exhaust gas emissions. A single-cylinder, four-stroke, water-cooled, VCR diesel engine was used as the test set-up. The experiment test was run with an engine speed of 1500 rpm and CR from 13:1 to 15:1 with a constant engine load of 10 Kg. was used to evaluate the investigation. The experimentation analysed the influence of CNG substitution ratios in the A/F mixture throughout combustion with diesel fuel. The CNG substitution ratio was calculated by Eq. (1).

$$CNG,\% = \frac{\dot{m}_{CNG} CV_{CNG}}{\dot{m}_{CNG} CV_{CNG} + \dot{m}_{Diesel} CV_{Diesel}} \times 100, \qquad (1)$$

where: $\dot{m}_{CNG}$ and $\dot{m}_{Diesel}$ are the fuel mass flow rate of CNG and diesel fuel in Kg/s, and CV_{CNG} and CV_{Diesel} are the net calorific value of CNG and diesel fuel.

Two steps were used to analyses the experiment. At CR of 18:1, the engine fired up using D100% fuel in the first step. The CR was adjusted from 13:1 to 15:1 in steps of 0.5:1 by tilting the head arrangements from the cylinder after achieving the thermal stabilization state and the engine speed and load varying from 1500 rpm and 10 kg. The combustion performance of single, double, and triple injections was also compared under these operating parameters. The engine exhaust gas emissions and smoke was recorded by the AVL Di Gas 444N gas analyzer and AVL 437 smoke meter. For DF combustion in the second step, diesel and CNG were used. CNG was mixed in the intake manifold's

air stream, and diesel was injected after the compression of the CNG air mixture and adjusting the CNG substitution ratios of CNG40%, CNG60%, and CNG80% from the CNG supply valve. The same instruments used in the first step in DF combustion recorded the engine exhaust gas emissions and smoke.

3. Results and discussion

3.1. Carbon Monoxide emission

The primary influence of CO emission is low flame temperature and poor air/fuel mixture from the engine. Generally, CO emission increases as increasing in the CNG substitution ratio [21, 22]. With increasing CNG supply, the ignition fuel quantity and spray combustion region decrease at a lower CNG substitution ratio. The lower equivalence ratio in the CNG/air mixture indicates that the quenching region expands, and the CO emissions increase more than pure diesel. For a higher CNG substitution ratio, the equivalence ratio rises by increasing the CNG quantity and reducing pilot fuel quantity, decreasing the ignition temperature and leading to CO emissions.

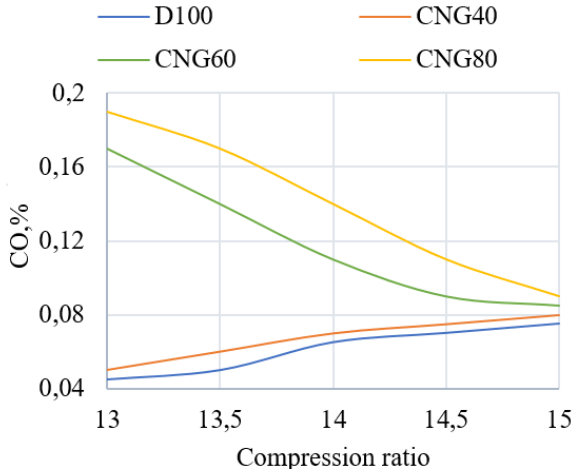


Fig. 2 Variation of CO emission versus compression ratio

Fig. 2 shows that CO emission increased as increased the CNG substitution ratio than D100% combustion. In comparison to D100, CO emissions increased by 11.1%, 20%, 7.6%, 7.1% and 7% at all CR for CNG40% DF combustion, as shown in Fig. 2. Similarly, The CO emissions raised by 277.7%, 180%, 69.2%, 28.5% and 13.3% at CNG60% substitution ratio than D100 fuel combustion. The same pattern was observed for CNG80% substitution ratio, and CO emission increased by 322.2%, 240%, 115.5%, 57.1% and 20% than D100% fuel at CR of 13:1 to 15:1.

3.2. Hydrocarbon emission

Hydrocarbon emission is imprimis from incomplete combustion as scarcity of oxygen and lower combustion rates in dual-fuel combustion. Fig. 3 shows that the HC emission increased with an increase in the CNG substitution ratio than D100 combustion. Usually, the HC emission is substantially higher compared to single diesel fuel combustion in DF mode [19, 25]. It is because ambient flame extinction zones in the CNG/air mixture are distant from the injection spray zones at a lower CNG substitution ratio; thus, the HC emissions are produced from the incomplete combustion of diesel/CNG DF combustion. In a higher CNG substitution ratio, the combustion throughout the spray zones is reformed because of the rich mixture; however, the quantity of CNG/air mixture compressed into the cylinder from the intake manifold during the compression stroke increases, increasing HC emissions. At higher CNG substitution, reduction in pilot fuel decreases the local temperature and ignition energy during expansion, which may lead to imperfect combustion and produce higher HC pollutants. The HC emissions are higher at CNG40% DF combustion compared to D100% fuel by 46%, 40.1%, 30.3%, 16.6% and 6.06% at all CR. At CNG60% substitution ratio, HC emissions increased by 98.6%, 66.6%, 70%, 105.5% and 142.4% compared to D100% combustion. The substitution ratio of CNG80% DF combustion increased HC emissions by 180%, 144.4%, 130%, 161.1% and 187.8%, respectively, compared to D100% fuel at CR of 13:1 to 15:1.

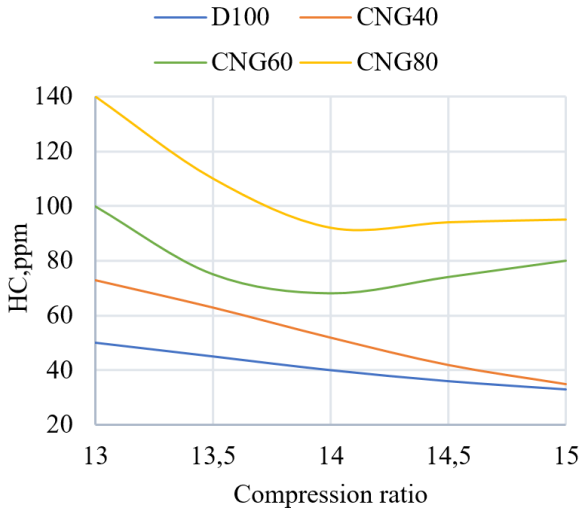


Fig. 3 Variation of HC emission versus compression ratio

3.3. Carbon Dioxide emission

Carbon dioxide is a leading element of greenhouse gases responsive to global warming. The scarcity of carbon atoms and oxygen content with CNG substitution influences CO_2 emission. Generally, the CO_2 emission is lower compared to single diesel fuel combustion in DF mode. CO_2 emission decrease with an increase in the CNG substitution ratios compared to D100% combustion [10, 26, 27].

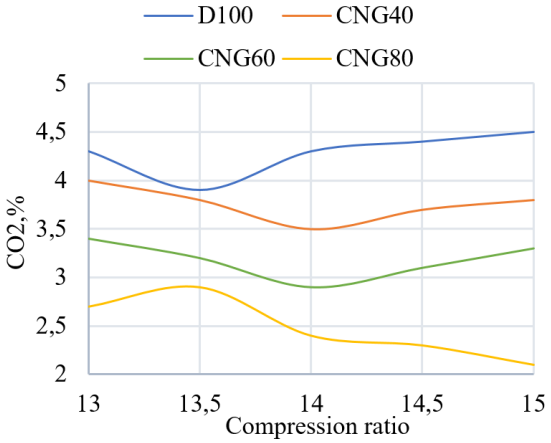


Fig. 4 Variation of CO_2 emission versus compression ratio

Fig. 4 shows CO_2 emission was lower in DF combustion at all CR. At the CNG40% substitution ratio, the CO_2 emissions are reduced by 7.5%, 2.6%, 22.4%, 18.9% and 18.4% at all CR from 13:1 to 15:1 compared to D100% fuel. The CO_2 emissions decreased by 26.4%, 21.8%, 48.2%, 41.9% and 36.3% for the CNG60% substitution ratio. The same pattern was observed for CNG80% substitution ratio, and CO_2 emissions were reduced by 59.2%, 34.4%, 79.1%, 91.3% and 114.3% than D100% fuel combustion.

3.4. Nitrogen oxides emission

The formation of NOx is facet by the rich oxygen level, high cylinder temperature and long reaction time resulting from combustion. Generally, NO_X emission decreases in DF mode [28, 29]. In diesel/CNG DF combustion, the combustion process occurs in the lean, premixed regime due to increased Stoichiometric air/fuel ratio, decreasing the cylinder temperature and decreasing NO_X emission. For single diesel combustion, maximum fuel is burned as a diffusion flame near the stoichiometric equivalence ratio resulting in stoichiometric combustion producing higher NO_X due to higher combustion temperatures. Fig. 5 shows that NO_X emission decreases with increases in CNG substitution ratios compared to D100% combustion. At CNG40% DF combustion substitution ratio, NO_X emissions are reduced by 3.02%, 10.5%, 25.9%, 19.1% and 16.9% at all CR. The NO_X emissions were reduced by 222.3%, 207%, 204%, 103.4% and 52.7% for CNG60% substitution ratio DF combustion. The same pattern was observed for the CNG80% substitution ratio, and NO_X emissions decreased by 396.8%, 746.7%, 990%, 904% and 840.5%, respectively, compared to D100% fuel combustion at CR of 13:1 to 15:1.

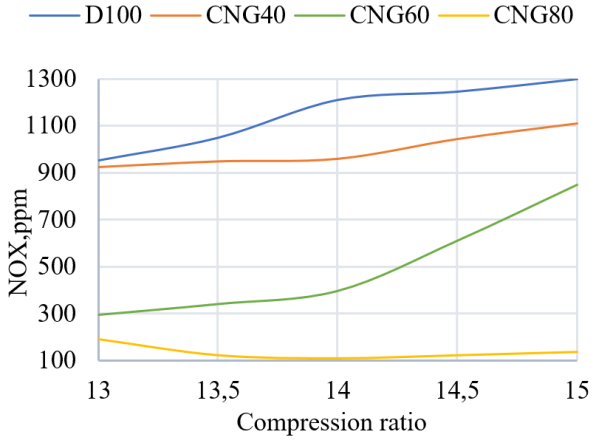


Fig. 5 Variation of NO_X emission versus compression ratio

3.5. Smoke emission

Fig. 6 shows that the smoke opacity changes for different CNG substitution ratios. The volume of liquid fuel injected is the most significant factor affecting the smoke opacity. A dual-fuel engine needs less diesel fuel since it replaces with a cleaner-burning CNG fuel and produces lower smoke. A higher CNG substitution ratio improves the fuel mixture combustion and lowers the C/H ratio, which leads to less smoke and more heat contribution from the CNG fuel. CNG's clean-burning gas quality helps reduce soot emissions and promotes soot oxidation at a higher CNG substitution ratio [30, 31]. Since CNG is mostly made of methane and belongs to the lower paraffin family, it does not tend to produce considerable soot. At a lower CNG substitution ratio, the heat input contribution of CNG fuel is reduced, and pilot liquid fuel raises the combustion carbon content and smoke levels compared to the high CNG substitution ratio. Fig. 6 demonstrates that as the CNG substitution increases, the smoke level decreases compared to D100% combustion. When the CNG40% substitution ratio is used in DF combustion, the smoke emissions are reduced by 35.2%, 64.7%, 14.2%, 38.6% and 50% at CR from 13:1 to 15:1 compared to D100% fuel combustion. The smoke emissions decreased by 17.1%, 16.6%, 23%, 16% and 125% for the CNG60% substitution ratio. Similarly, the same pattern was at the CNG80% substitution ratio, and the smoke emission was reduced by 68.4%, 194.6%, 33.5%, 55.3% and 350% than D100% fuel combustion for all CR.

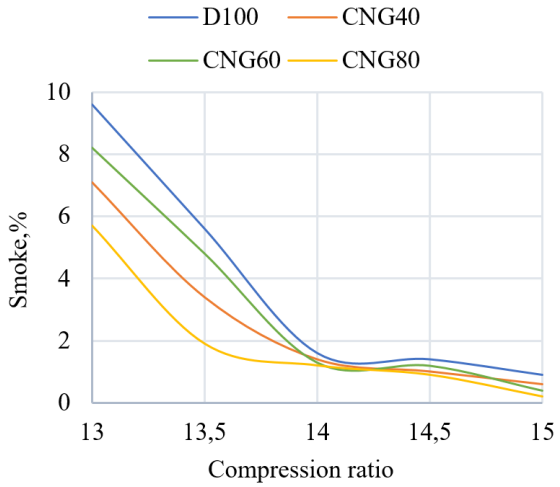


Fig. 6 Variation of smoke opacity versus compression ratio

4. Conclusion

The experiment investigates the effects of the CNG substitution ratio of CNG40%, CNG60%, and CNG80% in DF combustion and varying compression ratios from 13:1 to 15:1 of a CI engine to analyse the exhaust gas emissions characteristics. The investigation's main points are highlighted.

- Carbon Monoxide (CO) and hydrocarbon (HC) emissions were increased in DF combustion compared to single diesel fuel combustion. The substitution of CNG80% produced higher CO and HC emissions. When CNG is used more frequently, the equivalency ratios increase along with the CNG supply rate, which reduces the amount of liquid fuel and lowers the mixture's ignition temperature and spark energy. These facts increase CO and HC emissions in DF combustion compared to D100% combustion.
- The emission of CO_2 was lower in DF combustion than in pure diesel combustion at all CNG substitution ratios. CNG has a higher calorific value and lower carbon content making it a better alternative fuel to reduce CO_2 emissions than pure fuel diesel combustion. The maximum reduction of CO_2 emission was attained at the CNG80% substitution ratio at all CR.
- The amount of NOx and smoke emission in the exhaust gas is reduced when CNG is mixed with diesel fuel. At the CNG80% substitution ratio, the lowest NO_X and

smoke emissions were achieved. Higher CNG substitution ratio results in longer ignition delay times, slower flame propagation speeds, lower cylinder temperatures, and lower C/H ratios, which reduce NO_X emissions and smoke emissions during DF combustion.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. **Cheng, Q.; Kaario, O.; Ahmad, Z.; Vuorinen, V.; Larmi, M.** 2021. Effect of pilot fuel properties on engine performance and combustion stability in a tri-fuel engine powered by premixed methane-hydrogen and diesel pilot, International journal of hydrogen energy 46(75): 37469-37486. https://doi.org/10.1016/j.ijhydene.2021.09.053.
2. **Zwierzchowski, R.; Różycka-Wrońska, E.** 2021. Operational determinants of gaseous air pollutants emissions from coal-fired district heating sources, Archives of Environmental Protection. https://doi.org/10.24425/aep.2021.138469.
3. **Hooftman, N.; Messagie, M.; Van Mierlo, J.; Coosemans, T.** 2018. A review of the European passenger car regulations–Real driving emissions vs local air quality, Renewable and Sustainable Energy Reviews 86: 1-21. https://doi.org/10.1016/j.rser.2018.01.012.
4. **Shu, J.; Fu, J.; Zhang, Y.; Xie, M.; Liu, J.; Liu, J.; Zeng, D.** 2020. Influences of natural gas energy fraction on combustion and emission characteristics of a diesel pilot ignition natural gas engine based on a reduced chemical kinetic model, Fuel 261: 116432. https://doi.org/10.1016/j.fuel.2019.116432.
5. **Manisalidis, I.; Stavropoulou, E.; Stavropoulos, A.; Bezirtzoglou, E.** 2020. Environmental and health impacts of air pollution: a review, Frontiers in public health p. 14. https://doi.org/10.3389/fpubj.2020.00014.
6. **Roy, M. M.; Wang, W.; Alawi, M.** 2014. Performance and emissions of a diesel engine fueled by biodiesel–diesel, biodiesel–diesel-additive and kerosene–biodiesel blends, Energy conversion and management 84: 164-173. https://doi.org/10.1016/j.enconman.2014.04.033.
7. **Chen, H.; He, J.; Zhong, X.** 2019. Engine combustion and emission fuelled with natural gas: a review, Journal of the Energy Institute 92(4): 1123-1136. https://doi.org/10.1016/j.joei.2018.06.005.
8. **Gürbüz, H., Sandalcı, T.** 2023. Numerical analysis of diesel injection strategies on emissions and performance in CH4/diesel powered RCCI diesel engine with high ratio EGR, Alexandria Engineering Journal 64: 517-526. https://doi.org/10.1016/j.aej.2022.09.012.
9. **Shim, E., Park, H., & Bae, C.** 2018. Intake air strategy for low HC and CO emissions in dual-fuel (CNG-diesel) premixed charge compression ignition engine, Applied energy 225: 1068-1077. https://doi.org/10.1016/j.apenergy.2018.05.060.
10. **Karczewski, M., & Szamrej, G.** 2023. Experimental evaluation of the effect of replacing diesel fuel by CNG on the emission of harmful exhaust gas components and emission changes in a dual-fuel engine, Energies 16(1): 475. https://doi.org/10.3390/en16010475.
11. **Wang, Z.; Zhang, F.; Xia, Y.; Wang, D.; Xu, Y.; Du, G.** 2021. Combustion phase of a diesel/natural gas dual fuel engine under various pilot diesel injection timings, Fuel 289: 119869. https://doi.org/10.1016/j.fuel.2020.119869.
12. **Pathak, S. K.; Nayyar, A.; Goel, V.** 2021. Optimization of EGR effects on performance and emission parameters of a dual fuel (Diesel+ CNG) CI engine: An experimental investigation, Fuel 291: 120183. https://doi.org/10.1016/j.fuel.2021.120183.
13. **Sudarmanta, B.; Setiyawan, A.; Putra, A. B. K.; Yuvenda, D.; Silva, J. D.** 2019. Optimization of pilot diesel injection timing on load variation dual fuel diesel-CNG engine on combustions and emissions characteristics, Int Rev. Mech. Eng., p. 13. https://doi.org/10.15866/ireme.v13i1.16409.
14. **Rai, A. A.; Bailkeri, N. K.; BR, S. R.** 2021. Effect of injection timings on performance and emission Characteristics of CNG diesel dual fuel engine, Materials Today: Proceedings 46: 2758-2763. https://doi.org/10.1016/j.matpr.2021.02.509.
15. **Lyu, M.; Zhang, C.; Bao, X.; Song, J.; Liu, Z.** 2017. Effects of the substitution rate of natural gas on the combustion and emission characteristics in a dual-fuel engine under full load, Advances in Mechanical Engineering 9(12): 1687814017747158. https://doi.org/10.1177/1687814017747158.
16. **Lee, S.; Kim, C.; Lee, S.; Lee, J.; Kim, J.** 2020. Diesel injector nozzle optimization for high CNG substitution in a dual-fuel heavy-duty diesel engine, Fuel 262: 116607. https://doi.org/10.1016/j.fuel.2019.116607.
17. **Sun, L.; Liu, Y.; Zeng, K.; Yang, R.; Hang, Z.** 2015. Combustion performance and stability of a dual-fuel diesel–natural-gas engine, Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering 229(2): 235-246. https://doi.org/10.1177/0954407014537814.
18. **Ulishney, Christopher J.; Cosmin E. Dumitrescu**. 2022. Effect of gas composition on the performance and emissions of a dual-fuel diesel-natural gas engine at low load conditions, Fuel 324: 124531. https://doi.org/10.1016/j.fuel.2022.124531.
19. **Tripathi, G.; Sharma, P.; Dhar, A.** 2020. Effect of methane augmentations on engine performance and emissions, Alexandria Engineering Journal 59(1): 429-439. https://doi.org/10.1016/j.aej.2020.01.012.
20. **Jamrozik, A.; Tutak, W.; Grab-Rogaliński, K.** 2019. An experimental study on the performance and emission of the diesel/CNG dual-fuel combustion mode in a stationary CI engine, Energies 12(20): 3857. https://doi.org/10.3390/en12203857.
21. **Lounici, M. S.; Loubar, K.; Tarabet, L.; Balistrou, M.; Niculescu, D. C.; Tazerout, M.** 2014. Towards improvement of natural gas-diesel dual fuel mode: An experimental investigation on performance and exhaust emissions, Energy 64: 200-211. https://doi.org/10.1016/j.energy.2013.10.091.
22. **Bari, S.; Hossain, S. N.** 2019. Performance of a diesel engine run on diesel and natural gas in dual-fuel mode of operation, Energy Procedia 160: 215-222.

https://doi.org/10.1016/j.egypro.2019.02.139.
23. **Yousefi, A.; Guo, H.; Birouk, M.** 2019. Effect of diesel injection timing on the combustion of natural gas/diesel dual-fuel engine at low-high load and low-high speed conditions, Fuel 235: 838-846. https://doi.org/10.1016/j.fuel.2018.08.064.
24. **Lee, C. F.; Pang, Y.; Wu, H.; Nithyanandan, K.; Liu, F.** 2020. An optical investigation of substitution rates on natural gas/diesel dual-fuel combustion in a diesel engine, Applied Energy 261: 114455. https://doi.org/10.1016/j.apenergy.2019.114455.
25. **Rotaru, S.** 2022. Experimental study on combustion of an automotive diesel engine fuelled with CNG. In IOP Conference Series: Materials Science and Engineering 1262(1): 012072. https://doi.org/10.1088/1757-899X/1262/1/012072.
26. **Stettler, M. E.; Midgley, W. J.; Swanson, J. J.; Cebon, D.; Boies, A. M.** 2016. Greenhouse gas and noxious emissions from dual fuel diesel and natural gas heavy goods vehicles, Environmental science & technology 50(4): 2018-2026. https://doi.org/10.1021/acs.est.5b04240.
27. **Rotaru, S.; Pana, C.; Negurescu, N.; Cernat, A.; Fuiorescu, D.; Nuţu, C. N.** 2020. Effects of CNG quantity on combustion characteristics and emissions of a dual fuelled automotive diesel engine, In E3S Web of Conferences 180: 01008). https://doi.org/10.1051/e3sconf/202018001008.
28. **Santoso, A.; Sampurno, B.; Cahyono, B.** 2021. Modifying single-cylinder diesel engine become dual-fuel system to reduce NOx emission and maintain fuel efficiency, In IOP Conference Series: Materials Science and Engineering 1052(1): 012003. https://doi.org/10.1088/1757-899X/1052/1/012003.
29. **Nithyanandan, K.; Zhang, J.; Li, Y.; Meng, X.; Donahue, R.; Lee, C. F.; Dou, H.** 2016. Diesel-like efficiency using compressed natural gas/diesel dual-fuel combustion, Journal of Energy Resources Technology 138(5): 052201. https://doi.org/10.1115/1.4032621.
30. **Zhou, H.; Li, X.; Lee, C. F. F.** 2019. Investigation on soot emissions from diesel-CNG dual-fuel. international journal of hydrogen energy 44(18): 9438-9449. https://doi.org/10.1016/j.ijhydene.2019.02.012.
31. **Dimitrov, E.; Tashev, A.** 2022. CNG effect on performance parameters of diesel engine operated on dual-fuel mode, In AIP Conference Proceedings 2449(1): 050009. https://doi.org/10.1063/5.0091441.

N. Kumar, B. B. Arora, S. Maji

EXPERIMENTAL EVALUATION OF CNG SUBSTITUTION RATIO ON EXHAUST GAS EMISSIONS OF DIESEL/CNG DUAL FUEL COMBUSTION

S u m m a r y

The addition of Compressed natural gas as a complement to diesel in compression ignition engines in dual-fuel combustion mode is a viable technology for increasing efficiency and lowering emissions. This work investigates the impact of a dual-fuel operating mode on the engine exhaust pollutant emissions of a diesel engine using compressed natural gas as the principal fuel and neat diesel as the pilot fuel. Compressed natural gas was injected into an intake manifold of a single-cylinder diesel test engine under different engine operating parameters, and up to 80% substitution was attained. And diesel fuel was injected after the compressed natural gas air mixture was compressed. The tests were carried out at five different compression ratios ranging from 13:1 to 15:1 in steps of 0.5:1. The experiment study revealed that injecting CNG into diesel engines via dual fuel combustion significantly impacted exhaust gas emissions compared to pure diesel combustion. The Carbon monoxide (CO) and hydrocarbon (HC) emissions were increased, while carbon dioxide (CO_2), nitrogen oxide (NO_X) and smoke opacity were decreased in dual fuel combustion compared to single diesel fuel.

Keywords: diesel, CNG, dual fuel, emissions, combustion.

Received March 14, 2023
Accepted October 9, 2023

This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 (CC BY 4.0) License (http://creativecommons.org/licenses/by/4.0/).

This article has been accepted for publication in IEEE Transactions on Power Delivery. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRD.2023.3318689

Fault Location Algorithm for Multi-terminal Radial Medium Voltage DC Microgrid

Vaibhav Nougain, Sukumar Mishra, *Senior Member, IEEE*, Soumya Shubhra Nag, *Member, IEEE* and Aleksandra Lekić, *Senior Member, IEEE*

***Abstract*—Accurately locating the fault helps in the rapid restoration of the isolated line back into the system. This paper proposes a novel communication-based multi-terminal method to locate the fault in a radial medium voltage DC (MVDC) microgrid. A time-domain based algorithm is proposed which applies to an MVDC system with different possible combinations of lines and cables. Terminal measurements of voltages and currents, voltages across current limiting reactor (CLR), and node currents are used to propose a flexible online fault location method. Based on the availability of communication and sensors, different terminals can be used to increase the reliability of the proposed fault location method. This method is robust to variations of key implementation parameters like type of faults, fault resistance, fault location, sampling frequency, white Gaussian noise (WGN) in measurement, and different line/cable combinations. Further, the fault location calculation is analyzed with parameter variation. PSCAD/EMTDC based electromagnetic transient simulations are used to validate the performance of the algorithm.**

***Index Terms*—Power system protection, microgrids, Power distribution faults, fault location.**

I. Introduction

MVDC microgrid serves the requirement of supplying the shipping service and propulsion loads in shipboard electrical systems [1]-[3], onboard DC marine vessels [4]-[5], and remote area mine sites [6]. This is a result of various advantages [7] including (i) lesser losses, (ii) lesser complexity with a lesser number of conversion stages in the system, (iii) relaxed synchronization requirements, and (iv) easy integration of multiple energy storage systems. Such multi-terminal MVDC microgrids, when subjected to a fault contingency experience a high rate of rise of current due to discharging converter capacitor [1]. The interfacing converter is prone to get damaged in case of improper control logic [7]. Even if the control is proper, the converter control is bypassed to limit the current [8] affecting the power transfer in the system. Either way, there is an indispensable requirement for a fast, selective, and reliable protection approach that ensures continuous operation of power transfer even under a fault [1]. Protection has two broad challenges, (i) fault identification and (ii) fault location. Identifying the fault helps in the isolation of the faulty line and accurately locating the DC fault ensures rapid restoration of the isolated faulty line into the DC system.

Vaibhav Nougain and Aleksandra Lekić are with the Department of Electrical Sustainable Energy, Delft University of Technology, 2628 CD Delft, The Netherlands. (e-mail: V.Nougain@tudelft.nl, A.Lekic@tudelft.nl).

Sukumar Mishra and Soumya Shubhra Nag are with the Department of Electrical Engineering, Indian Institute of Technology, Delhi, New Delhi, 110016 India. e-mail: (sukumariitdelhi@gmail.com, ssnag@ee.iitd.ac.in).

Focussing on the aspect of DC fault location, the literature has mainly covered low voltage DC (LVDC) and high voltage DC (HVDC) fault location. Generally, LVDC systems have the attribute of short lines (up to 1km) for distribution and a low value of fault resistance (up to 5Ω) [9]-[17]. Since the cost can be a constraint for such a system, single terminal methods are suitable for fault location in LVDC systems [9]-[17]. Single terminal methods do not require communication i.e., other terminal data to locate the fault. However, the accuracy of single terminal fault location methods has inherent dependence on fault resistance whereas double terminal (or multi-terminal) methods eliminate the dependence [9]. This means double terminal (or multi-terminal) methods are accurate even for high resistance faults. They come with the inherent problem of communication complexity and extra cost [12]. Discussing single terminal fault location methods, the authors in [14]-[16] use a probe power unit (PPU) as an active external injection unit. The method in [14] neglects the damping coefficient in the second order RLC discharging circuit, taking the damped frequency equal to the natural frequency. This results in a deviation in the accuracy of fault distance. The method in [15] uses a least square algorithm to evaluate the damping i.e., attenuation constant. However, when the attenuation constant is high, the accuracy of the method is jeopardized. Authors in [16] use modified PPU to calculate the high value of attenuation constant and thereby, evaluate fault distance with higher accuracy in comparison to [14]-[16]. The authors in [13] use a fault location module, consisting of an inductor, two switches, and two thyristors to achieve an accuracy of 98.4% for low resistance faults in the LVDC system. However, the method in [13] is prone to be affected by WGN in measurement as it uses the differential term (*di/dt*) in the analysis of fault location. Authors in [17] propose an iterative method to evaluate fault location for pole-to-ground (PTG) faults without any external modification. However, this method has a dead zone for faults closer to the bus and for pole-to-pole (PTP) faults. The method in [9] proposes the use of CLR in the single terminal fault location method. Based on the ratio of transient voltages (ROTV), the analysis is robust and mathematically convincing. The authors in [10] use an improved solid-state circuit breaker (SSCB) to propose a curve-fitting method. However, the parameter extraction method presents low accuracy when the fault occurs close to the bus. The authors in [11] use multiple capacitance earthing as external modifications. However, the method is prone to WGN in measurement as it uses differential terms (*di/dt* and *dv/dt*) in the analysis of fault location.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 08:30:53 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Power Delivery. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRD.2023.3318689

TABLE I: Fault location methods in literature

Fault Location Methods	Method Type/Communication needed?	External modifications	Limitations
[11]	Time-domain LVDC/No	multiple capacitive earthing method	limited to low fault resistance and homogeneous lines
[13]	Time-domain LVDC/Yes	fault location module	limited to low fault resistance and homogeneous lines
[14]-[16]	Time-domain LVDC/No	probe power unit	limited to low fault resistance and homogeneous lines
[17]	Time-domain LVDC/No	No	limited to homogeneous lines
[20]	Time-domain HVDC/Yes	current limiting reactor	limited to homogeneous lines
[21]	Travelling-wave non-homogeneous HVDC/Yes	current limiting reactor	limited to fault resistance, $R_f < 100\,\Omega$
[24]	Time-domain HVDC/Yes	current limiting reactor	limited to homogeneous lines
[Proposed method]	Time-domain non-homogeneous MVDC/Yes	current limiting reactor	sensitive to parameter variation (much like other time-domain methods)

Generally, HVDC systems have long lines for transmission and the possibility of high fault resistance [19]-[20]. As a result, double-terminal (or multi-terminal) methods are suitable for fault location in such a system [20]-[24]. Long transmission lines come with the additional problem of communication delay, which needs consideration. Discussing double terminal methods, the authors in [20] use time-domain-based simplified R-L representation of transmission lines [18]-[19] to model overhead lines (OHLs) for fault location in HVDC systems. However, the mathematical scope of the work limits the application of the methods to homogeneous lines or cables. The authors in [21]-[22] address the problem of fault location for non-homogeneous DC transmission comprising segments of UGC and segments of OHL. The authors in [21] use the traveling wave-based double terminal method to detect the surge arrival time accurately. The method is limited to a fault resistance of 100Ω and requires a high sampling frequency. The authors in [22] propose a traveling wave type method using continuous wavelet transform applied to a series of line current measurements. The method addresses the problem of [21]. However, multiple distributed current sensors put the viability of the scheme to question. The authors in [24] propose fault location using synchronized currents of the conductors and sheaths of cables. The analysis in [24] is focused on pole-to-pole faults in underground transmission cables (where overhead lines are not included in the studies). The authors in [24] also successfully introduce a generalized fault location analysis for a bigger DC system. The thorough analysis does not use voltage measurements to evaluate fault location. This reduces the cost of the implementation of the method. However, the complexity of the computation of the algorithm is higher due to fewer measured quantities (e.g., evaluation of Z_{seen}). Additionally, the method does not exclusively consider a non-homogeneous transmission for fault location, which limits the application of the method to homogeneous transmission cables only.

MVDC systems have the attribute of short lines for distribution and have the possibility of a high value of fault resistance [1], [7]. This makes double-terminal (or multi-terminal) methods more suitable than single-terminal methods for accurate fault location. Additionally, due to shorter lines, communication delay in the MVDC system is not a concern [6]. To the best of the authors' knowledge, the literature has not covered the application of time-domain-based methods for non-homogeneous lines or cables in MVDC systems. This results from limited mathematical exploration in the area of time-domain-based fault location methods for MVDC systems. This work proposes a novel time-domain-based fault location algorithm for multi-terminal radial MVDC microgrids. The proposed work has the following attributes:

- Unlike single terminal methods [9]-[17], the proposed work is accurate even for high resistance faults.
- Unlike time-domain-based double terminal methods [18]-[20], the method is flexible for systems with either homogeneous or non-homogeneous combinations of cables and lines for distribution.
- Unlike traveling wave-based double-terminal methods [21]-[22], the proposed work requires a sampling frequency as low as 5 kHz (see Section V(D)) and is fairly accurate for high resistance faults.
- The proposed work generalizes the analysis for a larger MVDC system with *k* nodes and *p* buses for broader application. Based on the availability of communication at different terminals, the method can be applied to increase the reliability of fault location.
- The method is robust to variation of key parameters like type of faults, fault resistance, fault location, and white Gaussian noise in the measurement.

The rest of the proposed work is organized as follows. Section II introduces the test system configuration and section III explains the differential fault identification method. Section IV presents the fault location algorithm whereas section V gives the implementation of the proposed method. Section VI validates the proposed fault location algorithm whereas section VII concludes the work.

II. Test System Configuration

A multi-terminal MVDC microgrid operating at ±2kV formed by a radial configuration of 6 buses [24] is used as a test system (see Fig. 1). Bus 1 has a fuel cell, whereas bus 2 has a photovoltaic system interfaced to the system with a boost DC/DC converter. Bus 3 has an energy storage system (ESS) battery interfaced with a bidirectional DC/DC converter. Bus 4 has an AC utility grid connected to the DC system with a voltage source converter (VSC). Bus 5 has grid forming VSC interfaced AC loads, whereas bus 6 has buck DC/DC converter interfaced constant power loads (CPL) [1], [7].

A bipolar line configuration with a frequency-dependent transmission model (FDTL) for UGC and OHL is considered for the system. The UGC is modeled as single-core 280 mm^2 cables with copper conductor, XLPE insulation (2.8 mm thick), and PVC sheathing (2.5 mm thick), whereas

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 08:30:53 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Power Delivery. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRD.2023.3318689

the OHL is modeled as ACSR based stranded conductor with an outer radius of 0.02m and strand radius of 0.003m. The grounding capacitance for UGC and OHL is within 0.5μF/km and 0.01μF/km, respectively, while the equivalent DC capacitors of the converters are around 1-10mF. Hence, the fault contribution from the grounding capacitance of cables and lines can be ignored for short to medium lengths [1]. As a result, the R-L representation of UGC and OHL is used for the fault location analysis. The grounding scheme considered is TN-S which provides mid-point grounding at each converter terminal. The fault contingencies considered in the test system are of different types i.e., PTP, positive-pole to ground (P-PTG), and negative-pole to ground (N-PTG). The sources and loads in the setup are protected using intelligent electronic devices, IED_s which is a combination of DC current transducer, DC overcurrent relay, and DC circuit breakers as shown in Fig. 1. The cables and lines in the system are protected with IED_l using a differential current relay (DCR) as the fault identification scheme [1]. CLRs are used in Fig. 1 to limit the current rise rate during a fault contingency. The CLRs are placed between every IED_s and IED_l. Buses connected to their respective nodes, N_1 & N_2 are used to divide the system shown in Fig. 1 into two parts. Buses 3-6 of the system act as buses 1-4 for node, N_2 whereas buses 1-2 of the system act as buses 1-2 for node, N_1. As a result of which, $SL_{1,2}$ means source line connected at bus 1 of node 2 (or bus 3 of the system), $SL_{2,2}$ means source line connected at bus 2 of node 2 (or bus 4 of the system) and so on. The measurement points for each bus are DC output voltage, voltage after CLR, current through CLRs and adjacent node current (in the case of a multi-terminal system). Node current is important for fault identification using DCR. In case, the node current measurements are unavailable, a localized voltage-based fault identification method [1] can be used along with the DCR.

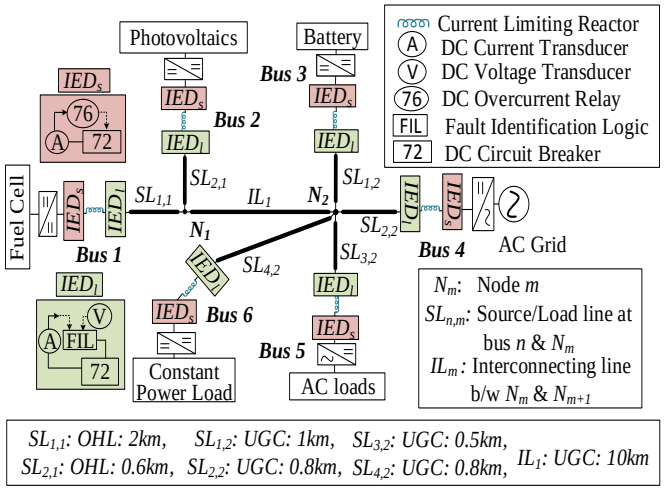


Fig. 1: Medium Voltage DC Test system

III. Fault Identification Algorithm

The fault is identified and classified into different types i.e., PTP, P-PTG and N-PTG at either $SL_{j,i}$ or IL_i. DCR uses bus and node current values to make a trip decision for a fault. To identify and classify faults at $SL_{j,i}$, the trip condition is given by equations (1a)-(1b).

$$\begin{cases} i_{Bj,i^+}(k) = \left| i_{j,i^+}(k) - i_{Ni^+}(k) \right| > I_{set} & \text{(1a)} \\ i_{Bj,i^-}(k) = \left| i_{j,i^-}(k) - i_{Ni^-}(k) \right| > I_{set} & \text{(1b)} \end{cases}$$

where $i_{j,i^+}(k)$ and $i_{j,i^-}(k)$ give the positive and negative pole currents from bus j with closest node i. $i_{Ni^+}(k)$ and $i_{Ni^-}(k)$ give the positive and negative pole currents at node i into the fault. The differential current feeds the fault and increases above a certain threshold, I_{set}. To identify and classify faults at IL_i, the trip condition is given by equations (2a)-(2b).

$$\begin{cases} i_{Nj,i^+}(k) = \left| i_{Ni^+}(k) - i_{N(i+1)^+}(k) \right| > I_{set} & \text{(2a)} \\ i_{Nj,i^-}(k) = \left| i_{Ni^-}(k) - i_{N(i+1)^-}(k) \right| > I_{set} & \text{(2b)} \end{cases}$$

where $i_{N(i+1)^+}(k)$ and $i_{N(i+1)^-}(k)$ give the positive and negative pole currents at node *(i+1)* into the fault.

IV. Fault Location Algorithm

Once the fault is identified, the proposed fault location algorithm takes each segment with different values of R-L per unit. This allows locating fault for a system with any combination of lines or cables (varying from a homogeneous line or n-segment non-homogeneous lines i.e., a combination of OHL and UGC). Any terminal and faulty terminal can be used to calculate accurate fault location. The algorithm's reliability increases if the number of terminals is used to calculate location. The method is initially explained for a 2-node, 3-bus system (see Fig. 2). Further, a more complex system with 4 nodes, and 8 buses is analyzed for the fault location. Finally, the analysis is generalized for a k node, p bus system for wider applications to different MVDC systems.

A. Fault Location for a three bus system

A three bus MVDC system is shown in Fig. 2. Two nodes, N_1 and N_2 with three source/load lines (SL_1, SL_2 and SL_3) and one interconnecting line (IL_1) give four segments in the system. The analysis is elaborated considering node, N_1. The same analysis can be extended using node, N_2. A fault can occur at any of the segments as shown as F_1-F_4 in Fig. 2. The subsection covers a detailed derivation of fault location for different fault cases.

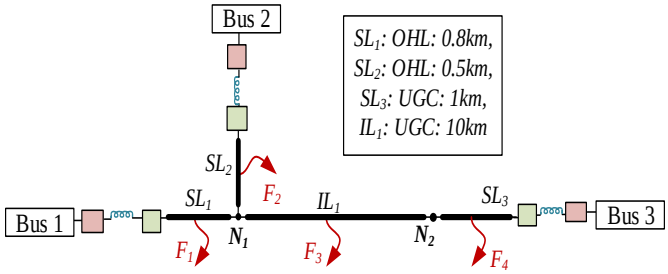


Fig. 2: Generalised *2 node, 3 bus* MVDC system

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 08:30:53 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Power Delivery. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRD.2023.3318689

1) Fault at SL_1: If the fault is located at SL_1, given as F_1 in Fig. 2, the equivalent three-terminal system is given as shown in Fig. 3. Here $v_n(t)$ gives n^{th} DC link bus voltage, $u_n(t)$ gives the voltage after CLR and $i_n(t)$ gives current fed into the fault for n^{th} bus. r_n is the resistance per unit length, l_n is the inductance per unit length and D_n is the total length of source/load line adjacent to n^{th} bus. r_{Mm} is the resistance per unit length, l_{Mm} is the inductance per unit length and M_m is the total length of interconnecting line between m^{th} to $(m+1)^{th}$ node. L is defined as value of CLR and R_f is defined as the fault resistance. Additionally, $x|_{B_n,N_m}$ is the fault distance calculated considering bus n, where $n \in \{1,2,3\}$. m is the node considered for reference along with bus n, where $m \in \{1,2\}$.

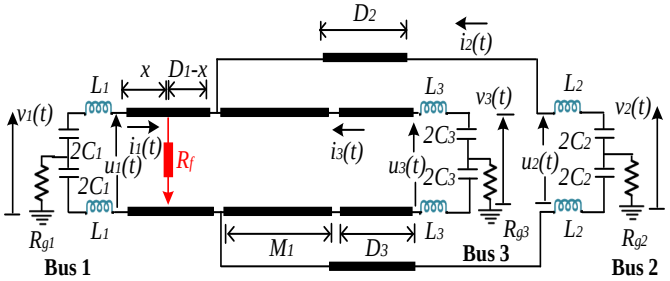


Fig. 3: Simplified *2 node, 3 bus* MVDC system

Using Kirchhoff's Voltage Law (KVL) for a PTP fault at SL_1 (shown in Fig. 3), equations (3)-(5) are obtained for bus 1 to bus 3.

$$2r_1 x i_1(t) + 2l_1 x\frac{di_1(t)}{dt} + R_f\sum i(t) = u_1(t) \quad (3)$$

$$\left.\begin{aligned} &2D_2 r_2 i_2(t) + 2l_2 D_2\tfrac{di_2(t)}{dt} + 2(D_1-x)r_1\left[i_2(t)+i_3(t)\right] \\ &+2l_1(D_1-x)\left(\tfrac{di_2(t)}{dt}+\tfrac{di_3(t)}{dt}\right) + R_f\sum i(t) = u_2(t) \end{aligned}\right\} \quad (4)$$

$$\left.\begin{aligned} &2(D_3 r_3 + M_1 r_{M1})i_3(t) + 2(l_3 D_3 + M_1 l_{M1})\tfrac{di_3(t)}{dt} + \\ &2(D_1-x)r_1\left[i_2(t)+i_3(t)\right] + \\ &2l_1(D_1-x)\left(\tfrac{di_2(t)}{dt}+\tfrac{di_3(t)}{dt}\right) + R_f\sum i(t) = u_3(t) \end{aligned}\right\} \quad (5)$$

Here $\sum i(t)=i_1(t)+i_2(t)+i_3(t)$. The current derivative terms can be obtained using the drop in voltage across the CLR, avoiding substitution errors due to differential calculations [9]. These terms are $\frac{di_1(t)}{dt} = \frac{v_1(t)-u_1(t)}{2L_1} = \frac{v_{L1}(t)}{2L_1}; \frac{di_2(t)}{dt} = \frac{v_2(t)-u_2(t)}{2L_2} = \frac{v_{L2}(t)}{2L_2}; \frac{di_3(t)}{dt} = \frac{v_3(t)-u_3(t)}{2L_3} = \frac{v_{L3}(t)}{2L_3}$. For the sake of simplicity in the analysis, $L_1=L_2=\cdots L_n$. Subtracting any two equations given in equations (3)-(5) negates the dependence of fault location on fault resistance, R_f. If the transition fault impedance has a resistive and an inductive component, equations (3)-(5) are given with an additional common inductive term $\left(l_f\frac{d\left[\sum i(t)\right]}{dt}\right)$. Since the term is common, it is canceled upon simplification. As a result, even if the transition fault impedance has an inductive component, the proposed fault location algorithm is still applicable without any changes. Rearranging equations (3) and (4) define the fault distance from bus 1 of node 1 considering bus 2 of node 1, given as equation (6).

$$x|_{B_2,N_1} = \frac{[u_1(t)-u_2(t)] + 2r_1D_1\left(i_2(t)+i_3(t)\right) + 2r_2D_2i_2(t)}{2r_1\sum i(t) + \frac{l_1\sum v_L(t)}{L}} + \frac{\frac{l_2D_2v_{L2}(t)}{L} + \frac{l_1D_1(v_{L2}(t)+v_{L3}(t))}{L}}{2r_1\sum i(t) + \frac{l_1\sum v_L(t)}{L}} \quad (6)$$

Here $\sum v_L(t) = v_{L_1}(t) + v_{L_2}(t) + v_{L_3}(t)$. Similarly, rearranging equations (3) and (5) define the fault distance from bus 1 considering bus 3 of node 1, given as equation (7).

$$x|_{B_3,N_1} = \frac{[u_1(t)-u_3(t)] + 2r_1D_1\left(i_2(t)+i_3(t)\right)}{2r_1\sum i(t) + \frac{l_1\sum v_L(t)}{L}} + \frac{2\left(r_3D_3 + M_1r_{M1}\right)i_3(t)}{2r_1\sum i(t) + \frac{l_1\sum v_L(t)}{L}} + \frac{\left(l_3D_3+M_1l_{M1}\right)v_{L3}(t)}{L} + \frac{l_1D_1(v_{L2}(t)+v_{L3}(t))}{L}}{2r_1\sum i(t) + \frac{l_1\sum v_L(t)}{L}} \quad (7)$$

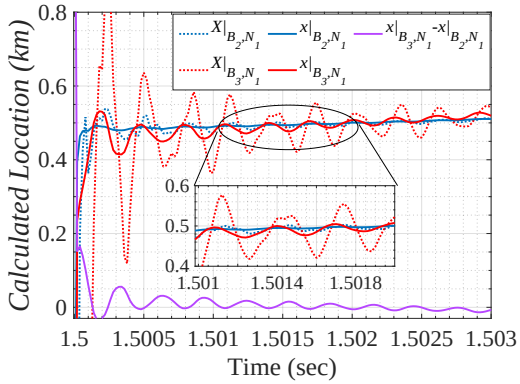


Fig. 4: Calculated location for a fault at SL_1

The equivalent calculated fault location is defined as $x|_{SL_1}$, given by $x|_{SL_1} = \frac{x|_{B_2,N_1}+x|_{B_3,N_1}}{2}$. This is done to reduce random uncertainty in measurement [25] and increase fault location reliability. Here $x|_{B_2,N_1} - x|_{B_3,N_1} \approx 0$ and $\boldsymbol{x} \leq D_1$: $\boldsymbol{x} \in \left\{x|_{B_2,N_1}, x|_{B_3,N_1}\right\}$ are two verification conditions for better reliability. These are validated using Fig. 4, showing a fault at 0.5km at SL_1. Either of the two terminals, along with the faulty terminal, can be used to calculate accurate fault location. The algorithm's reliability increases if the number of terminals is used to calculate fault location. $X|_{B_n,N_m}$ in Fig. 4 shows the calculated location for a short-circuit fault ($R_f = 0\Omega$) using bus n and node m whereas $x|_{B_n,N_m}$ shows the variation of $X|_{B_n,N_m}$ with a rolling mean filter having a moving window of 20 sample steps [16]. Throughout the analysis, $x|_{B_n,N_m}$ is used to represent the filtered value of calculated location, $X|_{B_n,N_m}$. Similar analysis can be carried out for faults at SL_2 and SL_3 respectively.

2) Fault at IL_1: If the fault is located at IL_1, given as F_3 in Fig. 2, equations (8)-(9) are obtained considering bus 3 and bus 1.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 08:30:53 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Power Delivery. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRD.2023.3318689

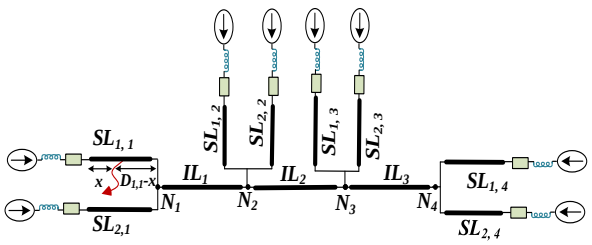


Fig. 5: *4 node, 8 bus* MVDC system

$$2r_3D_3i_3(t)+2l_3D_3\frac{di_3}{dt}+2dr_{M1}i_3(t)+2l_{M1}v_{L3}\frac{di_3}{dt}+R_f\sum i(t)=u_3(t) \tag{8}$$

$$2r_1D_1i_1(t)+l_1D_1\frac{di_1}{dt}+2(M_1-x)r_{M1}\Big(i_1(t)+i_2(t)\Big)+(M_1-x)l_{M1}\left[\frac{di_1}{dt}+\frac{di_2}{dt}\right]+R_f\sum i(t)=u_1(t) \tag{9}$$

Fault distance from bus 3 considering bus 1 is given as equation (10).

$$x|_{B_1,N_1}=\frac{\big[u_3(t)-u_1(t)\big]+2r_1D_1i_1(t)+2\Big[r_3D_3+M_1r_{M1}\Big]\Big(i_1(t)+i_2(t)\Big)}{2r_{M1}\sum i(t)+\frac{l_{M1}\sum v_L(t)}{L}}+\frac{\frac{l_1D_1v_{L1}}{L}+\frac{\Big[M_1l_{M1}+D_3l_3\Big]\Big(v_{L1}(t)+v_{L2}(t)\Big)}{L}}{2r_{M1}\sum i(t)+\frac{l_{M1}\sum v_L(t)}{L}}-\frac{2r_3D_3\sum i(t)+\frac{l_3D_3\sum v_L(t)}{L}}{2r_{M1}\sum i(t)+\frac{l_{M1}\sum v_L(t)}{L}} \tag{10}$$

Similar analysis can be carried out using KVL considering bus 2 and bus 3 to obtain $x|_{B_2,N_1}$. This gives the equivalent calculated fault location, $x|_{IL_1}=\frac{x|_{B_1,N_1}+x|_{B_2,N_1}}{2}$. Here $x|_{B_1,N_1}-x|_{B_2,N_1}\approx 0$ and $\boldsymbol{x}\leq M_1$: $\boldsymbol{x}\in\big\{x|_{B_1,N_1},x|_{B_2,N_1}\big\}$.

B. Fault location for a 4 node, 8 bus system

If the fault is located at $SL_{1,1}$, the single-line 4 node, 8 bus system is given as shown in Fig. 5. The unit resistance and inductance related to $SL_{j,i}$ is given as $r_{j,i}$ and $l_{j,i}$. $D_{j,i}$ is the total length of source/load line, $SL_{j,i}$. Using KVL for a PTP fault at $SL_{1,1}$, equations (11)-(15) are obtained considering bus 1 of node 1, bus 2 of node 1, bus 1 of node 2 and bus 1 of node 3.

$$2r_{1,1}xi_{1,1}(t)+2l_{1,1}x\frac{di_{1,1}(t)}{dt}+R_f\sum i(t)=u_{1,1}(t) \tag{11}$$

$$2D_{2,1}r_{2,1}i_{2,1}(t)+2l_{2,1}D_{2,1}\frac{di_{2,1}(t)}{dt}+2(D_{1,1}-x)r_{1,1}\Big[\sum i(t)-i_{1,1}(t)\Big]+2l_{1,1}(D_{1,1}-x)\Big(\frac{d\sum i(t)}{dt}-\frac{di_{1,1}(t)}{dt}\Big)+R_f\sum i(t)=u_{2,1}(t) \tag{12}$$

$$2D_{1,2}r_{1,2}i_{1,2}(t)+2M_1r_{M1}\Big[\sum i(t)-i_{1,1}(t)-i_{2,1}(t)\Big]+2l_{1,2}D_{1,2}\frac{di_{1,2}(t)}{dt}+2M_1l_{M1}\Big(\frac{d\big[\sum i(t)-i_{1,1}(t)-i_{2,1}(t)\big]}{dt}\Big)+2(D_{1,1}-x)r_{1,1}\Big[\sum i(t)-i_{1,1}(t)\Big]+2l_{1,1}(D_{1,1}-x)\Big(\frac{d\big[\sum i(t)-i_{1,1}(t)\big]}{dt}\Big)+R_f\sum i(t)=u_{1,2}(t) \tag{13}$$

$$2D_{1,3}r_{1,3}i_{1,3}(t)+2M_1r_{M1}\Big[\sum i(t)-i_{1,1}(t)-i_{2,1}(t)\Big]+2M_2r_{M2}\Big[i_{1,3}(t)+i_{2,1}(t)\Big]+2l_{1,3}D_{1,3}\frac{di_{1,3}(t)}{dt}+2M_1l_{M1}\Big(\frac{d\big[\sum i(t)-i_{1,1}(t)-i_{2,1}(t)\big]}{dt}\Big)+2M_2l_{M2}\Big(\frac{d\big[i_{1,3}(t)+i_{2,3}(t)\big]}{dt}\Big)+2(D_{1,1}-x)r_{1,1}\Big[\sum i(t)-i_{1,1}(t)\Big]+2l_{1,1}(D_{1,1}-x)\Big(\frac{d\big[\sum i(t)-i_{1,1}(t)\big]}{dt}\Big)+R_f\sum i(t)=u_{1,3}(t) \tag{14}$$

Rearranging (11) and (12) define the fault distance from bus 1 of node 1 considering bus 2 of node 1, given as equation (15).

$$x|_{B_2,N_1}=\frac{\big[u_{1,1}(t)-u_{2,1}(t)\big]+2r_{1,1}D_{1,1}\Big(\sum i(t)-i_{1,1}(t)\Big)+2r_{2,1}D_{2,1}i_{2,1}(t)}{2r_{1,1}\sum i(t)+\frac{l_{1,1}\sum v_L(t)}{L}}+\frac{\frac{l_{2,1}D_{2,1}v_{L2,1}(t)}{L}+\frac{l_{1,1}D_{1,1}(\sum v_L(t)-v_{L1,1}(t))}{L}}{2r_{1,1}\sum i(t)+\frac{l_{1,1}\sum v_L(t)}{L}} \tag{15}$$

Similarly, rearranging (11) and (13) defines the fault distance from bus 1 of node 1 considering bus 1 of node 2, given as equation (16).

$$x|_{B_1,N_2}=\frac{\big[u_{1,1}(t)-u_{1,2}(t)\big]+2r_{1,1}D_{1,1}\Big(\sum i(t)-i_{1,1}(t)\Big)+2r_{1,2}D_{1,2}i_{1,2}(t)}{2r_{1,1}\sum i(t)+\frac{l_{1,1}\sum v_L(t)}{L}}+\frac{\frac{l_{1,2}D_{1,2}v_{L1,2}(t)}{L}+\frac{l_{1,1}D_{1,1}(\sum v_L(t)-v_{L1,1}(t))}{L}}{2r_{1,1}\sum i(t)+\frac{l_{1,1}\sum v_L(t)}{L}} \tag{16}$$

On similar lines, rearranging equations (11) and (14) define the fault distance from bus 1 of node 1 considering bus 1 of node 3, given as equation (17). Further, similar analysis can be performed for faults at different $SL_{j,i}$ and IL_{i-1} respectively.

C. Fault location for a generalized k node, p bus system

Fig. 6(a) shows a generalized *p* bus, *k* node radial MVDC system. Fault at source/load line is shown to occur at node *i* *(N_i)*, bus *j* (given as $SL_{j,i}$, shown as F_1 in Fig. 6(b)). Fault at interconnecting line occurs between node *i-1* and *i* is shown as F_2 in Fig. 6(c). The algorithm can be applied based on a system's availability of communication and sensors with multiple terminal measurements available, and the location reliability increases, as highlighted earlier.
Furthermore, given the inductive nature of the fault location estimation described in subsections IV-A and IV-B, we can formulate the following theorem for the general location of the $SL_{j,i}$.

Theorem 1. *For the generalized MVDC power system depicted in Fig. 6, fault at the $SL_{j,i}$, for $j\in\{1,\ldots,N_p\}$ and $i\in\{1,\ldots,k\}$ can always be described with the equation* (18a)*.*

The proof for this theorem is in Appendix A. The results of Theorem 1 can be further elaborated to cover fault location at IL_{i-1}.

Theorem 2. *For the generalized MVDC power system depicted in Fig. 6, the fault placed at any interconnecting line between nodes $i-1$ and i, for $i\in\{1,\ldots,k\}$, can be determined using formula* (18b)*.*

The proof for this theorem is given in Appendix B.

V. Proposed Fault Location Scheme

To start with, the terminal voltage data, along with the voltage and current data across different CLRs, are measured. Further, a rolling mean filter (window of 50 samples) is used to filter out the high-frequency WGN in the measurement. Further, the DCR is used to identify if there is a fault in the system. If there is a violation of I_{set}, specific faulty segment is identified (between $SL_{j,i}$ and IL_i). For a fault at $SL_{j,i}$, equation (18a) is used for calculating the fault distance, whereas for

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 08:30:53 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Power Delivery. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRD.2023.3318689

$$x|_{B_1,N_3} = \frac{[u_{1,1}(t) - u_{1,3}(t)]}{2r_{1,1}\sum i(t) + \frac{l_{1,1}\sum v_L(t)}{L}} + \frac{\frac{l_{1,3}D_{1,3}v_{L1,3}(t)}{L} + \frac{l_{1,1}D_{1,1}(\sum v_L(t) - v_{L1,1}(t))}{L} + \frac{M_1 l_{M1}(\sum v_L(t) - v_{L1,1}(t) - v_{L2,1}(t))}{L} + \frac{M_2 l_{M2}(v_{L1,3}(t) - v_{L2,3}(t))}{L}}{2r_{1,1}\sum i(t) + \frac{l_{1,1}\sum v_L(t)}{L}} + \frac{2r_{1,1}D_{1,1}\left(\sum i(t) - i_{1,1}(t)\right) + 2r_{1,3}D_{1,3}i_{1,3}(t) + 2M_1 r_{M1}\left(\sum i(t) - i_{1,1}(t) - i_{2,1}(t)\right) + 2M_2 r_{M2}[i_{1,3}(t) + i_{2,3}(t)]}{2r_{1,1}\sum i(t) + \frac{l_{1,1}\sum v_L(t)}{L}} \quad (17)$$

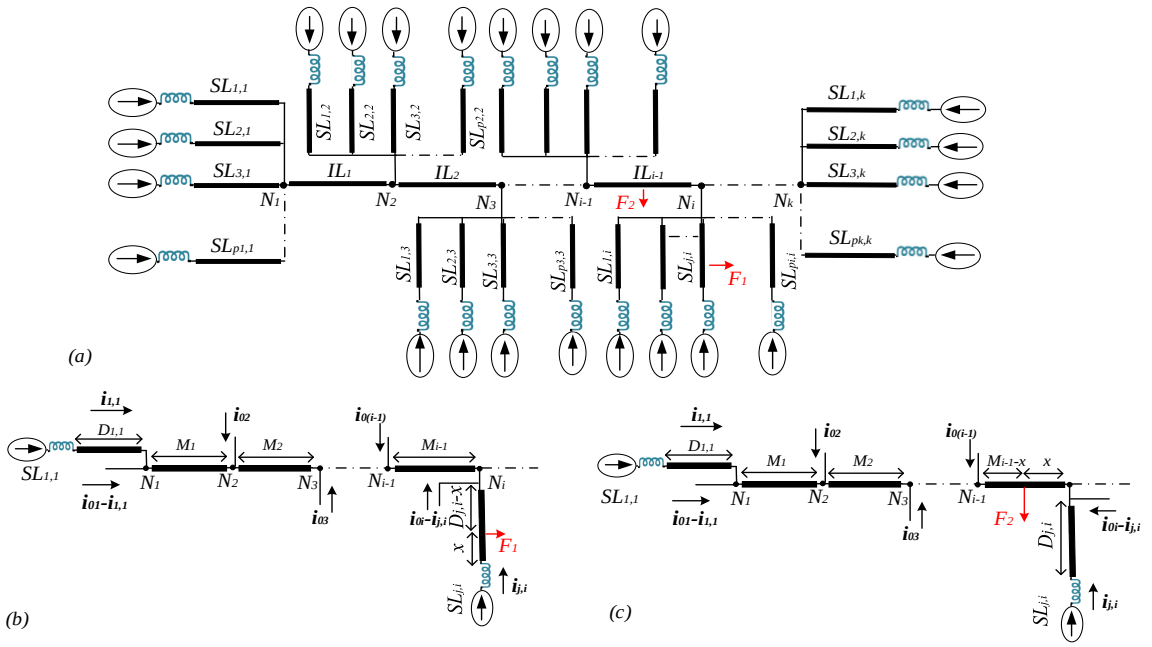


Fig. 6: (a) Generalised *k node, p bus* MVDC system, (b) effective circuit for fault, F_1, and (c) effective circuit for fault, F_2

$$x|_{B_n,N_m} = \frac{\left[u_{j,i}(t) - u_{n,m}(t)\right] + 2r_{n,m}D_{n,m}i_{n,m}(t) + 2\sum_{\beta=m}^{i-1} M_\beta r_{M_\beta} \sum_{\varepsilon=m}^{\beta} i_{0\varepsilon} + 2D_{j,i}r_{j,i} \sum_{\substack{(\alpha,\beta)=(1,1)\\(\alpha,\beta)\neq(j,i)}}^{p_\beta,k} i_{\alpha,\beta}(t) + l_{n,m}D_{n,m}\frac{v_{L_{n,m}}(t)}{L_{n,m}} + \sum_{\beta=m}^{i-1} l_{M_\beta} M_\beta \sum_{(\alpha,\varepsilon)=(1,m)}^{p_i-1,\beta} \frac{v_{L_{\alpha,\varepsilon}}(t)}{L_{\alpha,\varepsilon}} + D_{j,i}l_{j,i} \sum_{\substack{(\alpha,\beta)=(1,1)\\(\alpha,\beta)\neq(j,i)}}^{p_\beta,k} \frac{v_{L_{\alpha,\beta}}(t)}{L_{\alpha,\beta}}}{2r_{j,i}\sum i(t) + l_{j,i}\sum \frac{v_{L_{j,i}}}{L_{j,i}}} \quad (18a)$$

$$x|_{B_n,N_m} = \frac{\left[u_{j,i}(t) - u_{n,m}(t)\right] + 2r_{n,m}D_{n,m}i_{n,m}(t) + 2\sum_{\beta=m}^{i-2} M_\beta r_{M_\beta} \sum_{\varepsilon=m}^{\beta} i_{0\varepsilon} + 2M_{i-1}r_{M_{i-1}} \sum_{\varepsilon=m}^{i-1} i_{0\varepsilon}(t) + l_{n,m}D_{n,m}\frac{v_{L_{n,m}}(t)}{L_{n,m}} + \sum_{\beta=m}^{i-2} l_{M_\beta} M_\beta \sum_{(\alpha,\varepsilon)=(1,m)}^{p_i-2,\beta} \frac{v_{L_{\alpha,\varepsilon}}(t)}{L_{\alpha,\varepsilon}} + M_{i-1}l_{M_{i-1}} \sum_{(\alpha,\beta)=(1,1)}^{p_\beta,i-1} \frac{v_{L_{\alpha,\beta}}(t)}{L_{\alpha,\beta}}}{2r_{M_{i-1}}\sum i(t) + l_{M_{i-1}}\sum \frac{v_{L_{j,i}}}{L_{j,i}}} - \frac{2r_{j,i}D_{j,i}i_{j,i}(t) + l_{j,i}D_{j,i}\frac{v_{L_{j,i}}(t)}{L_{j,i}}}{2r_{M_{i-1}}\sum i(t) + l_{M_{i-1}}\sum \frac{v_{L_{j,i}}}{L_{j,i}}} \quad (18b)$$

a fault at IL_i, equation (18b) is used to calculate the fault distance. Fault location is calculated considering each bus and node (excluding the case where *(p,k)=(j,i)* i.e., the location from bus j considering bus j, node i). This means that the method entirely focuses on multi-terminal measurements for fault location. Once the fault location is calculated, conditions (22) are verified. If equation (22) is satisfied, the final fault location is evaluated using equation (21). Fig. 7 shows the layout of the proposed fault location scheme.

VI. Results and Validation

The test systems shown in Fig. 1 and Fig. 2 are implemented using PSCAD/EMTDC based electromagnetic transient simulations. A sampling frequency, f_s of 50 kHz is used for conclusive plots. However, section V(D) shows that a sampling frequency of 5 kHz is enough for satisfactory accuracy. The detailed models of OHL and UGC are considered in PSCAD/EMTDC. The parameters for the test system are given in Table II. The validation starts with fault identification and eventually the location of the fault in a three-bus system shown in Fig. 2. Once the process of fault location is validated

TABLE II: Parameters for a test system in PSCAD/EMTDC

DC Voltage *(kV)*	±2
DC link capacitance (2C) *(mF)*	15.6
Grounding resistance (R_g) (Ω)	0.2-0.5
Smoothing Reactor *(mH)*	1
Parameters of OHL	Resistance [Ω/km]: 0.132 Inductance [mH/km]: 0.201 Capacitance [μF/km]: 0.011
Parameters of UGC	Resistance [Ω/km]: 0.098 Inductance [mH/km]: 0.074 Capacitance [μF/km]: 0.272

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 08:30:53 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Power Delivery. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRD.2023.3318689

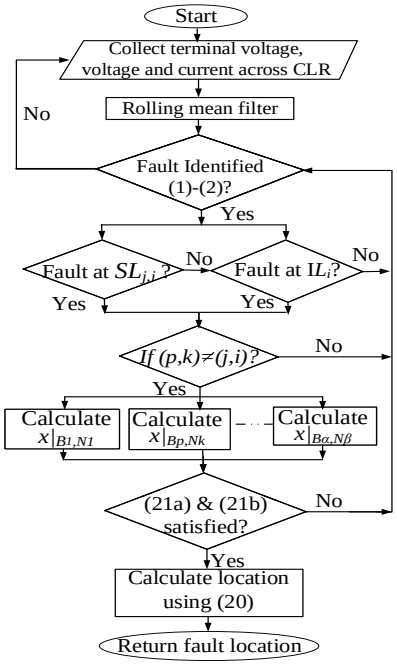


Fig. 7: Layout of the scheme

for a smaller system, it is then tested for a six-bus system shown in Fig. 1. The algorithm is further tested under different fault resistances and fault locations with different sampling frequencies and white Gaussian noise (WGN) in measurement. Finally, the effect of parameter variation on the proposed algorithm is discussed.

A. Validation for fault location for a three bus system

A three-bus system with SL_1-SL_3 and IL_1 is used (see Fig. 2) to check the performance of the fault identification and location algorithm. Different types of faults (PTP, P-PTG and N-PTG) occur at the source/load line and interconnecting line.

1) Source/Load line faults:

a) PTP fault: A PTP fault occurs at d=0.5km at SL_1 with fault resistance, R_f=1Ω. Both positive and negative pole differential current ($i_{B1,1^+}(t)$ and $i_{B1,1^-}(t)$) violate I_{set} indicating a PTP fault at SL_1. The differential current of other terminals does not violate I_{set}. Once the fault is identified in lesser than 1ms, the fault location algorithm calculates $X|_{B_2,N_1}$ and $X|_{B_3,N_1}$, $x|_{B_2,N_1}$ and $x|_{B_3,N_1}$ (rolling mean of $X|_{B_2,N_1}$ and $X|_{B_3,N_1}$). Over 1-2ms of fault inception, Fig. 8(a) shows $x|_{B_3,N_1}$≈0.496km, $x|_{B_2,N_1}$≈0.508km and $x|_{B_2,N_1}$-$x|_{B_3,N_1}$≈0.012km.

b) P-PTG faults: A P-PTG fault occurs at d=0.25km at SL_1 with fault resistance, R_f=1Ω and grounding resistance, R_g=0.5Ω. Only positive pole differential current ($i_{B1,1^+}(t)$) violates I_{set} indicating a P-PTG fault at SL_1. Fig. 8(b) shows $x|_{B_3,N_1}$≈0.251km, $x|_{B_2,N_1}$≈0.254km and $x|_{B_2,N_1}$-$x|_{B_3,N_1}$≈0.003km.

c) N-PTG fault: A N-PTG fault occurs at d=0.75km at SL_1 with fault resistance, R_f=1Ω and grounding resistance, R_g=0.5Ω. The differential negative pole current ($i_{B1,1^-}(t)$) violates I_{set} indicating a N-PTG. Once the fault is identified, the location is calculated as $x|_{B_3,N_1}$≈0.755km, $x|_{B_2,N_1}$≈0.763km and $x|_{B_2,N_1}$-$x|_{B_3,N_1}$≈0.008km.

2) Interconnecting line faults:

a) PTP fault: A PTP fault occurs at d=5km at IL_1. Both positive and negative pole differential current ($i_{B1,1^+}(t)$ and $i_{B1,1^-}(t)$) violate I_{set} indicating a PTP fault at IL_1. Fig. 9(a) shows $x|_{B_3,N_1}$≈4.995km, $x|_{B_2,N_1}$≈4.99km and $x|_{B_2,N_1}$-$x|_{B_3,N_1}$≈-0.005km.

b) P-PTG faults: A P-PTG fault occurs at d=3km at IL_1, with grounding resistance, R_g=0.5Ω. Only positive pole differential current ($i_{B1,1^+}(t)$) violates I_{set} indicating a P-PTG fault at IL_1. Fig. 9(b) shows $x|_{B_3,N_1}$≈3.005km, $x|_{B_2,N_1}$≈2.995km and $x|_{B_2,N_1}$-$x|_{B_3,N_1}$≈-0.01km.

c) N-PTG fault: A N-PTG fault occurs at d=6km at IL_1 with fault resistance, R_f=5Ω and grounding resistance, R_g=0.5Ω. Only the differential negative pole current ($i_{B1,1^-}(t)$) violates I_{set} indicating a N-PTG at IL_1. The location is calculated as $x|_{B_3,N_1}$≈6.002km, $x|_{B_2,N_1}$≈5.99km and $x|_{B_2,N_1}$-$x|_{B_3,N_1}$≈-0.012km.

B. Validation for fault location in a six bus system

For the system shown in Fig. 1, the inception of different faults at $SL_{1,1}$ and IL_1 occur to validate the fault location algorithm for a six-bus test system.

1) Source/Load line faults: Fig. 8 shows different faults at $SL_{1,1}$ (total length 2km). Fig. 10(a) shows the calculated fault location for a fault at d=0.05km. The fault location calculated from bus 1 considering different buses and nodes are $x|_{B_2,N_1}$≈0.051km, $x|_{B_3,N_1}$≈0.053km, $x|_{B_4,N_1}$≈0.055km, $x|_{B_5,N_1}$≈0.053km, and $x|_{B_6,N_1}$≈0.052km. Using equation (21), $x|_{SL_{1,1}}$=0.0528km. Fig. 10(b) shows the calculated fault location for a fault at d=1km. The location with different buses and nodes are $x|_{B_2,N_1}$≈1.001km, $x|_{B_3,N_1}$≈1.002km, $x|_{B_4,N_1}$≈1.0025km, $x|_{B_5,N_1}$≈1.002km, $x|_{B_6,N_1}$≈1.002km. Using equation (21), $x|_{SL_{1,1}}$=1.0019km. Fig. 10(c) shows the calculated fault location for a fault at d=1.95km. The location calculated with different buses and nodes are $x|_{B_2,N_1}$≈1.946km, $x|_{B_3,N_1}$≈1.9475km, $x|_{B_4,N_1}$≈1.95km, $x|_{B_5,N_1}$≈1.9475km, and $x|_{B_6,N_1}$≈1.947km. Using equation (21), $x|_{SL_{1,1}}$=1.9476km.

2) Interconnecting line faults: Fig. 11 shows different faults at IL_1 (total length 10km). Fig. 11(a) shows the calculated fault location for a fault at d=2km. Similar to (15), $x|_{IL_1}$=2.006km. Fig. 11(b) shows location for a fault at d=4km as $x|_{IL_1}$=4.008km and Fig. 11(c) shows location for a fault at d=6km as $x|_{IL_1}$=6.003km. Similar to faults at $SL_{1,1}$, the calculated location is accurate.

C. Validation for different fault resistances

The variation of calculated location with different fault resistances (up to R_f=200Ω) is explored in Table III. Since the method is multi-terminal, the algorithm's accuracy is independent of fault resistance. Table III shows faults at $SL_{1,2}$, and IL_2. Considering Fig. 1 as a test system, there are 5 buses which can be individually used to calculate the fault location for $SL_{1,2}$ (i.e., $x|_{B_4,N_2}$, $x|_{B_5,N_2}$, $x|_{B_6,N_2}$, $x|_{B_1,N_2}$ and $x|_{B_2,N_2}$). Similarly, for a fault at IL_1, $x|_{B_1,N_2}$, $x|_{B_2,N_2}$, $x|_{B_3,N_2}$, $x|_{B_4,N_2}$, $x|_{B_5,N_2}$, and $x|_{B_5,N_2}$ can be used. However, due to space

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 08:30:53 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Power Delivery. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRD.2023.3318689

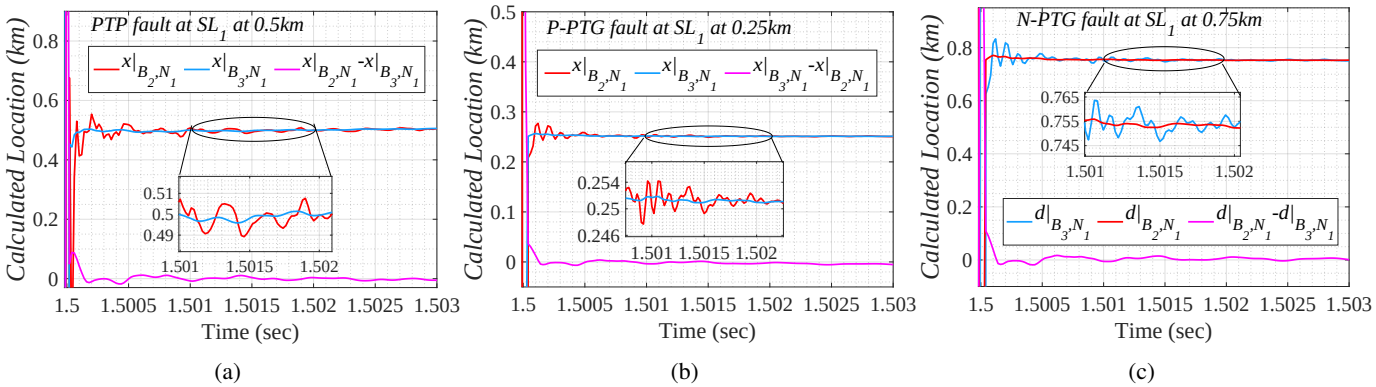


Fig. 8: Three-bus system: Calculated fault location for different faults at SL_1 at (a) d=0.5km; (b) d=0.25km; (c) d=0.75km.

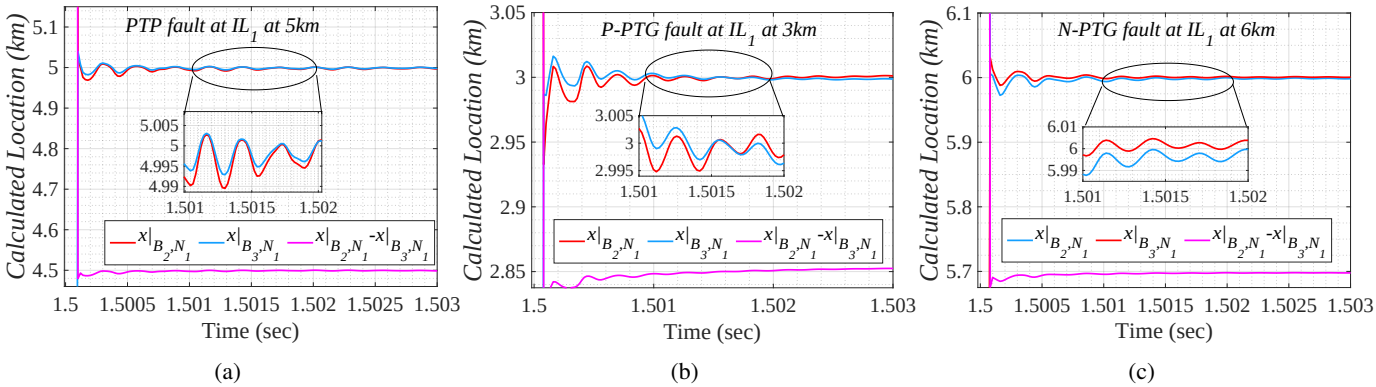


Fig. 9: Three-bus system: Differential current and calculated fault location for different faults at IL_1 at (a) d=5km; (b) d=3km; (c) d=6km.

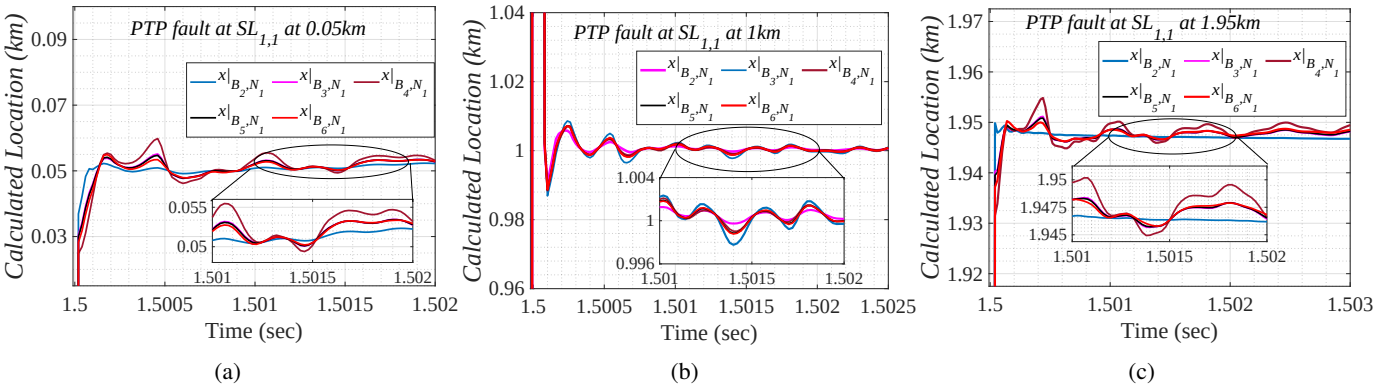


Fig. 10: Six-bus system: calculated fault location for faults at $SL_{1,1}$ at (a) d=0.25km, (b) d=0.5km, (c) d=1km, (d) d=1.25km, (e) d=1.5km, (f) d=1.75km.

constraints, Table III shows the fault location using 3 buses for each case. For high sampling frequencies (f_s>20kHz), there is no loss of accuracy for high fault resistances. For low sampling frequencies (f_s<20kHz), there is a marginal loss of accuracy (<1%) for high fault resistances. This is because as fault resistance increases, the signal damping increases. If the sampling frequency is less, useful samples are passed over.

D. Performance under different sampling frequencies

Fault location accuracy with low sampling frequency can help in cost-effective hardware implementation. Table IV shows that a 5kHz sampling frequency gives satisfactory results for the proposed fault location method. However, as evident from Table IV, higher f_s means more useful samples, which increase the accuracy of the fault location algorithm.

E. Fault location calculation with parameter variation

The sensitivity of the fault location algorithm with the variation of system parameters, i.e., r (unit resistance) and l (unit inductance) of cables and C (DC link capacitance) of converters is explored. The initial parameters are given in Table II. It is observed that using different values of r and C up to $\pm 200\%$ does not change the accuracy of the fault location algorithm. However, the accuracy of the algorithm varies with variations in unit inductance. Table V shows $\varepsilon(\%)=\varepsilon/x$ with $\%$ change in inductance value. For faults closer to the bus terminal, $\%$ drop in fault location accuracy is higher. As a

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 08:30:53 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Power Delivery. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRD.2023.3318689

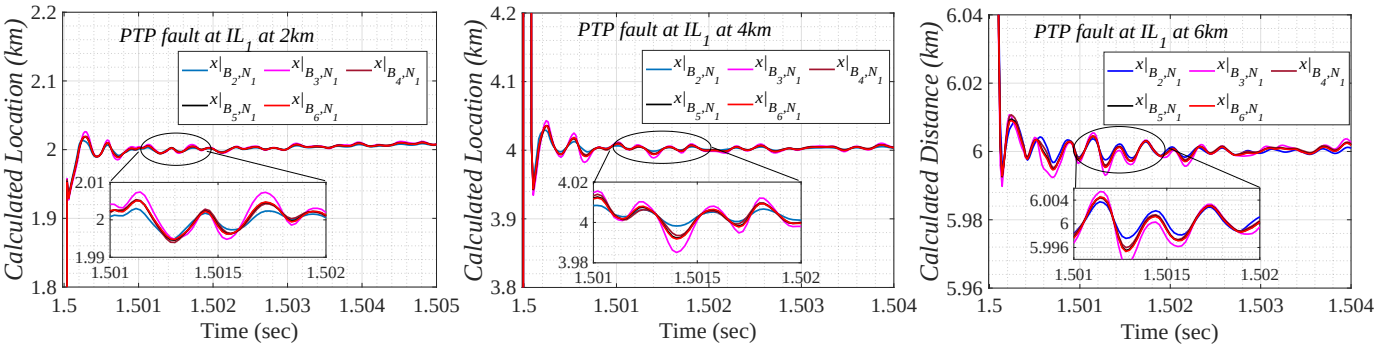


Fig. 11: Six-bus system: calculated fault location for faults at IL_1 at (a) d=2km, (b) d=4km, (c) d=6km.

TABLE III: Variation of calculated fault location (km) with fault resistance, R_f (Ω) for different faults (PTG, PTP) at different locations

	$SL_{1,2}$ (Total Length: 1km)									IL_1 (Total Length: 10km)								
	R_f=0.01Ω			R_f=50Ω			R_f=200Ω			R_f=0.01Ω			R_f=50Ω			R_f=200Ω		
	$x\|_{B_4,N_2}$	$x\|_{B_5,N_2}$	$x\|_{B_6,N_2}$	$x\|_{B_4,N_2}$	$x\|_{B_5,N_2}$	$x\|_{B_6,N_2}$	$x\|_{B_4,N_2}$	$x\|_{B_5,N_2}$	$x\|_{B_6,N_2}$	$x\|_{B_4,N_2}$	$x\|_{B_5,N_2}$	$x\|_{B_6,N_2}$	$x\|_{B_4,N_2}$	$x\|_{B_5,N_2}$	$x\|_{B_6,N_2}$	$x\|_{B_4,N_2}$	$x\|_{B_5,N_2}$	$x\|_{B_6,N_2}$
P-PTG																		
5%	0.052	0.053	0.05	0.052	0.052	0.052	0.054	0.055	0.052	0.505	0.502	0.502	0.510	0.505	0.505	0.515	0.51	0.512
30%	0.312	0.318	0.297	0.295	0.318	0.315	0.29	0.322	0.32	3.05	3.08	3.02	3.05	3.1	2.97	3.12	3.05	2.92
45%	0.45	0.452	0.462	0.455	0.448	0.462	0.455	0.446	0.46	4.505	4.51	4.492	4.51	4.51	4.495	4.51	4.51	4.485
60%	0.6	0.622	0.615	0.605	0.62	0.62	0.61	0.625	0.625	6.1	6.1	5.85	6.15	6.22	6.15	6.15	6.1	6.25
75%	0.755	0.76	0.748	0.762	0.762	0.755	0.765	0.77	0.762	7.51	7.5	7.525	7.51	7.51	7.52	7.52	7.51	7.62
95%	0.95	0.955	0.945	0.958	0.96	0.965	0.938	0.955	0.97	9.5	9.52	9.51	9.54	9.58	9.62	9.45	9.4	9.655
N-PTG																		
5%	0.052	0.053	0.05	0.052	0.052	0.052	0.054	0.055	0.052	0.505	0.502	0.502	0.510	0.505	0.505	0.515	0.51	0.512
30%	0.312	0.318	0.297	0.295	0.318	0.315	0.29	0.322	0.32	3.05	3.08	3.02	3.05	3.1	2.97	3.12	3.05	2.92
45%	0.45	0.452	0.462	0.455	0.448	0.462	0.455	0.446	0.46	4.505	4.51	4.492	4.51	4.51	4.495	4.51	4.51	4.485
60%	0.6	0.622	0.615	0.605	0.62	0.62	0.61	0.625	0.625	6.1	6.1	5.85	6.15	6.22	6.15	6.15	6.1	6.25
75%	0.755	0.76	0.748	0.762	0.762	0.755	0.765	0.77	0.762	7.51	7.5	7.525	7.51	7.51	7.52	7.52	7.51	7.62
95%	0.95	0.955	0.945	0.958	0.96	0.965	0.938	0.955	0.97	9.5	9.52	9.51	9.54	9.58	9.62	9.45	9.4	9.655
PTP																		
5%	0.051	0.0525	0.05	0.052	0.0535	0.052	0.054	0.0545	0.052	0.505	0.502	0.502	0.510	0.51	0.505	0.515	0.51	0.512
30%	0.32	0.32	0.297	0.29	0.318	0.315	0.33	0.322	0.32	3.06	3.12	3.12	3.05	3.1	2.97	3.12	3.05	2.92
45%	0.455	0.45	0.462	0.455	0.448	0.462	0.455	0.446	0.46	4.505	4.51	4.492	4.51	4.51	4.495	4.51	4.51	4.485
60%	0.62	0.622	0.615	0.605	0.62	0.62	0.61	0.625	0.625	6.05	5.95	5.85	6.15	6.22	6.15	6.15	6.1	6.25
75%	0.762	0.76	0.748	0.762	0.762	0.755	0.765	0.77	0.762	7.51	7.5	7.525	7.51	7.51	7.52	7.52	7.51	7.62
95%	0.962	0.955	0.945	0.958	0.96	0.965	0.938	0.955	0.97	9.55	9.52	9.51	9.54	9.58	9.62	9.45	9.4	9.655

TABLE IV: Location with different sampling frequencies for a PTP fault at $SL_{1,2}$

Error (m)=Fault Location-Calculated Fault Location				
Fault Location (m)	50kHz	20kHz	10kHz	5kHz
100	1.25	3.35	6.56	10.05
300	1.18	3.32	6.63	11.54
500	0.75	3.22	6.58	10.45
700	1.15	3.27	5.77	9.82
900	1.12	3.32	5.92	9.68

TABLE V: Variation of calculated fault location error (%) with variation in unit inductance (%) for a PTP fault at $SL_{1,1}$

Calculated Fault Location error (%)=ε/x						
	Fault Distance, x (m)					
Variation in l	100m	500m	900m	1300m	1700m	1900m
$\pm$5%	4.2%	1.24%	0.67%	0.45%	0.35%	0.31%
$\pm$10%	10.1%	2.35%	1.36%	0.88%	0.71%	0.62%
$\pm$15%	15.4%	3.94%	2.05%	1.32%	1.02%	0.97%
$\pm$20%	18.1%	4.84%	2.65%	1.81%	1.34%	1.28%

result, a $\pm$20% change in inductance reduces accuracy as high as $\pm$18%. This suggests that the fault location method is sensitive to the unit inductance of the line or cable. In contrast, unit resistance and DC link capacitance value do not affect it.

The future scope of work includes analyzing the variation of fault location accuracy concerning change in unit inductance in detail and proposing a location method with much lesser sensitivity due to its interpretation.

F. Fault detection and location time

For the application of the proposed location method, the fault identification scheme takes around 1 ms, which complements the location method. For f_s=20kHz, maximum identification time is recorded as 0.95ms. The fault location window is 1 to 2ms from the inception time. This makes the method's maximum detection and location time to be less than 2ms.

G. Performance with white Gaussian noise in measurement

White Gaussian Noise (WGN) in the measurement induces high-frequency components in the measured current and voltage data. This can cause inaccuracy in the results for the fault location of the proposed scheme. As a result, a rolling mean filter [16] is used with a moving window of 50 sample steps for voltage and current samples. Additionally, another rolling mean is used with a window of 20 samples to smoothen the fault location curve (see Fig. 4). WGN is random with the property of zero mean [16]. Using rolling mean samples effectively eliminates the effect of WGN to a great extent

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 08:30:53 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Power Delivery. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRD.2023.3318689

without affecting the accuracy of the proposed fault location algorithm. Table VI performs the algorithm with WGN in measurement for a PTP fault at SL_1. A first-order lag can also be used in place of a rolling mean if the delay due to the lag filter does not hinder the stringent requirement of fault detection and location time.

TABLE VI: Fault location in the presence of WGN

	Error (m)			
Fault Location (m)	10dB	20dB	30dB	40dB
100	7.25	4.42	4.76	2.55
300	6.18	4.35	3.85	2.62
500	6.75	5.02	4.52	2.55
700	7.15	4.34	3.82	2.71
900	7.12	5.3	2.92	2.35

VII. Conclusion

The work proposes a novel fault location algorithm for a multi-terminal radial MVDC microgrid. Based on the findings of the proposed work, it is concluded that:

- The proposed algorithm locates the fault distance in a system ranging from a simple homogeneous line configuration to n-segment non-homogeneous lines for distribution i.e., different line/cable combinations.
- Any of the bus terminals along with the faulty bus terminal can be used to accurately locate the fault. If other bus terminals are also utilized (subject to the availability of communication and sensors) for fault location, the reliability of the method increases.
- The proposed analysis of fault location is extrapolated and generalized for a comparatively larger MVDC system with k node and p bus for wider applications to different systems.
- The proposed method is sensitive to the use of the true value of the unit inductance of a line or cable. A 20% deviation in the true value of unit inductance can cause location error as high as 18%. The error is relatively higher for faults closer to the bus terminal. The accuracy of the method is not affected by even $\pm 200\%$ change in the unit resistance of the line or cable and the value of DC link capacitance.
- The performance of the location method is robust for different sampling frequencies (f_s=5-100kHz), under different WGN in measurement (SNR=0-40dB), for different types of faults (PTP, P-PTG, and N-PTG) at different fault resistances (validated up to R_f=200Ω) and location.

Acknowledgement

This project has received funding from the European Union's HORIZON-WIDERA-2021-ACCESS-03 under grant agreement No. 101079200.

References

[1] V. Nougain, S. Mishra and S. S. Jena, "Resilient Protection of Medium Voltage DC Microgrids against Cyber Intrusion," in IEEE Transactions on Power Delivery, doi: 10.1109/TPWRD.2021.3074879.

[2] N. Doerry, J. Amy and C. Krolick, "History and the Status of Electric Ship Propulsion, Integrated Power Systems, and Future Trends in the U.S. Navy," in Proceedings of the IEEE, vol. 103, no. 12, pp. 2243-2251, Dec. 2015, doi: 10.1109/JPROC.2015.2494159.

[3] J. D. Herbst, A. L. Gattozzi, A. Ouroua and F. M. Uriarte, "Flexible test bed for MVDC and HFAC electric ship power system architectures for Navy ships," 2011 IEEE Electric Ship Technologies Symposium, Alexandria, VA, 2011, pp. 66-71, doi: 10.1109/ESTS.2011.5770843.

[4] Z. Jin, L. Meng, J. M. Guerrero and R. Han, "Hierarchical Control Design for a Shipboard Power System With DC Distribution and Energy Storage Aboard Future More-Electric Ships," in IEEE Transactions on Industrial Informatics, vol. 14, no. 2, pp. 703-714, Feb. 2018, doi: 10.1109/TII.2017.2772343.

[5] K. Satpathi, A. Ukil, S. S. Nag, J. Pou and M. A. Zagrodnik, "DC Marine Power System: Transient Behavior and Fault Management Aspects," in IEEE Transactions on Industrial Informatics, vol. 15, no. 4, pp. 1911-1925, April 2019, doi: 10.1109/TII.2018.2864598.

[6] C. Yuan, M. A. Haj-ahmed and M. S. Illindala, "Protection Strategies for Medium-Voltage Direct-Current Microgrid at a Remote Area Mine Site," in IEEE Transactions on Industry Applications, vol. 51, no. 4, pp. 2846-2853, July-Aug. 2015, doi: 10.1109/TIA.2015.2391441.

[7] V. Nougain, S. Mishra and A. K. Pradhan, "MVDC Microgrid Protection Using a Centralized Communication With a Localized Backup Scheme of Adaptive Parameters," in IEEE Transactions on Power Delivery, vol. 34, no. 3, pp. 869-878, June 2019, doi: 10.1109/TPWRD.2019.2899768.

[8] R. M. Cuzner and V. Singh, "Future Shipboard MVdc System Protection Requirements and Solid-State Protective Device Topological Tradeoffs," in IEEE Journal of Emerging and Selected Topics in Power Electronics, vol. 5, no. 1, pp. 244-259, March 2017, doi: 10.1109/JESTPE.2016.2638921.

[9] S. Jiang, C. Fan, N. Huang, Y. Zhu and M. He, "A Fault Location Method for DC Lines Connected With DAB Terminal in Power Electronic Transformer," in IEEE Transactions on Power Delivery, vol. 34, no. 1, pp. 301-311, Feb. 2019, doi: 10.1109/TPWRD.2018.2867851.

[10] W. Liu, F. Liu, X. Zha, M. Huang, C. Chen and Y. Zhuang, "An Improved SSCB Combining Fault Interruption and Fault Location Functions for DC Line Short-Circuit Fault Protection," in IEEE Transactions on Power Delivery, vol. 34, no. 3, pp. 858-868, June 2019, doi: 10.1109/TPWRD.2018.2882497.

[11] A. Makkieh, V. Psaras, R. Peña-Alzola, D. Tzelepis, A. A. S. Emhemed and G. M. Burt, "Fault Location in DC Microgrids Based on a Multiple Capacitive Earthing Scheme," in IEEE Journal of Emerging and Selected Topics in Power Electronics, vol. 9, no. 3, pp. 2550-2559, June 2021, doi: 10.1109/JESTPE.2020.2995946.

[12] D. Wang, V. Psaras, A. A. S. Emhemed and G. M. Burt, "A Novel Fault Let-Through Energy Based Fault Location for LVDC Distribution Networks," in IEEE Transactions on Power Delivery, vol. 36, no. 2, pp. 966-974, April 2021, doi: 10.1109/TPWRD.2020.2998409.

[13] Y. Yang, C. Huang and Q. Xu, "A Fault Location Method Suitable for Low-Voltage DC Line," in IEEE Transactions on Power Delivery, vol. 35, no. 1, pp. 194-204, Feb. 2020, doi: 10.1109/TPWRD.2019.2930622.

[14] J. Park, J. Candelaria, L. Ma and K. Dunn, "DC Ring-Bus Microgrid Fault Protection and Identification of Fault Location," in IEEE Transactions on Power Delivery, vol. 28, no. 4, pp. 2574-2584, Oct. 2013, doi: 10.1109/TPWRD.2013.2267750.

[15] R. Mohanty, U. S. M. Balaji and A. K. Pradhan, "An Accurate Noniterative Fault-Location Technique for Low-Voltage DC Microgrid," in IEEE Transactions on Power Delivery, vol. 31, no. 2, pp. 475-481, April 2016, doi: 10.1109/TPWRD.2015.2456934.

[16] V. Nougain, V. Nougain and S. Mishra, "Low-voltage DC ring-bus microgrid protection with rolling mean technique," 2018 IEEMA Engineer Infinite Conference (eTechNxT), New Delhi, 2018, pp. 1-6.

[17] J. Yang, J. E. Fletcher and J. O'Reilly, "Short-Circuit and Ground Fault Analyses and Location in VSC-Based DC Network Cables," in IEEE Transactions on Industrial Electronics, vol. 59, no. 10, pp. 3827-3837, Oct. 2012, doi: 10.1109/TIE.2011.2162712.

[18] C. Li, C. Zhao, J. Xu, Y. Ji, F. Zhang and T. An, "A Pole-to-Pole Short-Circuit Fault Current Calculation Method for DC Grids," in IEEE Transactions on Power Systems, vol. 32, no. 6, pp. 4943-4953, Nov. 2017.

[19] J. Xu, S. Zhu, C. Li and C. Zhao, "The Enhanced DC Fault Current Calculation Method of MMC-HVDC Grid With FCLs," in IEEE Journal of Emerging and Selected Topics in Power Electronics, vol. 7, no. 3, pp. 1758-1767, Sept. 2019.

[20] J. Xu, Y. Lü, C. Zhao and J. Liang, "A Model-Based DC Fault Location Scheme for Multi-Terminal MMC-HVDC Systems Using a Simplified

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 08:30:53 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Power Delivery. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRD.2023.3318689

Transmission Line Representation," in IEEE Transactions on Power Delivery, vol. 35, no. 1, pp. 386-395, Feb. 2020.
[21] O. M. K. K. Nanayakkara, A. D. Rajapakse and R. Wachal, "Location of DC Line Faults in Conventional HVDC Systems With Segments of Cables and Overhead Lines Using Terminal Measurements," in IEEE Transactions on Power Delivery, vol. 27, no. 1, pp. 279-288, Jan.
[22] P. T. Lewis, B. M. Grainger, H. A. Al Hassan, A. Barchowsky and G. F. Reed, "Fault Section Identification Protection Algorithm for Modular Multilevel Converter-Based High Voltage DC With a Hybrid Transmission Corridor," in IEEE Transactions on Industrial Electronics, vol. 63, no. 9, pp. 5652-5662, Sept. 2016.
[23] D. Tzelepis, G. Fusiek, A. Dyśko, P. Niewczas, C. Booth and X. Dong, "Novel Fault Location in MTDC Grids With Non-Homogeneous Transmission Lines Utilizing Distributed Current Sensing Technology," in IEEE Transactions on Smart Grid, vol. 9, no. 5, pp. 5432-5443, Sept. 2018, doi: 10.1109/TSG.2017.2764025.
[24] H. Panahi, M. Sanaye-Pasand, R. Zamani and H. Mehrjerdi, "A Novel DC Transmission System Fault Location Technique for Offshore Renewable Energy Harvesting," in IEEE Transactions on Power Delivery, vol. 35, no. 6, pp. 2885-2895, Dec. 2020, doi: 10.1109/TPWRD.2020.3010160.
[25] "ISO/IEC Guide 98-3:2008." International Organization for Standardization, 10 Sept. 2007, www.iso.org/standard/50461.html.

APPENDIX A
GENERALITY FOR FAULT LOCATION AT $\mathrm{SL}_{j,i}$

Proof. Each node, *n* in the generalized system is considered to have p_n buses contributing a total fault current of $i_{0n}(t)$. Here fault current from bus 1 at node *n* is indicated as $i_{1,n}(t)$. Similarly, current from bus 2 at node *n* is indicated as $i_{2,n}(t)$ and so on. The total fault current from node 1, $i_{01} = \sum_{\alpha=1}^{p_1} i_{\alpha,1}$; from node 2, $i_{02} = \sum_{\alpha=1}^{p_2} i_{\alpha,2}$ and so on. The expressions, $\sum i(t) = \sum_{\alpha=1}^{p_\beta} \sum_{\beta=1}^{k} i_{\alpha,\beta}(t)$, $\frac{di_{1,1}}{dt} = \frac{v_{L_{1,1}}}{L_{1,1}}$, $\frac{di_{01}}{dt} = \sum_{\alpha=1}^{p_1} \frac{v_{L_{\alpha,1}}}{L_{\alpha,1}}$ and their extensions are used interchangeably in the analysis of generalised *k* node, *p* bus system ahead. If the fault takes place at source/load line at node *i*, bus *j* ($SL_{j,i}$), then the fault location is evaluated using the effective faulty circuit as shown in Fig. 11(b). Here the location is calculated using buses related to $SL_{1,1}$ and $SL_{j,i}$. $r_{j,i}D_{j,i}$ and $l_{j,i}D_{j,i}$ give the equivalent resistance and inductance from bus *j* to node *i* whereas $r_{n,m}D_{n,m}$ and $l_{n,m}D_{n,m}$ give the equivalent resistance and inductance from bus *n* to node *m*. Using KVL for a PTP fault at bus *j* of node *i* from bus 1 of node 1, equations (19)-(20) are obtained.

$$\begin{aligned} 2r_{1,1}D_{1,1}i_{1,1}(t) + 2\sum_{\beta=1}^{i-1} M_\beta r_{M_\beta} \sum_{\varepsilon=1}^{\beta} i_{0\varepsilon} + 2(D_{j,i}-x)r_{j,i} \sum_{\substack{(\alpha,\beta)=(1,1)\\(\alpha,\beta)\neq(j,i)}}^{p_\beta,k} i_{\alpha,\beta}(t) + \\ R_f \sum i(t) + l_{1,1}D_{1,1}\frac{v_{L_{1,1}}(t)}{L_{1,1}} + \sum_{\beta=1}^{i-1} l_{M_\beta}M_\beta \sum_{(\alpha,\varepsilon)=(1,1)}^{p_{i-1},\beta} \frac{v_{L_{\alpha,\varepsilon}}(t)}{L_{\alpha,\varepsilon}} + \\ (D_{j,i}-x)l_{j,i} \sum_{\substack{(\alpha,\beta)=(1,1)\\(\alpha,\beta)\neq(j,i)}}^{p_\beta,k} \frac{v_{L_{\alpha,\beta}}(t)}{L_{\alpha,\beta}} = u_{1,1}(t) \end{aligned} \quad (19)$$

$$2r_{j,i}xi_{j,i}(t) + l_{j,i}x\frac{v_{L_{j,i}}}{L_{j,i}} + R_f \sum i(t) = u_{j,i}(t) \quad (20)$$

Similarly, as explained in previous subsections, rearranging equations (19)-(20) give the generalised fault distance expression from bus *m* of node *n* as given in equation (18a), which completes the proof. □

Further, the equivalent fault distance using all available terminal measurements can be defined as (21).

$$x|_{SL_{j,i}} = \frac{1}{pk} \sum_{\substack{\alpha=1\\(\alpha,\beta)\neq(j,i)}}^{p} \sum_{\beta=1}^{k} x|_{B_\alpha,N_\beta} \quad (21)$$

Reliability conditions are defined as (22a)-(22b). Here (22a) suggests that fault locations calculated using different terminal buses are close to each other in magnitude. (22b) suggests that each calculated fault location is lesser than the total line length, $D_{j,i}$.

$$\sum_{\substack{\alpha=2\\(\alpha,\beta)\neq(j,i)}}^{p} \sum_{\beta=2}^{k} x|_{B_\alpha,N_\beta} - \sum_{\substack{\alpha=1\\(\alpha,\beta)\neq(j,i)}}^{p} \sum_{\beta=1}^{k} x|_{B_\alpha,N_\beta} < D_{threh} \quad (22a)$$

$$\boldsymbol{x} \leq D_{j,i}\text{: } \boldsymbol{x} \in \left\{ \sum_{\substack{\alpha=1\\(\alpha,\beta)\neq(j,i)}}^{p} \sum_{\beta=1}^{k} x|_{B_\alpha,N_\beta} \right\} \quad (22b)$$

D_{thresh} is a static threshold for practical implementation of (22a). The lower limit of setting D_{thresh} is restricted by the time to locate the fault. For faults with lower fault resistance, the response damps slowly. As a result of which, it is possible that a very low threshold does not satisfy (22a) until the response is absolutely damped. This can delay the time to locate the fault. Additionally, a high value of D_{thresh} may compromise the accuracy of fault location. A nominal value of around $\pm 5\% - 8\%$ can be selected as the static threshold for D_{thresh}. Throughout the study, the calculated fault distance for a fault at $SL_{j,i}$ is noticed to be composed of (a) $\left[u_{j,i}(t) - u_{n,m}(t)\right]$, (b) contribution of resistive and inductive component of $SL_{n,m}$, (c) contribution of resistive and inductive component of all *ILs* between $SL_{n,m}$ and $SL_{j,i}$, and (d) numerator and denominator contributions of resistive and inductive component of $SL_{j,i}$.

APPENDIX B
GENERALITY FOR FAULT LOCATION AT IL_{i-1}

Proof. If the fault takes place at the interconnecting line between nodes *i-1* and *i*, then the fault location is evaluated using the effective faulty circuit as shown in Fig. 11(c). Using KVL for a PTP fault at IL_{i-1} with bus 1 of node 1 and bus *j* of node *i*, (23)-(24) are obtained respectively.

$$\begin{aligned} 2r_{1,1}D_{1,1}i_{1,1}(t) + 2\sum_{\beta=1}^{i-2} M_\beta r_{M_\beta} \sum_{\varepsilon=1}^{\beta} i_{0\varepsilon} + 2(M_{i-1}-x)r_{M_{i-1}} \sum_{\varepsilon=1}^{i-1} i_{0\varepsilon}(t) + \\ R_f \sum i(t) + l_{1,1}D_{1,1}\frac{v_{L_{1,1}}}{L_{1,1}} + \sum_{\beta=1}^{i-2} l_{M_\beta}M_\beta \sum_{(\alpha,\varepsilon)=(1,1)}^{p_\beta,\beta} \frac{v_{L_{\alpha,\beta}}}{L_{\alpha,\beta}} + (M_{i-1}-x)l_{M_{i-1}} \\ \sum_{(\alpha,\beta)=(1,1)}^{p_\beta,i-1} \frac{v_{L_{\alpha,\beta}}}{L_{\alpha,\beta}} = u_{1,1}(t) \end{aligned} \quad (23)$$

$$\begin{aligned} 2r_{j,i}D_{j,i}i_{j,i}(t) + 2xr_{M_{i-1}} \sum_{\varepsilon=i}^{k} i_{0\varepsilon}(t) + R_f \sum i(t) + l_{j,i}D_{j,i}\frac{v_{L_{j,i}}}{L_{j,i}} + \\ xl_{M_{i-1}} \sum_{(\alpha,\beta)=(1,i)}^{p_\beta,k} \frac{v_{L_{\alpha,\beta}}}{L_{\alpha,\beta}} = u_{j,i}(t) \end{aligned} \quad (24)$$

Rearranging equations (23)-(24) give the generalized fault location expression from bus *m* of node *n* as given in equation (18b). This completes the proof of theorem 2. □

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 08:30:53 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Power Delivery. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRD.2023.3318689

The equivalent fault distance and the reliability conditions can be worked out similarly to equations (21)-(22). Similar to the case of fault at $SL_{j,i}$, throughout the study, the calculated fault distance for a fault at IL_{i-1} is noticed to be composed of (a) $\left[u_{j,i}(t) - u_{n,m}(t)\right]$, (b) contribution of resistive and inductive component of $SL_{n,m}$, (c) contribution of resistive and inductive component of all *ILs* between $SL_{n,m}$ and IL_{i-1}, (d) contributions of resistive and inductive component of $SL_{j,i}$, and (e) numerator and denominator contributions of resistive and inductive component of IL_{i-1}.

Vaibhav Nougain received a B.Tech degree in electrical engineering in 2017 from Delhi Technological University (Formerly Delhi College of Engineering), and a Ph.D. degree in electrical engineering, from Indian Institute of Technology Delhi, New Delhi, India in 2022. He was a Visiting Student with The Center for Electric Power and Energy (CEE), Technical University of Denmark in 2019. Since January 2023, Vaibhav has been working as a Postdoc Researcher at TU Delft, Faculty of Electrical Engineering, Mathematics, and Computer Science in the group Intelligent Electrical Power Grids. His research interests include the protection and control of power electronics interfaced power systems.

Sukumar Mishra (Senior Member, IEEE) received his M.Tech and Ph.D. in Electrical Engineering from the National Institute of Technology, Rourkela in 1992 and 2000 respectively. Presently, Dr. Mishra is a Professor at the Indian Institute of Technology Delhi and has been its part for the past 17 years. He has won many accolades such as the INSA Medal for Young Scientist (2002), the INAE Young Engineer Award (2009, 2002), the INAE Silver Jubilee Young Engineer Award (2012), Samanta Chandra Shekhar Award (2016), IETE Bimal Bose award (2019), National Mission Innovation Champion award (2019) and NASI-Reliance Industries Platinum Jubilee Award for Application Oriented Innovation in Physical Sciences (2019). He has been granted fellowships from academies like NASI (India), INAE (India), and professional societies like IET (U.K.), IETE (India), and IE (India). He has also been recognized as the INAE Industry Academic Distinguished Professor. His research interests lie in the fields of Power Systems, Power Quality Studies, Renewable Energy, and Smart Grid.

Prof. Mishra is currently acting as the ABB Chair professor and has previously delegated as the NTPC, INAE, and Power Grid Chair professor. He has also served as an Independent Director of the Cross Border Power Transmission Company Ltd. and the River Engineering Pvt. Ltd. And has carried out many important industrial consultations with TATA Power, Microtek, and others. Prof. Mishra has been working in close association with the IEEE Delhi Section Executive Committee for the past few years and is currently serving as an Editor for the IEEE Transactions on Smart Grid, and IEEE Transactions on Sustainable Energy and was an Area Editor for the IET Generation, Transmission & Distribution journal.

Soumya Shubhra Nag (Member, IEEE) received the B.E. degree from Jadavpur University, Kolkata, India, in 2009, and the M.Tech. and Ph.D. degrees from the Indian Institute of Technology (IIT) Kanpur, Kanpur, India, in 2011 and 2017, respectively, all in electrical engineering. From January 2017 to July 2018, he was a Research Fellow with the Rolls-Royce Corporate Laboratory, at Nanyang Technological University (NTU), Singapore. He is currently an Assistant Professor at IIT Delhi, New Delhi, India. His research interests include power electronic converters for transportation electrification, DC distribution systems, and renewable energy systems.

Aleksandra Lekić (M'18, SM'22) received the B.Sc, M.Sc, and Ph.D. degrees in electrical engineering from the School of Electrical Engineering, University of Belgrade, Serbia, in 2012, 2013, and 2017, respectively. Between 2012 and 2018 she has been a Teaching Assistant with the School of Electrical Engineering, University of Belgrade, and an Assistant Professor from 2018 to 2019. In 2019 she worked as a Postdoctoral Researcher at the Department of Electrical Engineering (ESAT), KU Leuven, and at the Institute EnergyVille, Genk, Belgium. Since January 2020, Aleksandra works as a tenured Assistant Professor at TU Delft, Faculty of Electrical Engineering, Mathematics and Computer Science in the group Intelligent Electrical Power Grids.

Aleksandra leads a team of researchers specializing in the Control of HVDC/AC power systems. She is an Associate Editor in the International Journal of Electrical Power & Energy Systems, Elsevier. Aleksandra represents TU Delft in the General Assembly in the CRESYM organization, committed to open-source software development for the electrical grids.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 08:30:53 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

Check for updates

Received: 7 March 2023 | Accepted: 28 June 2023

DOI: 10.1002/hcs2.65

ORIGINAL ARTICLE

HEALTH CARE SCIENCE

Financial burden of seeking diabetes mellitus care in India: Evidence from a Nationally Representative Sample Survey

Mehak Nanda[1] | **Rajesh Sharma**[2]

[1]University School of Management and Entrepreneurship, Delhi Technological University, Delhi, India

[2]Department of Humanities and Social Sciences, National Institute of Technology Kurukshetra, Kurukshetra, Haryana, India

Correspondence
Rajesh Sharma, Department of Humanities and Social Sciences, National Institute of Technology Kurukshetra, Haryana, India.
Email: rajesh.sharma@nitkkr.ac.in

Funding information
None

Abstract

Background: Diabetes mellitus (DM) is a major public health concern in India, and entails a severe burden in terms of disability, death, and economic cost. This study examined the out-of-pocket health expenditure (OOPE) and financial burden associated with DM care in India.

Methods: The study used data from the latest round of the National Sample Survey on health, which covered 555,115 individuals from 113,823 households in India. In the present study, data of 1216 individuals who sought inpatient treatment and 6527 individuals who sought outpatient care for DM were analysed.

Results: In India, 10.04 per 1000 persons reported having DM during the last 15 days before the survey date, varying from 6.94/1000 in rural areas to 17.45/1000 in urban areas. Nearly 38% of Indian households with diabetic members experienced catastrophic health expenditure (at the 10% threshold) and approximately 10% of DM-affected households were pushed below the poverty line because of OOPE, irrespective of the type of care sought. 48.5% of households used distressed sources to finance the inpatient costs of DM. Medicines constituted one of the largest proportion of total health expenditure, regardless of the type of care sought or type of healthcare facility visited. The average monthly OOPE was over 4.5-fold and 2.5-fold higher for households who sought inpatient and outpatient care, respectively, from private health facilities, compared with those treated at public facilities. Notably, the financial burden was more severe for households residing in rural areas, those in lower economic quintiles, those belonging to marginalised social groups, and those using private health facilities.

Abbreviations: CHE, catastrophic health expenditure; DALYs, disability-adjusted life years; DM, diabetes mellitus; GDP, gross domestic product; LMICs, low and middle-income countries; NCDs, noncommunicable diseases; OBCs, other backward classes; OOPE, out-of-pocket health expenditure; OR, odds ratio; PHCs, primary health centres; PMJAY, Pradhan Mantri Jan Arogya Yojana; SCs, scheduled castes; STs, scheduled tribes; UHC, universal health coverage.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2023 The Authors. *Health Care Science* published by John Wiley & Sons Ltd on behalf of Tsinghua University Press.

Conclusion: The burden of DM and its associated financial ramifications necessitate policy measures, such as prioritising health promotion and disease prevention strategies, strengthening public healthcare facilities, improved regulation of private healthcare providers, and bringing outpatient services under the purview of health insurance, to manage the diabetes epidemic and mitigate its financial impact.

KEYWORDS
out-of-pocket health expenditure, distressed financing, catastrophic health expenditure, diabetes mellitus

1 | INTRODUCTION

Diabetes mellitus (DM) is a growing public health concern and is ranked among the top-10 causes of mortality and disability-adjusted life years (DALYs) worldwide [1]. In 2021, 537 million people globally had DM, and this number is expected to increase to 783 million by 2045 [2]. Notably, over 80% of individuals with DM reside in low and middle-income countries (LMICs) [2]. India has been described as the 'Diabetes Capital of the World' [3], accounting for one in every seven adults living with DM globally [2]. In 2021, India was ranked second in terms of the number of adults living with DM, and third in terms of DM-related mortality worldwide [2]. Between 1990 and 2016, India experienced a 29.3% increase in age-standardised prevalence of DM and 39.6% increase in age-standardised DALYs rate for DM, the highest increase among major non-communicable diseases (NCDs) [4, 5]. The economic consequences of DM are dire, resulting in substantial medical and non-medical expenses as well as productivity losses in LMICs such as India [6–10]. As per the International Diabetes Federation, the total DM-related health expenditure in India was USD 8.5 billion in 2021 and is projected to rise to USD 10.3 billion by 2030 [11]. Moreover, India has one of the lowest public health expenditure (1.15% of GDP) [12], and one of the highest out-of-pocket health expenditure (OOPE) (58.7% of total health expenditure) in the world [13]. Heavy dependence on OOPE impedes access to healthcare services, as well as exposes households to financial catastrophe and falling into or intensifying poverty.

The rising burden of DM in India [2, 14], accompanied by dismally low health insurance coverage [15], a significant reliance on OOPE, and the dominant presence of the fee-for-service private health sector [16], highlights the substantial economic burden imposed on DM-affected households. To mitigate this economic impact, it is crucial to conduct a comprehensive examination of the financial strain associated with diabetes care in the country. However, limited evidence is available regarding the financial impact of seeking DM care in India. Previous studies were restricted to specific geographical areas [17–21] or relied on older data sets [22, 23], focusing on a few aspects such as assessing the OOPE and associated financial burden due to DM only across economic quintiles [23] or only for inpatient care [22].

Our study aims to fill the gap in the existing literature by comprehensively examining the OOPE and financial burden associated with DM care in India. For this purpose, we employed the latest nationally representative survey data conducted by the National Sample Survey Organisation (NSSO). Specifically, our study was guided by the following objectives. First, we estimated the number of people who reported having DM in India. Second, we estimated OOPE and determined the proportion of various components (e.g., doctors' fees, the cost of medicines, diagnostic tests) of total health expenditure. Third, we estimated the financial burden of OOPE using three metrics: incidence of experiencing catastrophic health expenditure (CHE), percentage of households falling below the poverty line because of OOPE, and the incidence of relying on distressed sources. Fourth, we estimated the percentage of diabetic individuals who did not seek treatment, and the percentage of diabetic individuals who did not seek treatment on medical advice and their reasons for doing so. Finally, we assessed the loss of household income resulting from seeking hospitalisation and outpatient care due to DM. All measures were estimated separately by the type of care sought (inpatient and outpatient) and type of healthcare facility visited (public and private) to facilitate a nuanced and detailed analysis. Examining the financial burden of DM care is crucial from a policy standpoint to ameliorate the financial hardships experienced by diabetic patients and their families. Our analysis using

27711757, 2023, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hcs2.65 by Delhi Technological University, Wiley Online Library on [02/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

nationally representative estimates of costs and financial impact are expected to assist in designing strategies to enhance financial risk protection for Indian households.

2 | DATA AND METHODOLOGY

2.1 | Overview of data source

The study employed data from the 75th round of the National Sample Survey on health, named 'Household Social Consumption: Health'. The survey was conducted by the NSSO during 2017–2018 and covered 555,115 individuals from 113,823 households in India. The survey used a stratified multistage sampling design, with villages and urban blocks as the first stage units in rural and urban areas, respectively, and households as the second stage units. In addition to collecting information about socioeconomic and demographic characteristics, the survey gathered extensive information about the nature of ailments, utilisation patterns of healthcare facilities, and expenses incurred on inpatient and outpatient treatment. In this study, we analysed data of 1216 individuals (from 1215 households) who sought inpatient treatment for DM and 6527 individuals (from 5719 households) who sought outpatient care for DM.

2.2 | Outcome variables

2.2.1 | Out-of-pocket health expenditure (OOPE)

The NSSO survey collected detailed information about the total health expenditure under three subcategories: medical, nonmedical, and transportation expenses. To calculate the OOPE, we subtracted any reimbursements received from the total health expenditure. The survey recorded inpatient and outpatient expenses incurred during the last 365 days and 15 days, respectively, which were converted into monthly figures.

2.2.2 | Catastrophic health expenditure (CHE)

A household is defined to incur CHE if OOPE exceeds a certain threshold of the household's income or expenditure [24, 25].

$$\mathrm{CHE}_i = \begin{cases} 1, \text{ if } \dfrac{\mathrm{OOPE}_i}{\mathrm{TE}_i} > \mathrm{Z} \\ 0, \text{ otherwise} \end{cases}$$

where, OOPE_i represents the out-of-pocket health expenditure of the i^{th} household, TE_i represents the total consumption expenditure of the i^{th} household, and Z is the threshold.

In the existing literature, there is no consensus regarding the threshold. Hence, we computed the CHE at three thresholds: 10%, 20%, and 40% [26].

The proportion of households experiencing CHE was estimated using the following formula.

$$\text{Incidence of CHE} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{CHE}_i$$

where N is the total number of households.

2.2.3 | Poverty headcount ratio

The poverty headcount ratio estimates the proportion of households pushed below the poverty line because of OOPE [27].

$$\text{Poverty Headcount}_i = \begin{cases} 1, \text{ if } \mathrm{TE}_i \geq \mathrm{PL} \text{ and } (\mathrm{TE}_i - \mathrm{OOPE}_i) < \mathrm{PL} \\ 0, \text{ otherwise} \end{cases}$$

where PL is the inflation-adjusted official poverty line as defined by the Tendulkar Committee [28].

$$\text{Poverty Headcount Ratio} = \frac{1}{N}\sum_{i=1}^{N} \text{Poverty Headcount}_i$$

2.2.4 | Distressed financing

We defined a household as using distressed financing if it relied on borrowings, sale of physical assets, contributions from friends and relatives, or other sources to pay for OOPE [29]. In the case of inpatient care, the NSSO survey grouped sources of finance into major and second major categories, as households might have used more than one source. Therefore, if a household used borrowings and contributions as its major and second major sources of finance, respectively, we included such household in both categories [29]. Hence, the total incidence of all sources of finance could be greater than 100% for inpatient care.

2.3 | Statistical analysis

Descriptive statistics and multivariable logistic regression were employed in the study.

27711757, 2023, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hcs2.65 by Delhi Technological University, Wiley Online Library on [02/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Multivariable logistic regression was used to gauge the socioeconomic and demographic factors associated with experiencing CHE, impoverishment, and distressed financing due to OOPE for DM.

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n.$$

In the above equation, π_i is the probability of occurrence of a binary outcome variable, that is, incurring CHE, impoverishment, and distressed financing for the i^{th} household. $X_1 \ldots X_n$ represent variables, namely economic quintiles, social group[1] (i.e., Scheduled Tribes [STs], Scheduled Castes [SCs], Other Backward Classes [OBCs], and Others), household's principal source of income, religion (i.e., Hinduism, Islam, Others), place of residence (i.e., rural or urban areas), gender and educational status of household head, any elderly member in household, health insurance status, and type of healthcare facility visited.

Sample weights provided by the NSSO were applied as applicable. All analyses were performed using Stata Version 14.1.

3 | RESULTS

In India, 10.04 per 1000 persons reported having DM during the last 15 days before the survey date, varying from 6.94/1000 in rural areas to 17.45/1000 in urban areas (Supporting Information: Table 1). Females reported a slightly higher burden of DM (10.43/1000) compared with males (9.68/1000). Individuals aged 60 years and above reported the highest burden of DM (65.73/1000). The number of individuals reporting DM was higher among those belonging to other social groups (16.35/1000) compared with STs (2.09/1000), SCs (6.62/1000), and OBCs (9.43/1000). Notably, the burden of DM ranged from 3.55/1000 among individuals belonging to the poorest economic quintile to 21.91/1000 among those in the wealthiest economic quintile. The number of persons reporting DM was substantially higher among individuals primarily earning through regular wages or salaries (14.01/1000) and other work (38.98/1000) compared with those working as casual labourers (6.07/1000) and those who were self-employed (8.25/1000).

[1]STs and SCs are the two most socially backward and economically disadvantaged social groups in India [30, 31]. OBCs include backward classes of citizens other than SCs and STs as specified in the lists prepared by the Government of India, which are periodically updated [32].

3.1 | OOPE and share of various components

Supporting Information: Table 2 shows the average monthly OOPE of households by the type of care sought. The average OOPE was INR 1555.4[2] for inpatient care and INR 1380.2 for outpatient care. Urban households reported higher OOPE than their rural counterparts, irrespective of the type of care sought. The OOPE was invariably higher for households who sought DM treatment from private facilities compared with those treated in public facilities in the case of both inpatient (INR 2139.6 vs. INR 459.8) and outpatient care (INR 1760.3 vs. INR 690.0) (Supporting Information: Tables 3–6). 62.6% of inpatient cases and 66.3% of outpatient cases were sought at private health facilities (Figure 1; Supporting Information: Table 7). The key reasons for not choosing the public health facilities (by those who chose private health facilities) were nonavailability of doctors or the quality of public health facilities being unsatisfactory, and a preference for a trusted doctor or hospital, regardless of the type of care sought (Supporting Information: Figure 1).

Medicines constituted one of the largest share of expenses in the case of both inpatient care (private facilities: 27.6%; public facilities: 24.7%) and outpatient care (private facilities: 76.3%; public facilities: 43.5%) (Figure 2). Notably, the relative burden of nonmedical expenses was higher at public facilities compared with private facilities, both for inpatient (32.9% vs. 8.8%) and outpatient care (13.2% vs. 2.4%). Likewise, the share of transportation expenses was over 3.5 times (7.5 times) higher for public facilities compared with private facilities in the event of inpatient (outpatient care). Conversely, when hospitalisation was sought at private facilities, bed charges (14.1%) and doctors' fees (13.1%) were the leading contributors to the total health expenditure after medicines.

3.2 | Catastrophic health expenditure

Of all the households who sought inpatient care for DM, 37.9%, 20.3%, and 8.8% of households experienced CHE at the 10%, 20%, and 40% thresholds, respectively (Table 1). The CHE incidence in the case of outpatient care for DM was 37.9%, 20.7%, and 8.9% at the 10%, 20%, and 40% thresholds,

[2]USD 1 = INR 68.3 using average 2018 exchange rate.

27711757, 2023, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hcs2.65 by Delhi Technological University, Wiley Online Library on [02/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

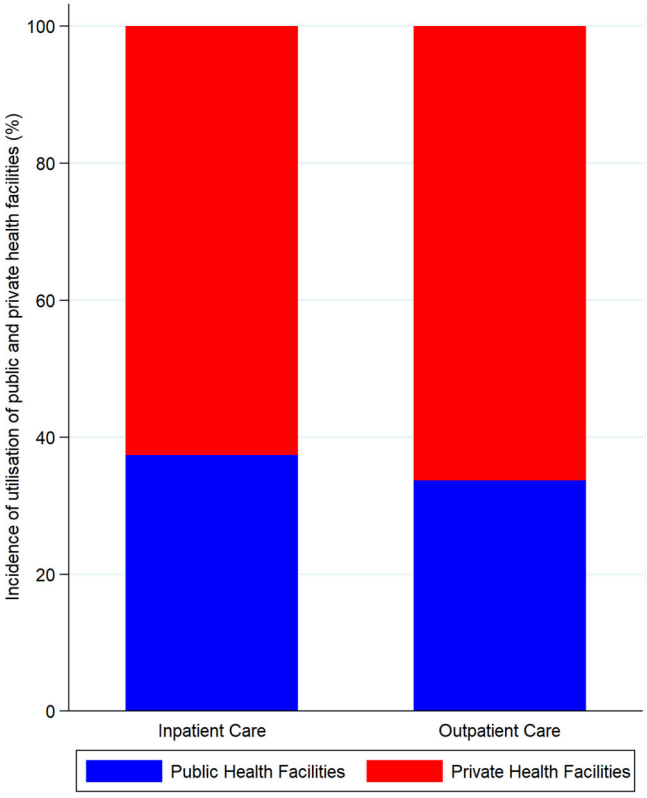


FIGURE 1 Incidence of utilisation of public and private health facilities.

respectively. Rural households experienced higher CHE incidence compared with their urban counterparts, regardless of the type of care sought and across all thresholds. Households belonging to marginalised social groups (STs and SCs) reported higher CHE incidence compared with those belonging to OBCs and other social groups across all thresholds for both inpatient and outpatient care. In the case of outpatient care for DM, households belonging to the bottom two economic quintiles reported higher CHE incidence at the 10% threshold (Quintile 1: 59.7%; Quintile 2: 41.9%) than those belonging to the wealthiest quintile (Quintile 5: 30.8%), with a similar pattern observed across all thresholds. Furthermore, households whose principal source of earnings was from regular wages or salaries reported the lowest CHE incidence at all thresholds for inpatient care. The CHE incidence (at all thresholds) was over 3-fold and 2-fold higher for households who sought DM care from private health facilities compared with those who were treated in public facilities in the event of inpatient and outpatient care, respectively (Supporting Information: Tables 3–6).

3.3 | Poverty headcount ratio

Table 2 shows that 10.4% and 9.9% of all DM-affected households who sought inpatient and outpatient care, respectively, were pushed below the poverty line because of OOPE. The poverty headcount ratio was substantially higher among rural households (inpatient care: 13.6%; outpatient care: 13.6%) compared with their urban counterparts (inpatient care: 6.6%; outpatient care: 6.2%). Households belonging to STs (20.5%), SCs (15.8%), and OBCs (11.0%) reported higher poverty headcount ratios than those belonging to other categories (6.1%) in the case of inpatient care. The percentage of households falling below the poverty line was higher among lower economic quintiles compared with wealthier quintiles, regardless of the type of care sought. Furthermore, households primarily earning from regular wages or salaries reported the lowest poverty headcount ratio for inpatient care. The poverty headcount ratio was higher among households who sought DM care from private facilities compared with those treated in public facilities in the case of both inpatient (13.1% vs. 4.7%) and outpatient care (11.3 vs. 8.1%) (Supporting Information: Tables 3–6).

3.4 | Incidence of distressed financing

Of all the households who sought inpatient treatment, 48.5% relied on distressed sources either as the major or second major source to cope with the cost of DM care (Table 3). Borrowings (25.3%) and contributions from friends and relatives (18.5%) were the most common distressed sources for financing inpatient care, whereas outpatient care was predominantly financed through income and savings (94.9%) (Supporting Information: Figure 2). Notably, more than 50% of households residing in rural areas, those belonging to lower economic quintiles, those belonging to OBCs, and those earning from casual work used distressed sources to pay for inpatient-related OOPE (Table 3). The incidence of using distressed sources was higher among households who sought inpatient services from private health facilities (50.9%) than those treated in public facilities (43.3%) (Supporting Information: Tables 3 and 4).

3.5 | Proportion of diabetic individuals not seeking treatment

Of all the individuals who reported having DM during the last 15 days before the survey date, 0.8% did not seek treatment (Supporting Information: Table 8). The incidence of not seeking treatment

27711757, 2023, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hcs2.65 by Delhi Technological University, Wiley Online Library on [02/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

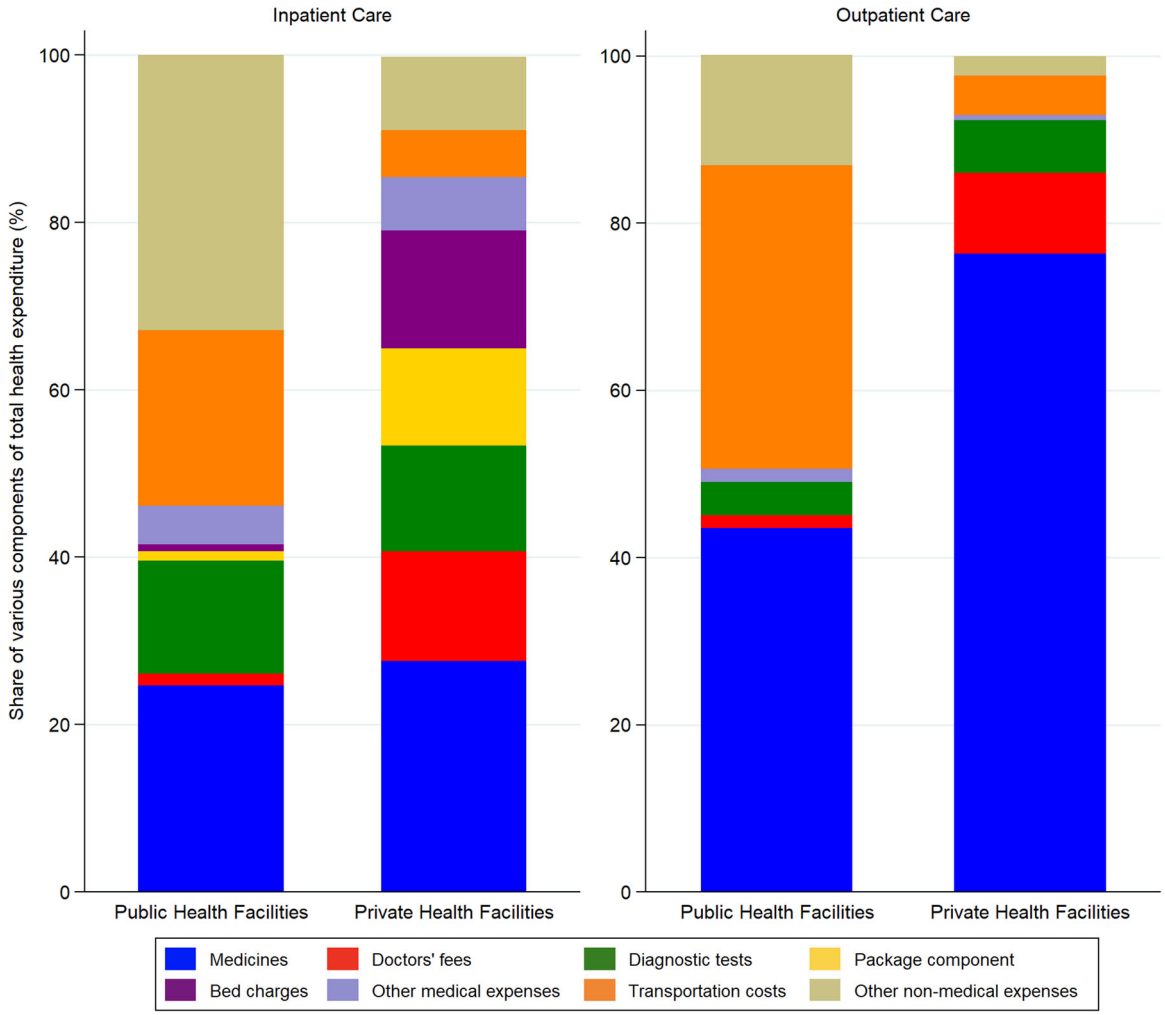


FIGURE 2 Share of various components of total health expenditure by the type of care and health facility.

was the highest among individuals belonging to the lowest economic quintile (5.2%), those belonging to OBCs (1.5%), and those residing in rural areas (1.4%). We also found that 2.6% of diabetic individuals did not seek treatment on medical advice during the last 15 days (Supporting Information: Table 8). The primary reasons for not seeking treatment on medical advice were that the ailment was not considered severe (43.0% of cases) and other reasons (50.1% of cases). In only 0.9% of cases, individuals reported financial reasons (i.e., expensive facilities) for not seeking treatment on medical advice, ranging from 0.1% in rural areas to 2.2% in urban areas. In contrast, a considerable proportion of individuals from the lowest economic quintile (20.5%) and OBCs (10.4%) reported the nonavailability of medical facilities in their neighbourhood as one of the reasons for not seeking treatment on medical advice (Supporting Information: Figure 3).

3.6 | Loss of household income

Households not only bear the financial cost of healthcare but also experience indirect costs, such as loss of earnings, when the patient or caregiver is unable to work. We found that the average loss of household income caused by hospitalisation for DM was INR 3330.9, and that caused by outpatient care was INR 88.2 (Supporting Information: Table 9). Households belonging to STs (INR 4769.7) and other social groups (INR 4355.6), those in the middle-economic quintile (INR 4226.3), those primarily earning through self-employment (INR 4031.2), and those residing in urban areas (INR 3976.3) reported the highest loss of household income because of inpatient treatment for DM.

3.7 | Multivariable logistic regression

Supporting Information: Table 10 shows the results of multivariable logistic regression to reveal the impact

27711757, 2023, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hcs2.65 by Delhi Technological University, Wiley Online Library on [02/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 Incidence of incurring catastrophic health expenditure.

Socioeconomic characteristics	Inpatient care			Outpatient care		
	At 10% threshold (%)	At 20% threshold (%)	At 40% threshold (%)	At 10% threshold (%)	At 20% threshold (%)	At 40% threshold (%)
Region						
Rural	44.7 (40.5–48.8)	22.6 (19.2–26.1)	9.8 (7.3–12.3)	42.9 (40.9–44.9)	25.7 (23.9–27.5)	11.2 (9.9–12.5)
Urban	29.6 (26.1–33.1)	17.4 (14.4–20.3)	7.6 (5.5–9.6)	33.0 (31.4–34.6)	15.8 (14.6–17.0)	6.6 (5.8–7.4)
Social group						
Scheduled tribes	60.1 (51.2–69.0)	41.9 (32.9–50.9)	16.3 (9.6–23.0)	60.8 (54.3–67.3)	36.7 (30.3–43.1)	18.6 (13.4–23.8)
Scheduled castes	43.2 (35.6–50.9)	27.4 (20.5–34.3)	11.2 (6.3–16.1)	40.3 (36.7–43.9)	24.2 (21.0–27.3)	8.3 (6.3–10.4)
Other backward classes	32.1 (28.0–36.1)	20.2 (16.7–23.7)	9.4 (6.9–11.9)	36.0 (34.1–38.0)	19.5 (17.8–21.1)	8.7 (7.5–9.8)
Other	41.9 (37.2–46.6)	14.8 (11.4–18.2)	5.9 (3.7–8.2)	38.0 (36.1–39.9)	20.2 (18.6–21.7)	8.8 (7.7–10.0)
Major source of household income						
Self-employment	42.4 (38.2–46.6)	23.8 (20.1–27.4)	10.2 (7.6–12.7)	38.8 (36.9–40.7)	21.4 (19.8–23.1)	10.0 (8.8–11.2)
Regular wage or salary	23.8 (19.3–28.4)	12.7 (9.1–16.2)	3.7 (1.7–5.7)	33.9 (31.6–36.2)	18.2 (16.3–20.0)	6.0 (4.8–7.1)
Casual labour	40.7 (33.4–47.9)	18.5 (12.8–24.3)	8.7 (4.5–12.9)	33.8 (30.4–37.2)	16.4 (13.8–19.1)	6.3 (4.6–8.0)
Other	43.0 (35.7–50.4)	26.2 (19.6–32.7)	14.0 (8.8–19.2)	45.6 (42.2–48.9)	26.8 (23.8–29.8)	12.9 (10.7–15.1)
Economic quintiles						
Quintile 1	55.3 (47.4–63.2)	18.4 (12.3–24.6)	8.9 (4.4–13.4)	59.7 (55.1–64.3)	42.5 (37.9–47.2)	16.1 (12.7–19.6)
Quintile 2	35.6 (28.7–42.5)	21.0 (15.1–26.9)	8.8 (4.8–12.9)	41.9 (38.5–45.4)	24.1 (21.2–27.1)	14.9 (12.5–17.4)
Quintile 3	36.6 (30.0–43.2)	25.1 (19.2–31.0)	12.1 (7.7–16.6)	39.5 (36.3–42.7)	21.6 (18.9–24.3)	7.8 (6.0–9.5)
Quintile 4	39.7 (33.6–45.8)	23.8 (18.5–29.1)	6.7 (3.6–9.8)	41.7 (38.9–44.5)	22.6 (20.2–24.9)	9.5 (7.8–11.2)
Quintile 5	31.7 (27.2–36.2)	15.3 (11.9–18.8)	8.4 (5.7–11.0)	30.8 (28.9–32.7)	15.0 (13.5–16.4)	5.9 (5.0–6.9)
Total	37.9 (35.2–40.6)	20.3 (18.0–22.5)	8.8 (7.2–10.4)	37.9 (36.7–39.2)	20.7 (19.7–21.8)	8.9 (8.2–9.6)

Note: () 95% confidence interval.

27711757, 2023, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hcs2.65 by Delhi Technological University, Wiley Online Library on [02/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 2 Percentage of households falling below the poverty line because of out-of-pocket health expenditure on diabetes mellitus.

Socioeconomic characteristics	Inpatient care (%)	Outpatient care (%)
Region		
Rural	13.6 (10.8–16.5)	13.6 (12.2–15.0)
Urban	6.6 (4.7–8.5)	6.2 (5.4–7.0)
Social group		
Scheduled tribes	20.5 (13.2–27.9)	7.8 (4.2–11.4)
Scheduled castes	15.8 (10.1–21.5)	12.7 (10.3–15.2)
Other backward classes	11.0 (8.3–13.7)	10.0 (8.8–11.3)
Other	6.1 (3.8–8.4)	9.0 (7.8–10.1)
Major source of household income		
Self-employment	12.0 (9.3–14.8)	11.7 (10.4–12.9)
Regular wage or salary	4.6 (2.3–6.8)	8.0 (6.7–9.3)
Casual labour	11.5 (6.7–16.2)	7.7 (5.8–9.6)
Other	13.8 (8.6–18.9)	10.2 (8.2–12.2)
Economic quintiles		
Quintile 1	7.8 (3.5–12.0)	19.6 (15.9–23.4)
Quintile 2	20.3 (14.5–26.1)	24.2 (21.2–27.2)
Quintile 3	16.4 (11.3–21.5)	14.2 (11.9–16.5)
Quintile 4	8.2 (4.8–11.6)	9.4 (7.8–11.1)
Quintile 5	4.6 (2.6–6.6)	2.4 (1.8–3.0)
Total	10.4 (8.7–12.2)	9.9 (9.1–10.6)

Note: () 95% confidence interval.

of various socioeconomic and demographic factors on the likelihood of incurring CHE (at the 10% threshold) due to DM. Rural households were significantly more likely to experience CHE for both inpatient (odds ratio [OR]: 2.59 [1.56–4.29]; $p < 0.05$) and outpatient care (OR: 2.64 [1.92–3.63]; $p < 0.05$) compared with their urban counterparts. Utilisation of private health facilities significantly increased the likelihood of experiencing CHE, irrespective of the type of care sought ($p < 0.05$). Households belonging to STs were more likely to experience CHE (OR: 4.32 [1.93–9.70]; $p < 0.05$) for inpatient care compared with other social groups. In the case of outpatient care, female-headed households (OR: 1.59 [1.07–2.37]; $p < 0.05$) and households primarily earning from other work (OR: 1.72 [1.11–2.65]; $p < 0.05$) were significantly more likely to incur CHE. Conversely, the odds of experiencing CHE were significantly lower among households belonging to higher economic quintiles, regardless of the type of care sought ($p < 0.05$). Insurance coverage significantly decreased the odds of incurring CHE for inpatient care (OR: 0.57 [0.33–0.96]; $p < 0.05$). Households earning primarily from regular wages or salaries were significantly less likely to incur CHE for inpatient care (OR: 0.46 [0.26–0.81]; $p < 0.05$) compared with those earning from self-employment. Similarly, place of residence, economic quintile, and type of healthcare facility visited emerged as statistically significant predictors of households' likelihood of falling into poverty and using distressed financing (Supporting Information: Table 10).

4 | DISCUSSION

Universal Health Coverage (UHC), which aims to ensure that everyone has access to quality healthcare without facing financial hardship [33], has been adopted as one of the Sustainable Development Goals [34]. Financial protection is only achieved when there is no financial hardship caused by OOPE and no financial

27711757, 2023, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hcs2.65 by Delhi Technological University, Wiley Online Library on [02/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 3 Incidence of using distressed financing.

Socioeconomic characteristics	Inpatient care %	Outpatient care %
Region		
Rural	54.4 (50.2–58.5)	5.5 (4.6–6.5)
Urban	41.4 (37.6–45.2)	4.7 (3.9–5.4)
Social group		
Scheduled tribes	44.4 (35.3–53.6)	9.7 (5.5–13.9)
Scheduled castes	47.8 (40.0–55.6)	3.9 (2.4–5.4)
Other backward classes	50.1 (45.7–54.4)	6.0 (5.0–7.0)
Other	47.0 (42.3–51.8)	4.4 (3.5–5.2)
Major source of household income		
Self-employment	43.2 (39.0–47.5)	2.6 (2.0–3.3)
Regular wage or salary	36.8 (31.5–42.0)	4.7 (3.7–5.8)
Casual labour	69.9 (63.0–76.7)	4.4 (2.9–6.0)
Other	42.9 (35.5–50.3)	12.6 (10.3–14.9)
Economic quintiles		
Quintile 1	59.0 (51.2–66.8)	4.8 (2.7–6.8)
Quintile 2	57.0 (49.8–64.1)	4.6 (3.1–6.1)
Quintile 3	61.7 (55.0–68.4)	5.7 (4.2–7.3)
Quintile 4	39.1 (33.0–45.2)	1.8 (1.0–2.6)
Quintile 5	39.3 (34.6–44.0)	6.7 (5.7–7.7)
Total	48.5 (45.7–51.4)	5.1 (4.5–5.7)

Note: () 95% confidence interval.

barriers to accessing care [35]. The prominent parameters used in the existing literature to capture the undesirable effects of OOPE are CHE and impoverishment [36–38]. Households may even resort to distressed coping strategies, such as borrowing money, sale of assets, and seeking contributions from friends and relatives to finance OOPE [29, 39]. Furthermore, households might forgo necessary healthcare because of financial constraints, thereby exacerbating health problems and putting the concerned families into a downward spiral of ill-health and poverty [39–41]. As per a recent review [39], comprehensive financial risk studies must measure four indicators to provide a holistic picture of the financial hardships experienced by households: (i) CHE, (ii) impoverishment, (iii) adoption of distressed coping strategies and (iv) forgone care due to financial reasons. Therefore, in the current study, we estimated the financial burden by examining all four parameters in respect of DM care in India.

We found that 37.9% of Indian households with diabetic members experienced CHE (at 10% threshold) and nearly 10% of DM-affected households fell below the poverty line because of OOPE, regardless of the type of care sought. Outpatient care was primarily financed through income or savings, whereas 48.5% of households used distressed sources to finance inpatient care for DM. Additionally, 0.8% of individuals with DM did not seek treatment and 2.6% of individuals did not seek treatment on medical advice. Medicines were the main constituent of health expenditure, regardless of the type of care sought and healthcare facility visited. Moreover, OOPE and the associated financial burden were substantially higher for households seeking DM care from private health facilities compared with those treated in the public facilities.

We observed that the number of persons reporting having DM was higher among the elderly, urban residents, individuals belonging to wealthier quintiles, other social groups, those with higher educational status, and those primarily earning from regular wages or salaries and other work. The higher burden of DM among individuals with higher socioeconomic status might reflect greater exposure to risk factors, such as

27711757, 2023, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hcs2.65 by Delhi Technological University, Wiley Online Library on [02/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

excessive calorie intake, sedentary lifestyles, and low physical activity [42–48]. Additionally, the higher burden could also be attributed to greater awareness and access to healthcare among such individuals [49]. Importantly, knowledge and awareness about DM and its risk factors is very low in India, particularly in rural areas [50, 51]. According to the Indian Council of Medical Research India Diabetes Study [50], only 43.2% of the study population had heard of diabetes, and the level of awareness was significantly lower among rural individuals (36.8%) compared with urban individuals (58.4%). If DM is not detected early and treated adequately, there is an increased risk of developing DM-related complications, increased healthcare use, and associated costs [2, 51].

In accord with previous studies [21, 23, 52–54], we found that DM imposes a substantial financial burden on Indian households. We observed that households residing in rural areas, those belonging to lower economic quintiles, those belonging to marginalised social groups (STs and SCs), and those primarily earning through casual, self-employment and other work, experienced higher financial burden caused by DM care. According to a case–control study, individuals with DM reported 2 times more days of inpatient treatment, 1.3 times more outpatient visits, and 9.7 times more medications than nondiabetic individuals; thereby, imposing severe financial strain on DM-affected households [55]. The high cost of DM care forces poor and vulnerable people to delay treatment [56], reduce their medication dosage to make drugs last for a longer period [57] or resort to borrowing or selling assets to finance care [23, 58]. The situation is further complicated by inadequate public healthcare facilities and a scarcity of healthcare workers, particularly specialists (for instance, 80% shortfall of specialists at community health centres) in the Indian rural healthcare system [59]. Additionally, paucity of basic items such as glucometers or test strips for glucometers is a major concern. For instance, a previous study reported that only 20% of primary health centres (PHCs) in the state of Odisha had glucometers [60]. This situation forces people to travel long distances to access medical care [61–63], leading to substantial costs associated with travel and lodging and income loss caused by it. Therefore, it is crucial to strengthen the public healthcare system to ensure timely access to care, improve health outcomes, and reductions in OOPE. In this regard, the creation of 150,000 Health and Wellness Centres by transforming existing subhealth centres and PHCs to provide comprehensive primary healthcare for NCDs is a laudable initiative by the Indian government [64]. Moreover, telemedicine has the potential to overcome barriers and enhance the quality and accessibility of DM care in remote regions, as evidenced by successful programmes such as the Diabetes Tele Management System and the Chunampet Rural Diabetes Prevention Project Model [65–67].

Interestingly, we found that outpatient care was as burdensome as inpatient care, causing 37.9% of DM-affected households to incur CHE (at the 10% threshold) and pushing 9.9% of households below the poverty line. The financial burden caused by outpatient care can be attributed to recurring visits, relatively small but ongoing expenses, and over-dependence on the private healthcare sector [68, 69]. Moreover, in India, the rising prevalence of NCDs, such as DM and cardiovascular conditions, leads to a greater need for outpatient facilities as chronic illnesses demand regular health check-ups, doctors' visits, diagnostic tests, and long-term drug support [69, 70]. Despite this, the majority of the government-sponsored health insurance schemes in India cover only inpatient treatment, leaving outpatient care outside the purview of health insurance [71, 72]. Recently, the government of India launched the largest government-sponsored health insurance scheme, 'Pradhan Mantri Jan Arogya Yojana (PMJAY)', with the objective of providing annual cover of INR 500,000 per household for secondary and tertiary care inpatient services to over 100 million poor and vulnerable families [73]. However, it must be noted that the PMJAY scheme does not cover outpatient care, which is a significant contributor to the financial burden in India.

In accord with previous studies [23, 74], we found that medicines are among the largest contributors of healthcare expenses in the case of both inpatient (private facilities: 27.6%; public facilities: 24.7%) and outpatient care (private facilities: 76.3%; public facilities: 43.5%). The poor availability of free or subsidised essential drugs in public health facilities compels individuals to purchase medicines from open markets, leading to higher OOPE or forgone treatments [75]. Previous studies reported that the availability of essential diabetes medicines in both public and private sectors in the state of Kerala and insulin availability in the public sector in Bengaluru fell short of the World Health Organization target of 80% availability of essential medicines for NCDs [76, 77]. A recent study found that 51.4% and 24.6% of Indian households with DM members were unable to afford insulin and metformin, respectively, with the highest unaffordability reported in rural areas and among the lowest income tertile [78]. Moreover, as per a recent study examining the availability and affordability of essential medicines for diabetes in 22 countries, it was found that among the 2972 surveyed DM patients in India, 71.5% did not take any diabetes medicine [78]. The key impediments to accessibility to insulin are

27711757, 2023, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hcs2.65 by Delhi Technological University, Wiley Online Library on [02/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

unaffordable prices, excessive reliance on foreign-made insulin, and a problematic alliance between Indian doctors and foreign manufacturers, which forces patients to buy expensive insulin instead of cheaper local brands [79, 80]. To curtail expenditure on medicines, the Indian government introduced the Pradhan Mantri Bhartiya Janaushadhi Pariyojana to provide access to quality generic medicines at affordable prices [81]. Adopting generic medicines can significantly reduce the financial burden caused by DM [82]. Furthermore, it is imperative to improve drug procurement and supply chain systems to ensure a steady supply of free medicines in the public health facilities [76, 83].

The current study underscores the crucial role of public health facilities in delivering DM care, as the financial burden was substantially lower among households who sought treatment from public hospitals compared with those treated in private facilities. Nevertheless, we found that the shortcomings in the public health sector, such as unsatisfactory quality or unavailability of doctors and long waiting times, prompt individuals to seek care from private facilities, resulting in financial hardships. Physicians from public healthcare facilities report a lack of specialised training, patient overload, and poor follow-up as operational gaps in the delivery of DM care in India [84]. The low level of public health expenditure has limited the capacity and quality of healthcare services in India's public health sector [85], highlighting the necessity to increase government health spending. We found that over 60% of inpatient and outpatient care for DM was sought at private health facilities. In India, the private health sector employs more than 80% of doctors involved in treating DM and its associated complications, which is one of the reasons of patients prefering private facilities [86]. The private sector delivers a substantial proportion of NCD-related care in India [60, 87, 88], exacerbating concerns about cost. Previous studies have documented instances of overcharging, unnecessary tests and treatments, and malpractices in the private health sector, largely because of inadequate monitoring by the government [83, 89–92]. A recent report suggests that improved regulation is a key driver of decreasing costs and improving quality of care [83].

4.1 | Strengths and limitations of the study

To the best of our knowledge, this is the first comprehensive study examining all four indicators of financial burden (CHE, impoverishment, distressed financing, and foregone care), disaggregated by the type of care sought (inpatient or outpatient) and type of healthcare facility visited (public or private), to provide a holistic picture of the financial hardships experienced by Indian households with any member suffering from DM. The strength of this study also lies in its use of the latest nationally representative survey data on health, which enhances the generalisability of the study results.

This study also involved few limitations that should be considered while interpreting the findings of the study. First, the NSSO survey collects information on self-reported ailments, which can be susceptible to under-reporting [93]. In the case of a few ailments, such as DM, a reported diagnosis by a qualified healthcare practitioner was required. However, it is estimated that over 50% of individuals with DM remain undiagnosed in India [2]; therefore, under-reporting of DM cases is a potential limitation of our study. Second, the higher burden of DM among individuals belonging to higher socioeconomic status should be interpreted with caution, because it might be attributed to either greater exposure to risk factors or better awareness and access to healthcare among such individuals. However, since the NSSO survey does not provide details about the individual-specific risk factors associated with DM (such as body mass index and tobacco use) and the exact cause of visiting the healthcare facility, it limited our ability to conduct a more nuanced analysis. Finally, the expenditure data were self-reported and may be subject to recall bias, particularly for inpatient care, where the recall period is 365 days.

5 | CONCLUSION

The current study highlights the colossal financial burden of OOPE experienced by DM-affected households in India. The burden of DM and its associated financial impact necessitate the need for multisectoral efforts, such as increasing government health expenditure, strengthening the public health sector, and improving regulatory implementation for private healthcare providers to augment financial risk protection. Policy measures to increase the uptake of health insurance and consider outpatient services under the purview of health insurance are also necessary to alleviate financial hardship. Moreover, ensuring the affordability of insulin and other essential drugs is imperative. Finally, for long-term sustainability, policymakers must prioritise health promotion and disease prevention strategies to contain the diabetes epidemic and associated financial burden.

27711757, 2023, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hcs2.65 by Delhi Technological University, Wiley Online Library on [02/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

AUTHOR CONTRIBUTIONS

Both the authors contributed to the conception and design of the study. Mehak Nanda conducted the data analysis and wrote the initial draft of the manuscript. Rajesh Sharma reviewed the manuscript and supervised the work. All authors contributed to the study and approved the submitted version.

ACKNOWLEDGMENTS

This study has received no specific grant from any funding agency in public, commercial or not-for-profit sectors.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

This study used data from the 75th round of the National Sample Survey on Health. Data are available in the public domain https://www.mospi.gov.in/.

ETHICS STATEMENT

This study is based upon secondary data, which is available in the public domain. Hence, it does not require ethics committee approval.

INFORMED CONSENT

Not applicable.

ORCID

Rajesh Sharma http://orcid.org/0000-0001-7468-1249

REFERENCES

1. World Health Organization (WHO). The global health observatory. 2019 [cited 2023 Jan 28]. Available from: https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates
2. International Diabetes Federation (IDF). Diabetes Atlas, 10th edition. 2021 [cited 2023 Jan 28]. Available from: https://diabetesatlas.org/idfawp/resource-files/2021/07/IDF_Atlas_10th_Edition_2021.pdf
3. Pandey S, Sharma V. World diabetes day 2018: battling the emerging epidemic of diabetic retinopathy. Indian J Ophthalmol. 2018;66(11):1652. https://doi.org/10.4103/ijo.IJO_1681_18
4. ICMR, PFHI, IHME. India: health of the nation's states. The India state-level disease burden initiative. 2017 [cited 2023 Jan 28] Available from: https://www.healthdata.org/sites/default/files/files/policy_report/2017/India_Health_of_the_Nation%27s_States_Report_2017.pdf
5. Tandon N, Anjana RM, Mohan V, Kaur T, Afshin A, Ong K, et al. The increasing burden of diabetes and variations among the states of India: the Global Burden of Disease Study 1990–2016. Lancet Global Health. 2018;6(12):e1352–62. https://doi.org/10.1016/S2214-109X(18)30387-5
6. Bansode B, Jungari DS. Economic burden of diabetic patients in India: a review. Diabetes Metab Syndr. 2019;13(4):2469–72. https://doi.org/10.1016/j.dsx.2019.06.020
7. Moucheraud C, Lenz C, Latkovic M, Wirtz VJ. The costs of diabetes treatment in low-and middle-income countries: a systematic review. BMJ Glob Health. 2019;4(1):e001258. https://doi.org/10.1136/bmjgh-2018-001258
8. Oberoi S, Kansra P. Economic menace of diabetes in India: a systematic review. Int J Diabetes Dev Ctries. 2020;40:464–75. https://doi.org/10.1007/s13410-020-00838-z
9. Sathyanath S, Kundapur R, Deepthi R, Poojary SN, Rai S, Modi B, et al. An economic evaluation of diabetes mellitus in India: a systematic review. Diabetes Metab Syndr. 2022;16:102641. https://doi.org/10.1016/j.dsx.2022.102641
10. Bommer C, Heesemann E, Sagalova V, Manne-Goehler J, Atun R, Bärnighausen T, et al. The global economic burden of diabetes in adults aged 20–79 years: a cost-of-illness study. Lancet Diabetes Endocrinol. 2017;5(6):423–30. https://doi.org/10.1016/S2213-8587(17)30097-9
11. IDF. India. diabetes report 2000–2045. 2021 [cited 2023 Jan 28]. Available from: https://diabetesatlas.org/data/en/country/93/in.html
12. Ministry of Health and Family Welfare, Government of India. National health policy, 2017. 2017 [cited 2023 Jan 29]. Available from: https://www.nhp.gov.in//NHPfiles/national_health_policy_2017.pdf
13. National Health Systems Resource Centre. National health accounts estimates for India (2016–17). New Delhi: Ministry of Health and Family Welfare, Government of India; 2019. https://nhsrcindia.org/sites/default/files/2021-06/FINAL%20National%20Health%20Accounts%202016-17%20Nov%202019-for%20Web%20%281%29.pdf
14. Nanda M, Sharma R, Mubarik S, Aashima A, Zhang K. Type-2 diabetes mellitus (T2DM): spatial-temporal patterns of incidence, mortality and attributable risk factors from 1990 to 2019 among 21 world regions. Endocrine. 2022;77(3):444–54. https://doi.org/10.1007/s12020-022-03125-5
15. Government of India. Key indicators of social consumption in India: Health. Ministry of Statistics & Programme Implementation, National Statistical Office; 2019. https://www.mospi.gov.in/sites/default/files/publication_reports/KI_Health_75th_Final.pdf
16. Shahrawat R, Rao KD. Insured yet vulnerable: out-of-pocket payments and India's poor. Health Policy Plan. 2012;27(3):213–21. https://doi.org/10.1093/heapol/czr029
17. Jacob AM, Devarajan A, Nachimuthu S, Datta M, Viswanathan V. Cost of diabetes treatment in private facilities for low resource urban community in South India. Int J Diabetes Dev Ctries. 2023;43:208–13.
18. Khongrangjem T, Phadnis S, Kumar S. Cost of illness (COI) of type-II diabetes mellitus in Shillong, Meghalaya. Int J Diabetes Dev Ctries. 2019;39:201–5. https://doi.org/10.1007/s13410-018-0636-z
19. Acharya LD, Rau NR, Udupa N, Rajan MS, Vijayanarayana K. Assessment of cost of illness for diabetic patients in South Indian tertiary care hospital. J Pharm Bioallied Sci. 2016;8(4):314–20. https://doi.org/10.4103/0975-7406.199336
20. Thakur A, Ray TK, Goel MK. Expenditure pattern on diabetes care: a community based longitudinal study in resettlement

27711757, 2023, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hcs2.65 by Delhi Technological University, Wiley Online Library on [02/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

colony of East Delhi. Indian J Community Health. 2017;29(2):209–12. https://doi.org/10.47203/IJCH.2017.v29i02.016
21. Kansra P. Economic burden of diabetes among women: a comparison of outpatient and inpatient care. J Health Manag. 2018;20(3):401–9. https://doi.org/10.1177/0972063418779916
22. Rao KD, Bhatnagar A, Murphy A. Socio-economic inequalities in the financing of cardiovascular & diabetes inpatient treatment in India. Indian J Med Res. 2011;133(1):57–63. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3100147/pdf/IJMR-133-57.pdf
23. Tripathy JP, Prasad BM. Cost of diabetic care in India: an inequitable picture. Diabetes Metab Syndr Obes. 2018;12(3):251–5. https://doi.org/10.1016/j.dsx.2017.11.007
24. Berki SE. A look at catastrophic medical expenses and the poor. Health Aff. 1986;5(4):138–45. https://doi.org/10.1377/hlthaff.5.4.138
25. Wagstaff A, Doorslaer E. Catastrophe and impoverishment in paying for health care: with applications to Vietnam 1993–1998. Health Econ. 2003;12(11):921–33. https://doi.org/10.1002/hec.776
26. Rajpal S, Kumar A, Joe W. Economic burden of cancer in India: evidence from cross-sectional nationally representative household survey. PLoS One. 2018;13(2):e0193320. https://doi.org/10.1371/journal.pone.0193320.
27. Yadav J, Menon GR, John D. Disease-specific out-of-pocket payments, catastrophic health expenditure and impoverishment effects in India: an analysis of National Health Survey data. Appl Health Econ Health Policy. 2021;19:769–82. https://doi.org/10.1007/s40258-021-00641-9
28. Planning Commission. Report of the expert group to review the methodology for measurement of poverty. 2014 [cited 2023 Jan 29]. Available from: https://niti.gov.in/planningcommission.gov.in/docs/reports/genrep/pov_rep0707.pdf
29. Sangar S, Dutt V, Thakur R. Distress financing of out-of-pocket health expenditure in India. Rev Dev Econ. 2019;23(1):314–30. https://doi.org/10.1111/rode.12540
30. National Commission for Scheduled Tribes. Background. 2023 [cited 2023 May 2]. Available from: https://ncst.nic.in/content/background
31. Biswas S, Pramanik KR, Sonowal CJ. Marginalised social groups differentials in nutritional status (BMI) among reproductive-aged women in West Bengal. BMC Public Health. 2023;23(1):842. https://doi.org/10.1186/s12889-023-15635-6
32. Ministry of Social Justice and Empowerment, Government of India. Department of social justice and empowerment. Welfare of the other backward classes, about the division. 2023 [cited 2023 May 2]. Available from: https://socialjustice.gov.in/common/31548
33. World Health Organization (WHO). Universal health coverage. 2021 [cited 2023 May 3]. https://www.who.int/news-room/fact-sheets/detail/universal-health-coverage-(uhc)
34. United Nations Development Programme (UNDP). Goal 3: good health and well-being. United Nations Development Programme; 2022. https://www.undp.org/sustainable-development-goals?utm_source=EN&utm_medium=GSR&utm_content=US_UNDP_PaidSearch_Brand_English&utm_campaign=CENTRAL&c_src=CENTRAL&c_src2=GSR&gclid=CjwKCAjw9e6SBhB2EiwA5myr9uiGZfMo0bJcdr_CwHw4cE6_TIMEdupc3L_816AMWbjZEY-ElPW93BoC6YYQAvD_BwE
35. World Health Organization (WHO). Global monitoring report on financial protection in health 2021. 2021 [cited 2023 May 3]. https://www.who.int/publications/i/item/9789240040953
36. World Health Organization & World Bank. Monitoring progress towards universal health coverage at country and global levels: framework, measures and targets. World Health Organization; 2014. https://apps.who.int/iris/handle/10665/112824
37. Saksena P, Hsu J, Evans DB. Financial risk protection and universal health coverage: evidence and measurement challenges. PLoS Med. 2014;1111(9):e1001701. https://doi.org/10.1371/journal.pmed.1001701
38. Hsu J, Flores G, Evans D, Mills A, Hanson K. Measuring financial protection against catastrophic health expenditures: methodological challenges for global monitoring. Int J Equity Health. 2018;17(1):69. https://doi.org/10.1186/s12939-018-0749-5
39. Rahman T, Gasbarro D, Alam K. Financial risk protection from out-of-pocket health spending in low-and middle-income countries: a scoping review of the literature. Health Res Policy Syst. 2022;20(1):83. https://doi.org/10.1186/s12961-022-00886-3
40. Petrovic D, Marcus K, Sandoval J, Cullati S, Piumatti G, Bodenmann P, et al. Health-related biological and non-biological consequences of forgoing healthcare for economic reasons. Prev Med Rep. 2021;24:101602. https://doi.org/10.1016/j.pmedr.2021.101602
41. Wagstaff A. Poverty and health sector inequalities. Bull World Health Organ. 2002;80:97–105.
42. Kibria GMA. Prevalence and factors associated with diabetes among Bangladeshi adults: an analysis of demographic and health survey 2017–18. Diabetes Epidemiol Manage. 2021;2:100012. https://doi.org/10.1016/j.deman.2021.100012
43. Seiglie JA, Marcus ME, Ebert C, Prodromidis N, Geldsetzer P, Theilmann M, et al. Diabetes prevalence and its relationship with education, wealth, and BMI in 29 low-and middle-income countries. Diabetes Care. 2020;43(4):767–75. https://doi.org/10.2337/dc19-1782
44. Singh S, Puri P, Subramanian SV. Identifying spatial variation in the burden of diabetes among women across 640 districts in India: a cross-sectional study. J Diabetes Metab Disord. 2020;19:523–33. https://doi.org/10.1007/s40200-020-00545-w
45. Sujata, Thakur R. Unequal burden of equal risk factors of diabetes between different gender in India: a cross-sectional analysis. Sci Rep. 2021;11(1):22653. https://doi.org/10.1038/s41598-021-02012-9
46. Weaver LJ, Narayan KM. Reconsidering the history of type 2 diabetes in India: emerging or re-emerging disease. Natl Med J India. 2008;21(6):288–91. Available from: https://www.researchgate.net/profile/K-M-V-Narayan/publication/26751982_Reconsidering_the_history_of_type_2_diabetes_in_India_Emerging_or_re-emerging_disease/links/02e7e520b5c5495c9d000000/Reconsidering-the-history-of-type-2-diabetes-in-India-Emerging-or-re-emerging-disease.pdf
47. Akhtar SN, Dhillon P. Prevalence of diagnosed diabetes and associated risk factors: evidence from the large-scale surveys in India. J Soc Health Diabetes. 2017;5(1):28–36.
48. Krishnamoorthy Y, Rajaa S, Verma M, Kakkar R, Kalra S. Spatial patterns and determinants of diabetes mellitus in

27711757, 2023, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hcs2.65 by Delhi Technological University, Wiley Online Library on [02/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Indian adult population: a secondary data analysis from nationally representative surveys. Diabetes Ther. 2023;14(1):63–75. https://doi.org/10.1007/s13300-022-01329-6

49. Maiti S, Akhtar S, Upadhyay AK, Mohanty SK. Socioeconomic inequality in awareness, treatment and control of diabetes among adults in India: evidence from National Family Health Survey of India (NFHS), 2019–2021. Sci Rep. 2023;13(1):2971. https://doi.org/10.1038/s41598-023-29978-y
50. Deepa M, Bhansali A, Anjana R, Pradeepa R, Joshi SR, Joshi PP, et al. Knowledge and awareness of diabetes in urban and rural India: the Indian Council of Medical Research India diabetes study (phase I): Indian Council of Medical Research India diabetes 4. Indian J Endocrinol Metab. 2014;18(3):379. https://doi.org/10.4103/2230-8210.131191
51. Joshi SR. Diabetes care in India. Ann Glob Health. 2018;81(6):830–38.65. https://doi.org/10.1016/j.aogh.2016.01.002
52. Ramachandran A, Ramachandran S, Snehalatha C, Augustine C, Murugesan N, Viswanathan V, et al. Increasing expenditure on health care incurred by diabetic subjects in a developing country. Diabetes Care. 2007;30(2):252–6. https://doi.org/10.2337/dc06-0144
53. Kumar D, Mukherjee K. Economic impact of type-2 diabetes mellitus on households in Hisar district of Haryana state, India. 2014 [cited 2023 Feb 3]. Available from: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=553df706e059ea25dfa7d247b858192846ef30a1
54. Rohilla L, Gujjar N, Kaur G, Walia P, Dayal D. Financial burden for families of children with type 1 diabetes: a cross-sectional survey from North India. Diabetol Int. 2022;13(4):665–71. https://doi.org/10.1007/s13340-022-00589-8
55. Shariful Islam SM, Lechner A, Ferrari U, Laxy M, Seissler J, Brown J, et al. Healthcare use and expenditure for diabetes in Bangladesh. BMJ Glob Health. 2017;2(1):e000033. https://doi.org/10.1136/bmjgh-2016-000033
56. Mendenhall E, Shivashankar R, Tandon N, Ali MK, Narayan KM, Prabhakaran D. Stress and diabetes in socio-economic context: a qualitative study of urban Indians. Soc Sci Med. 2012;75(12):2522–9. https://doi.org/10.1016/j.socscimed.2012.09.040
57. Bhojani U, Mishra A, Amruthavalli S, Devadasan N, Kolsteren P, De Henauw S, et al. Constraints faced by urban poor in managing diabetes care: patients' perspectives from South India. Glob Health Action. 2013;6(1):22258. https://doi.org/10.3402/gha.v6i0.22258
58. Little M, Humphries S, Patel K, Dewey C. Decoding the type 2 diabetes epidemic in rural India. Med Anthropol. 2017;36(2):96–110. https://doi.org/10.1080/01459740.2016.1231676
59. Government of India. Rural health statistics 2020–2021. New Delhi: Government of India, Ministry of Health and Family Welfare, Statistics Division; 2021. https://main.mohfw.gov.in/sites/default/files/rhs20-21_1.pdf
60. Institute for Health Metrics and Evaluation (IHME) and Public Health Foundation of India (PHFI). Health service provision in Odisha: assessing facility capacity, costs of care, and patient perspectives. Seattle, WA: IHME and PHFI; 2018. https://www.healthdata.org/policy-report/health-service-provision-odisha-assessing-facility-capacity-costs-care-and-patient
61. Shaw S, Sahoo H. Accessibility to primary health centre in a tribal district of Gujarat, India: application of two step floating catchment area model. GeoJournal. 2020;85:505–14. https://doi.org/10.1007/s10708-019-09977-1
62. Mohan V, Deepa M, Pradeepa R, Prathiba V, Datta M, Ravikumar S, et al. Prevention of diabetes in rural India with a telemedicine intervention. J Diabetes Sci Technol. 2012;6(6):1355–64. https://doi.org/10.1177/193229681200600614
63. Olickal J, Devasia JT, Thekkur P, Chinnakali P, Suryanarayana B, Kumar Saya G, et al. How far persons with diabetes travel for care? Spatial analysis from a tertiary care facility in Southern India. Ann GIS. 2021;27(4):341–9. https://doi.org/10.1080/19475683.2021.1966504
64. National Health Portal. Ayushman Bharat—Health and Wellness Centre. Ministry of Health and Family Welfare, Government of India; 2022. https://ab-hwc.nhp.gov.in/
65. Kesavadev J, Saboo B, Shankar A, Krishnan G, Jothydev S. Telemedicine for diabetes care: an Indian perspective-feasibility and efficacy. Indian J Endocrinol Metab. 2015;19(6):764. https://doi.org/10.4103/2230-8210.167560
66. Mohan V, Prathiba V, Pradeepa R. Tele-diabetology to screen for diabetes and associated complications in rural India: the Chunampet Rural Diabetes Prevention Project Model. J Diabetes Sci Technol. 2014;8(2):256–61. https://doi.org/10.1177/1932296814525029
67. Nanda M, Sharma R. A review of patient satisfaction and experience with telemedicine: a virtual solution during and beyond COVID-19 pandemic. Telemed E-Health. 2021;27(12):1325–31. https://doi.org/10.1089/tmj.2020.0570
68. Berman P, Ahuja R, Bhandari L. The impoverishing effect of healthcare payments in India: new methodology and findings. Econ Polit Wkly. 2010;45(16):65–71.
69. Mukherjee S, Chaudhuri A. Outpatient care and expenses: can they be ignored in health insurance programmes. 2020 [cited 2023 Feb 5]. Available from: http://idsk.edu.in/wp-content/uploads/2020/06/OP-68.pdf
70. Selvaraj S, Farooqui HH, Karan A. Quantifying the financial burden of households' out-of-pocket payments on medicines in India: a repeated cross-sectional analysis of National Sample Survey data, 1994–2014. BMJ Open. 2018;8(5):e018020. https://doi.org/10.1136/bmjopen-2017-018020
71. Selvaraj S, Karan AK. Why publicly-financed health insurance schemes are ineffective in providing financial risk protection. Econ Polit Wkly. 2012;47(11):60–8.
72. Hooda SK. Penetration and coverage of government-funded health insurance schemes in India. Clin Epidemiology Glob Health. 2020;8(4):1017–33. https://doi.org/10.1016/j.cegh.2020.03.014
73. National Health Authority. Ayushman Bharat—Pradhan Mantri Jan Arogya Yojana. 2022 [cited 2023 Mar 1]. Available from: https://pmjay.qcin.org/pm-jay
74. Grover S, Avasthi A, Bhansali A, Chakrabarti S, Kulhara P. Cost of ambulatory care of diabetes mellitus: a study from north India. Postgrad Med J. 2005;81(956):391–5. https://doi.org/10.1136/pgmj.2004.024299
75. Maiti R, Bhatia V, Padhy BM, Hota D. Essential medicines: an Indian perspective. Indian J Community Med. 2015;40(4):223–32. https://doi.org/10.4103/0970-0218.164382

27711757, 2023, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hcs2.65 by Delhi Technological University, Wiley Online Library on [02/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

76. Satheesh G, Sharma A, Puthean S, Ansil TP M, E J, Raj Mishra S, et al. Availability, price and affordability of essential medicines for managing cardiovascular diseases and diabetes: a statewide survey in Kerala, India. Trop Med Int Health. 2020;25(12):1467–79. https://doi.org/10.1111/tmi.13494
77. Satheesh G, Unnikrishnan MK, Sharma A. Challenges constraining availability and affordability of insulin in Bengaluru region (Karnataka, India): evidence from a mixed-methods study. J Pharm Policy Pract. 2019;12:31. https://doi.org/10.1186/s40545-019-0190-1
78. Chow CK, Ramasundarahettige C, Hu W, AlHabib KF, Avezum A Jr., Cheng X, et al. Availability and affordability of essential medicines for diabetes across high-income, middle-income, and low-income countries: a prospective epidemiological study. Lancet Diabetes Endocrinol. 2018;6(10):798–808. https://doi.org/10.1016/S2213-8587(18)30233-X
79. Sharma A, Kaplan WA. Challenges constraining access to insulin in the private-sector market of Delhi, India. BMJ Glob Health. 2016;1(3):000112. https://doi.org/10.1136/bmjgh-2016-000112
80. Jayaraman KS. India's diabetics have poor access to insulin, says study. Nature India. 2016. Available from: https://www.nature.com/articles/nindia.2016.134#:~:text=Access%20to%20life%2Dsaving%20insulin,says%20a%20new%20study1
81. Government of India (GOI). Quality generic medicines at affordable prices. Pharmaceuticals & Medical Devices Bureau of India, Department of Pharmaceuticals, Government of India; 2022. http://janaushadhi.gov.in/index.aspx
82. Chaudhary RK, Philip MJ, Santhosh A, Karoli SS, Bhandari R, Ganachari MS. Health economics and effectiveness analysis of generic anti-diabetic medication from jan aushadhi: an ambispective study in community pharmacy. Diabetes Metab Syndr. 2021;15(6):102303. https://doi.org/10.1016/j.dsx.2021.102303
83. Selvaraj S, Karan KA, Srivastava S, Bhan N, Mukhopadhyay I. India health system review. New Delhi: World Health Organization, Regional Office for South-East Asia; 2022. Available from: https://apo.who.int/publications/i/item/india-health-system-review
84. Tripathy JP, Sagili KD, Kathirvel S, Trivedi A, Nagaraja SB, Bera OP, et al. Diabetes care in public health facilities in India: a situational analysis using a mixed methods approach. Diabetes Metab Syndr Obes. 2019;12:1189–99. https://doi.org/10.2147/DMSO.S192336
85. Kumar A, Sarwal R. Health insurance for India's missing middle. 2021 [cited 2023 Jan 4]. Available from: https://www.niti.gov.in/sites/default/files/2021-10/HealthInsuranceforIndiasMissingMiddle_28-10-2021.pdf
86. Das AK, Saboo B, Maheswari A, Banerjee S, Jayakumar C, Mohan AR, et al. Health care delivery model in India with relevance to diabetes care. Heliyon. 2022;8(10):e10904. https://doi.org/10.1016/j.heliyon.2022.e10904
87. Patel V, Chatterji S, Chisholm D, Ebrahim S, Gopalakrishna G, Mathers C, et al. Chronic diseases and injuries in India. Lancet. 2011;377(9763):413–28. https://doi.org/10.1016/S0140-6736(10)61188-9
88. Institute for Health Metrics and Evaluation (IHME) and Public Health Foundation of India (PHFI). Health service provision in Andhra Pradesh and Telangana: assessing facility capacity, costs of care, and patient perspectives. Seattle, WA: IHME and PHFI; 2018. Available from: https://www.healthdata.org/research-analysis/library/health-service-provision-andhra-pradesh-and-telangana-assessing-facility
89. Phadke A. Regulation of doctors and private hospitals in India. Econ Polit Wkly. 2016;51(6):46–55.
90. Dehury RK, Samal J, Coutinho S, Dehury P. How does the largely unregulated private health sector impact the Indian mass? J Health Manag. 2019;21(3):383–93. https://doi.org/10.1177/0972063419868561
91. The Times of India. Greed and the rise of medical malpractice in India. 2016 [cited 2023 Mar 3]. Available from: https://timesofindia.indiatimes.com/blogs/the-irreverent-lawyer/greed-and-the-rise-of-medical-malpractice-in-india/
92. Chakravarthi I. Regulation of private health care providers in India: current status, future directions. Indian J Public Admin. 2018;64(4):587–98. https://doi.org/10.1177/0019556118780093
93. Bhise MD, Patra S, Chaudhary M. Geographical variation in prevalence of non-communicable diseases (NCDs) and its correlates in India: evidence from recent NSSO survey. J Public Health. 2018;26(5):559–67. https://doi.org/10.1007/s10389-017-0889-x

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Nanda M, Sharma R. Financial burden of seeking diabetes mellitus care in India: evidence from a Nationally Representative Sample Survey. Health Care Sci. 2023;2:291–305. https://doi.org/10.1002/hcs2.65

27511757, 2023, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hcs2.65 by Delhi Technological University, Wiley Online Library on [02/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

MAPAN-Journal of Metrology Society of India
https://doi.org/10.1007/s12647-023-00690-6

ORIGINAL PAPER

Finishing and Wear analysis of 3D-Printed Workpiece Through Centrifugal Force-Assisted Abrasive Flow Machining

A. Bhardwaj[1] and P. Ali[2]*

[1]Mechanical Engineering Department, Delhi Technological University, Delhi 110042, India

[2]Mechanical Engineering Department, Gautam Buddha University, Greater Noida, U.P 201312, India

Received: 17 May 2023 / Accepted: 13 September 2023

© Metrology Society of India 2023

Abstract: The rapid population growth has raised the need of industrialization which has opened a new horizon for the additive manufacturing, especially in the area of 3D printing technology. 3D printing has immersed as the additive manufacturing technique which finds its application in every domain including toys, confectionaries, biomedical, electronics, aerospace and automobile. But achieving the desirable surface finish is a difficult task for 3D-printed material especially poly lactic acid (PLA). Due to less material removal in traditional abrasive flow machining, the scope of the current investigation is to finish 3D-printed hollow cylindrical workpiece made of PLA material by Fused deposition method using Centrifugal force-assisted abrasive flow Machining process. The investigation was carried out to determine the optimum percentage improvement in surface finish and surface micro-hardness using Response surface methodology. Using an optical profilometer and a micro-hardness tester, the surface characteristics of the workpiece were examined. Additionally, wear modeling was carried out using the Ansys® software, and the outcomes were confirmed through tests using Reye–Archard–Khrushchev wear law. The simulation results were in good confirmation of experimental results with total debris of 1.4 mg. The findings of the investigation indicate an average improvement in surface quality of 39.27%. The experimental results show that the major influence on enhanced R_a and micro-hardness comes from CFG rod rotation. Also, the ideal percentage improvements in surface finish and micro-hardness are found to be 35.34% and 41.29 HV, respectively.

Keywords: Abrasive flow machining; Centrifugal force-assisted abrasive flow machining; Three-dimensional printing; Polylactic acid; Wear; Modeling

1. Introduction

3D printing has emerged as a modern tool in the additive manufacturing scenario where the requisite job is achieved by layer-by-layer addition of material over each other. The process has many applications in the field of aerospace, toys, automobile, cosmetics and biomedical. The process is putting its foot in the area of constructions where the new 3D-printed houses and bridges are quite common. The process finds its applicability in the toy manufacturing where intrinsic shapes may be manufacture with the real ease. Considering wide application of the process, it is important to analyze the mechanical property of 3D-printed material and analyze its applicability. Though 3D-printed objects are popular, they are incomplete until the aesthetic value of the product is not complete which comes from surface finishing. Surface finish is a major challenge in 3D-printed products and that become more challenging when applied to the internal surface of the difficult to reach intrinsic part. The present investigation deals with the above challenges. Over the period of time, wide range of research has been done on the 3D printing pertaining to its improvement in process efficiency and applicability. Petrovskaya et al. [1] synthesized polylactide hydroxyapatite-based material for the bone implants using 3D printing technique and commented on the successful implementation in the field of biomedical and bone implants. The material PLA (Polylactic Acid) finds its application in the synthesis of composite materials. Many researchers are putting their efforts in the composites synthesizing. Chaitanaya et al. [2] fabricated PLA-based

*Corresponding author, E-mail: parvesh1133@gmail.com

Published online: 18 October 2023

bio-composite using extrusion and direct injection molding, whereas Rimašauskas et al. [3] used 3D printing for the processing of carbon fiber for continuous composite structure. Liu et al. [4] developed a PLA based hydrophobic calcium oxide composite. The researchers blended the CaO with the PLA and observed the hydrophobic property of the composites. Apart from composites, many researchers directed their research toward the computation of the mechanical property of the PLA materials. Harris et al. [5] studied the mechanical property of PLA component made by FDM. The researchers performed set of experiments and improved the property of joint formed by the beads of PLA. The researchers gave an additive free solution for 3D printing using FDM material. Miroslav et al. [6] improved the property of PLA material by doing annealing up to 100 °C by 90 min and also by adding wood flour and kaoline to give mechanical strength and highlighted its positive application. The recent challenge in the 3D-printed material is in the area of machining. Many researchers have worked in the field of finishing as the aesthetic value of the product and its functional validity increases with it [7–12]. Kechagias et al. [13] used the analysis of means and ANOVA to compute the surface roughness. Through the computation of 3D-printed PLA plates cut by laser, it was observed that the velocity and power were important parameters for the computation of the surface roughness. Kuo et al. [14] developed an acetone-vapor polishing method to finish the acrylonitrile butadiene styrene (ABS) parts manufactured through fused deposition methods (FDM), which resulted the improvement in surface finish. Kumar et al. [15] employed the method of ball end magnetorheological finishing process to produce the nano-finished component developed through FDM technique. Some of the studies are based on fabrication parameters for the 3D printing jotted the successful applicability of the process in modern industrial application [16]. Many researchers contributed in optimization of process parameter for FFF process and jotted its successful application for printing of various materials [17–20]. It was found that AFM is prominent process for finishing of internal surfaces [21].

Fused filament fabrication (FFF) is one of the most popular 3D printing techniques that may be used to create a variety of parts and is compatible with strong, production-ready materials. Despite the fact that FFF 3D-printed parts offer a number of appealing qualities, they frequently have poor surface morphology, which is a crucial factor in printing high-quality conducting traces for electronics devices. The literature review also reveals the importance of 3D printing with the PLA materials and also highlights the limitation of low surface finish. The shear yield strength depends heavily on the texture of the workpiece. As the hardness of the surface increases, so does the strength of the material. Thus, to cater this issue, CFAAFM is used for improving the surface roughness of 3D-printed FDM part to form a smooth surface for accurate conductive trace formation.

2. Materials and Methods

2.1. Principle of Centrifugal Force-Assisted Abrasive Flow Machining (CFAAFM) Process

AFM is an nontraditional process which is capable of achieving nano-level finishing, as the abrasive laden media is pressurized to pass through an restrictive passage of the workpiece. CFAAFM is the hybrid of conventional AFM process. This process uses the advantage of centrifugal force produced by the rotation of rod placed coaxially along the axis of hollow workpiece whose internal surface is to finish, along with the conventional reciprocation of media due to the difference in extrusion pressure [22]. The mechanism of the CFAAFM process is depicted in Fig. 1. The CFG rod develops a centrifugal force in the media flow path which increases the active abrasive particles near the walls. This causes more amount of material removal. Figure 2 shows the flowchart for the stages used in the present investigation for analysis.

2.2. Fused Deposition Modeling

The most well-liked 3D printing process now is undoubtedly Fused Deposition Modeling (FDM). Moreover, it is known as "fused filament fabrication" (FFF). Many people associate this technology with 3D printing. FDM is a practical, affordable, and suitable option for many applications. The most often used material in FDM 3D printing is PLA (Polylactic Acid). Organic substance PLA enables

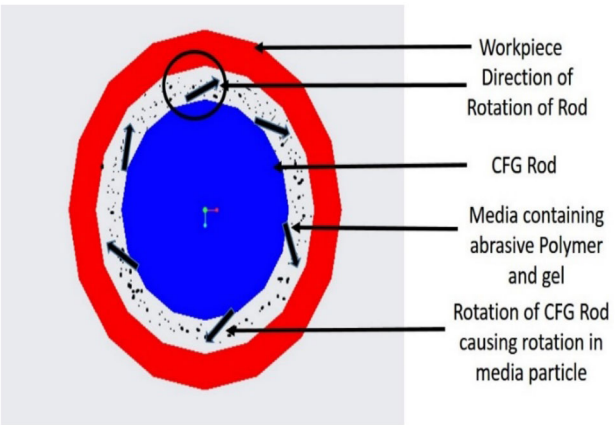


Fig. 1 Mechanism of CFAAFM process

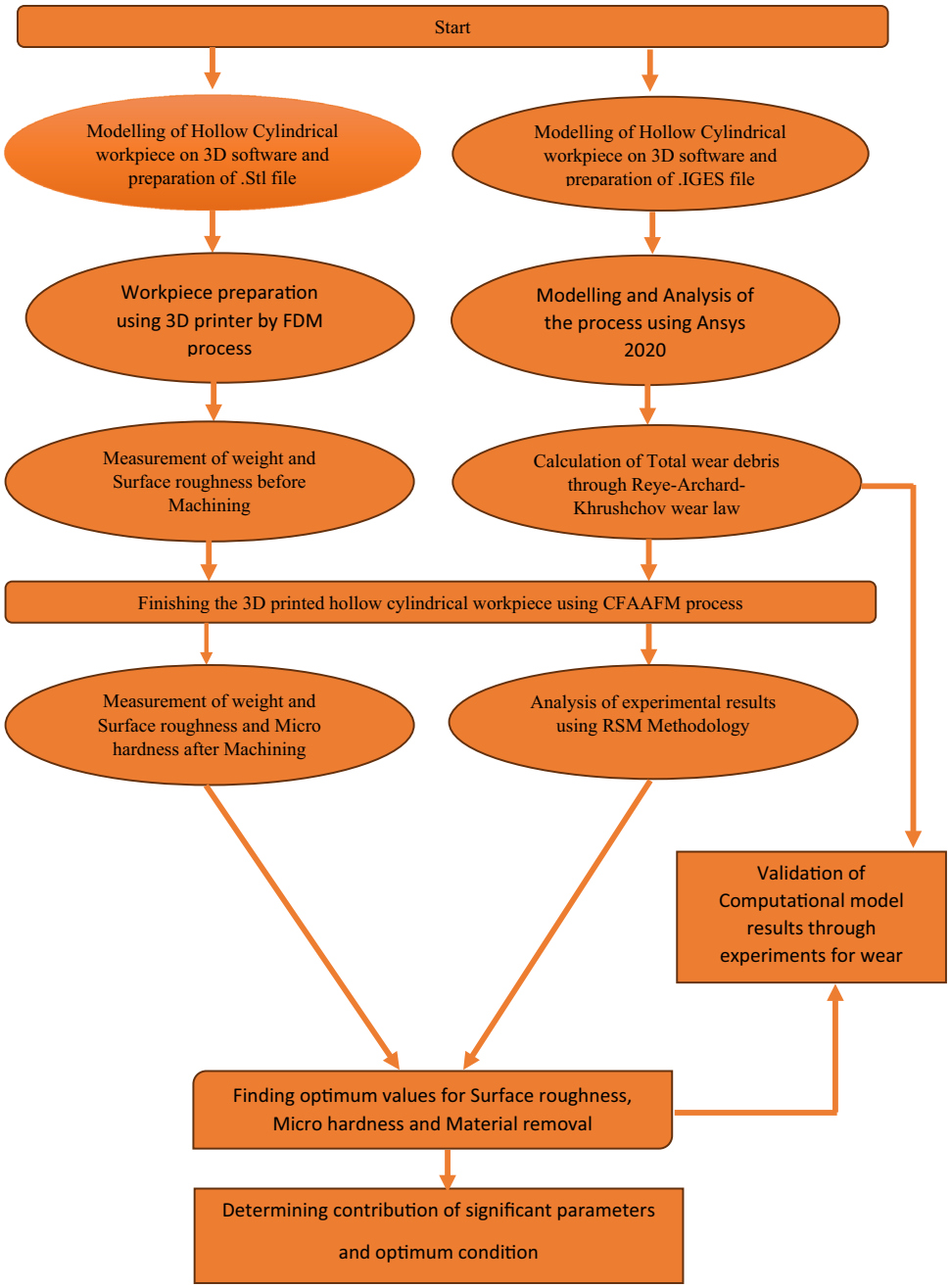


Fig. 2 Flowchart of the process

more accurate printing. A polymer known as PLA is created using organic substances like maize starch. PLA material becomes harder and little bit more liquid while printing. As a result, print looks more detailed and shinier. In the present invention, PLA workpiece was made through Shaper jet 3D printing whose description is given in Table 1. Melting a string of plastic and layer-by-layer putting it to a construction platform is how the FDM process operates. As the material is completely melted, 3D-printed components with an adequate thickness resemble injection moulded ones mechanically.

The interior of solid models is typically filled with a stabilizing space-fame rather than being printed solid because of the somewhat lengthy printing process. The Infill Rate—a measure of how much of the volume is actually filled with solid material—defines that area. This is usually set at 20 to 25% of the interior. The heated plastic is poured via a circular nozzle during FDM. Little grooves are thus formed between the layers, creating a rough surface. Since the layers stick to one another better in PLA, the consistency of the groove size is improved. The nozzle size limits the level of detail while using FDM. Smaller nozzles enable more detail, but they significantly

Table 1 Description of 3D printer

Sr. no.	Property	Description
1	Company Name	Shaper Jet 3D printer
2	Printing Technique	Fused deposition modeling (FDM)
3	Maximum Print Volume	180 × 200 × 160
4	Layer Thickness	0.06 to 0.32 mm
5	Connectivity	USB, SDHC card
6	Compatible Materials	Poly Lactic Acid (PLA)
7	File Format	.Stl,.amf
8	Operating System	Windows, MAC, Linux
9	Printing Speed	200 mm/s
10	Maximum temperature	290 °C

lengthen the printing process. To achieve the optimal balance between detail and print time, nozzles with diameters between 300 and 500 µm are typically employed. The typical layer height is between 100 and 200 µm, while 50 µm is only occasionally possible. The dimensions of the workpiece were taken as 10 mm OD, 8 mm ID and the height of 16 mm. Figure 3 shows the setup used for the preparation of the workpiece. Figure 3a, b shows the 3D printer whose specification is listed in Table 1. Figure 3c, d shows the preparation of the workpiece. Figure 4 shows the PLA workpiece. Table 2 highlights the property of PLA material used for the manufacturing of the workpiece.

2.3. Experimental Setup

The centrifugal force generating rod (CFG rod), which rotates through an induction motor, is used in the experimental setup for the CFAAFM method. The experimental setup for CFAAFM process is shown in Fig. 5. The experimental equipment included a nylon fixture, bearings, a 3-phase induction motor, a power drive to regulate the rotation of the CFG rod, and gear arrangements for power transmission. The fixture held the workpiece and guided the media between the two media cylinders. The driving motor in the current experimental setup was a three-phase induction motor controlled with power drive that was coupled to the motor shaft using a mild steel gear. This mild steel gear was connected with an aluminum gear at the intermediate position, which was further connected with two aluminum gears for CFG rod rotation. The fixture was a combination of three parts in which the lower part of fixture consists the CFG rod coupled with bearing and gears. The other end of the CFG rod was fitted into a slot made in the upper part of fixture. This slot was made to support the end of the CFG rod and to guide its rotation. The middle part and upper part of fixture had slot to support the workpiece. The CFG rod is rotated symmetrically

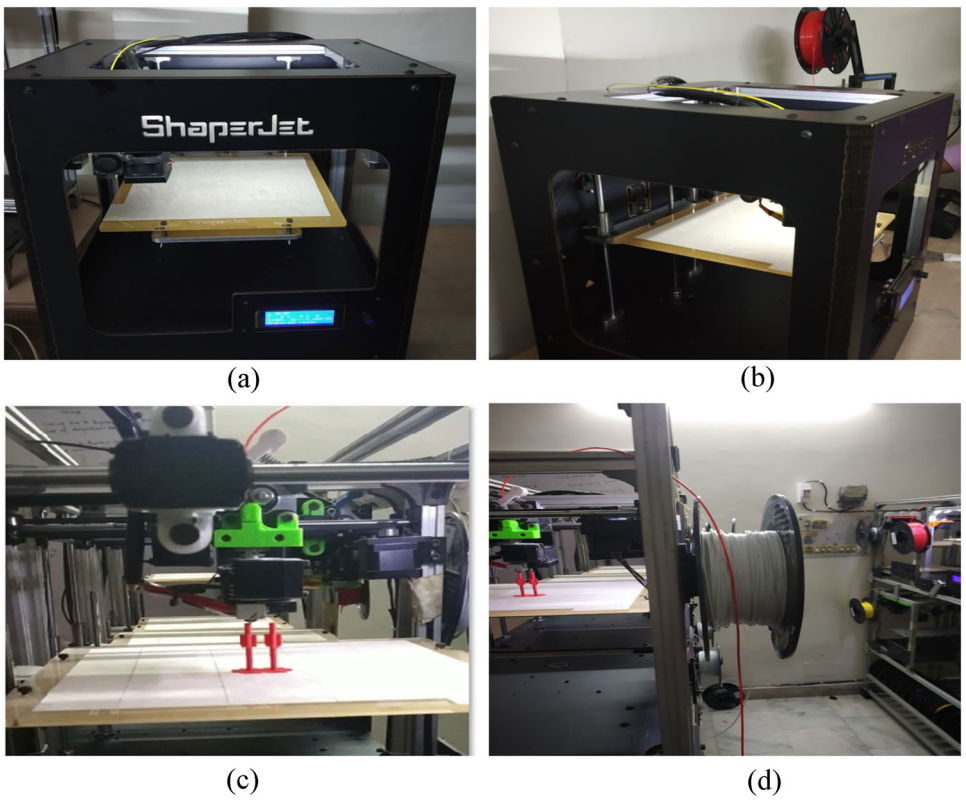


Fig. 3 Experimental setup for fabrication of workpiece

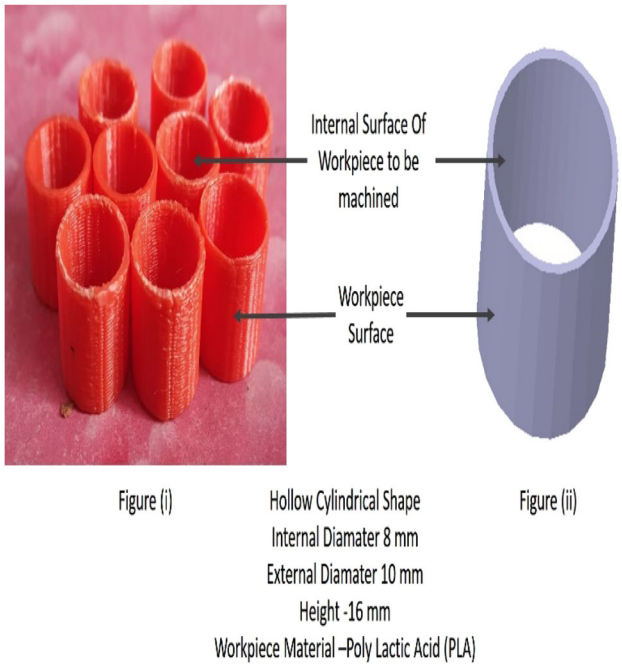


Fig. 4 Workpiece

inside the workpiece for the uniform distribution of centrifugal force. This enhanced the dynamic abrasive particles and increased the material removal. Figure 5a shows the conventional AFM setup with 3 phase induction motor. Figure 5b shows the side view of the experimental set up with the mechanism of power transmission. Figure 5c, d shows the fixture containing the workpiece, arrangement for CFG rod rotation, and Fig. 5e shows the power drive for controlling the CFG rod rotation. The functions and specifications of the components used in CFAAFM process are mentioned in Table 3, which is as follows:

2.4. Calculation of Total Wear Debris

To understand the impact of crucial parameters (temperature variation, pressure, media velocity) over the surface of work piece, CFAAFM process modeling was conducted. This will increase the process's machining efficiency. Reye–Archard–Khrushchev wear law was used in conjunction with this analysis to further anticipate wear. The law provides the following equation [23]:

$$Q = \frac{K.W.L}{H}$$

where Q = Wear debris aggregate volume, K = Constant, W = Cumulative Radial force, L = Sliding distance, H = Hardness of the surfaces in contact, whichever is softest.

The wear obtained through the computational model analysis was contrasted with the outcomes of the experiments.

Considering that asperities have a circular shape,

the asperity's area $= (\pi).\, a^2$

(a defines the radius of asperity).

Cumulative Radial Force

$$\mathrm{W} = (P).(\pi).a^2.N_\mathrm{a}$$

P represents the pressure employed over the work piece, N_a = No. of abrasive particles taking participation in the machining process.

Sliding length value for a single asperity $(L) = 2a$

Total no. of abrasive particles used in the abrasive laden media (N), $[24] = \dfrac{V_\mathrm{a}}{V_\mathrm{o}}$

where V_a represents total volume of abrasive particles, V_o = Volume of a single abrasive

$$= \frac{4}{3}.(\pi).\left(\frac{d}{2}\right)^3$$

$$V_\mathrm{a} = \frac{C.V_\mathrm{m}.d_\mathrm{m}}{d_\mathrm{a}}$$

where C = abrasive concentration in abrasive laden media,

Table 2 Property of PLA material

Sr.no	Physical property	Units	Numerical value
1	ρ	g/cm^3	1.24 g/cm^3
2	S_{yt}	MPa	50 MPa
3	σ	MPa	80 MPa
4	KC (Un-notched) IZOD	J/m	96.1
5	Shrink rate	in/in	0.0037–0.0041 in/in
6	Heat Deflection Temperature	°F	126°F (52 °C)
7	Thermal Conductivity	Wm/K	0.16 Wm/K
8	Specific heat	J/kg/K	1200
9	Thermal Diffusivity	mm^2/s	0.110

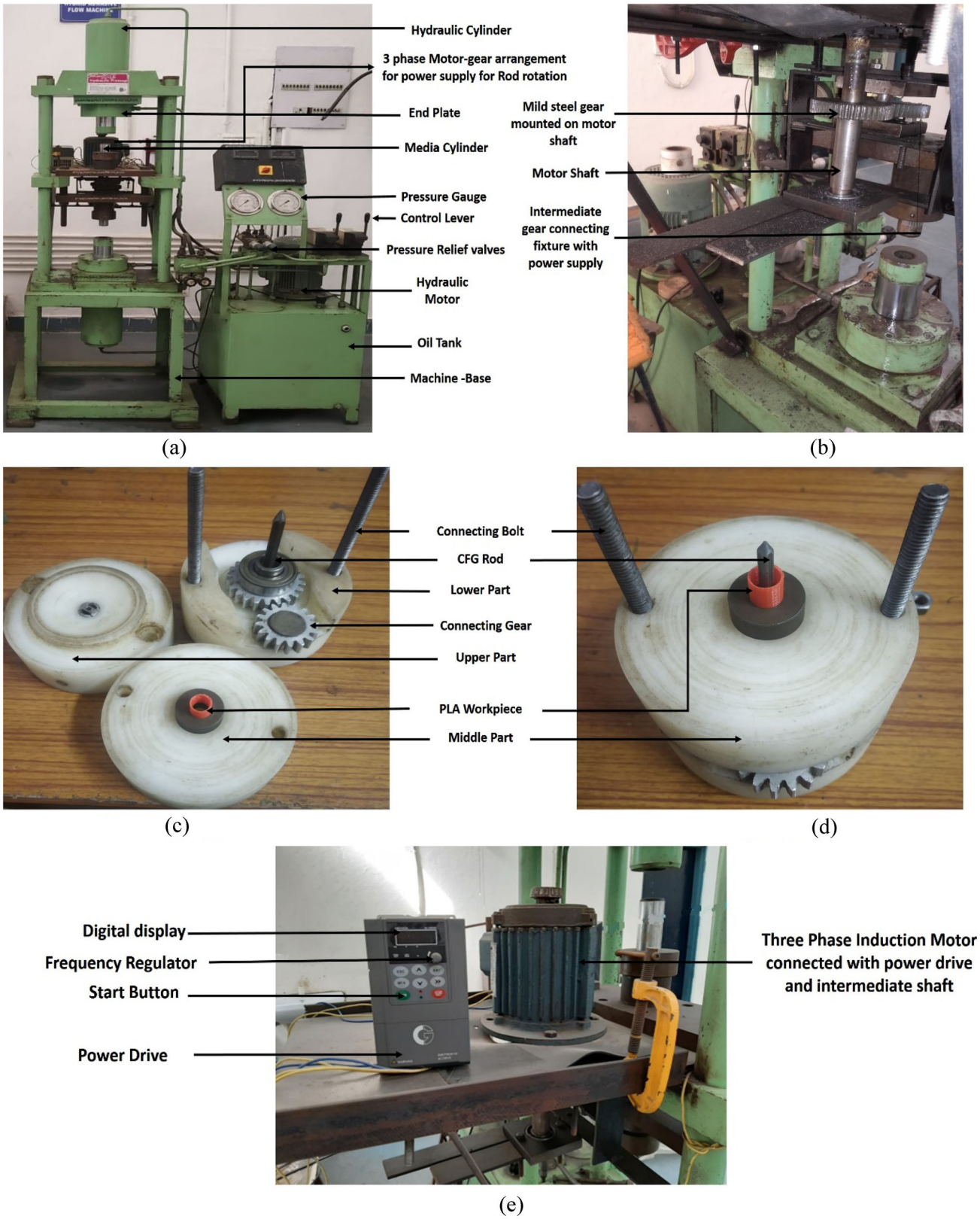


Fig. 5 Experimental setup **a** CFAAFM setup **b** power transmission mechanism **c**, **d** fixture containing workpiece **e** three phase induction motor attached with power drive

Table 3 Details of the components used in CFAAFM process

Sr.no	Components	Description
1	Hydraulic Power pack	Maximum Extrusion pressure = 210 kg/cm^2
		Enables the piston in the hydraulic cylinder to move back and forth
		It comprises of a hydraulic pump, reservoir, filter, and motor in addition to an accompanying hydraulic circuit
2	Hydraulic Cylinder	Diameter of Bore = 130 mm
		Stroke length of Hydraulic Cylinder = 96 mm
		Responsible for the movement of piston for creating pressure difference
3	Conventional AFM Setup	Media Flow Volume = 290 cm^3
4	Power drive	Power = 0.75 KW, 1 HP
		Input = AC Single Phase, Voltage (V): 200–240
		Used to change rotation speed of 3 Phase induction motor by frequency regulation
5	Three-Phase induction motor	Power = 1 HP,
		Rotational Speed = 800 rpm
		Used for rotating the CFG Rod

d_m = media density, V_m = media volume, d_a = abrasive density

Total active number of abrasive particles (Na),

$$[24] = \frac{4.V_a(D.d - d^2)}{V_o.D^2}$$

where D represents workpiece diameter

Fraction of abrasive particle staking participation in material removal $= \frac{N_a}{N}$

2.5. Modeling of CFAAFM Process

The initial solid modeling of the process was done on Creo 3.0, and further analysis of the model was performed on Ansys® 2020. The initial geometry was modified on space claim, and the meshing was done with proper sizing ($6e^{-3}$ m) and skewness with proper edge preparation.

The properties considered for the model analysis are shown in Table 4.

Figure 6 shows the model components of CFAAFM, and Table 5 shows the meshing details of various components used in the model. Figure 6a shows the top view of entire assembly consisting of workpiece, rod and fluid (mixture of abrasive, polymer and gel). Figure 6b shows the side view of the model, Fig. 6c shows the meshing of the whole model, and Fig. 6d shows the meshing of CFG rod.

The assumptions for the computational model analysis are stated as follows:

- Flow is taken as laminar and steady in nature
- Flowing media behaves likes homogeneous and isotropic
- Media flows in a symmetric path inside the three-dimensional hollow surface
- Analysis was pressure and Green Gauss cell based.
- The surface of the work piece and the abrasive particle slide past one another.
- A circular form for the asperities is assumed.
- The asperities will entirely let through abrasive particles.
- Hemispherical-shaped wear takes place.
- Every asperity will not take part in the removal process.

The laminar model with energy equation is consider for the analysis of flow parameters. At the inlet, 40 MPa of extrusion pressure was taken and at the outlet it was taken as 20 MPa to create a pressure difference for media flow. The workpiece was kept stationary, and the rod was rotated with 100 rpm speed. The standard solution parameters are kept, and the solutions are run for some iterations. The results obtained were grid tested and independent of the mesh size. Table 6 shows the boundary condition description. Figure 7 shows the boundary conditions of the computational model.

2.6. Optimization of CFAAFM Process

All of the points in the current design have the same standard error and are situated at equal distances from the region's center. The region center is represented by the point (0, 0, ….,0) for which X and Y relationship was studied. The experimental data can be used to define the errors of Y at any location over the surface. Due to the requirement of rotatability, the standard error will be the same for every point located at a distance of m from the region's center. Draper [25] has given an equation for this condition which is as follows:

Table 4 Properties of the materials

Component	Material	Properties
CFG Rod	Cu	C_p = 376 J/kg-K
		ρ = 8939 kg/m^3
		K = 384 W/m-K
Work piece	PLA material	K = 108 W/m-K
		C_p = 400 J/kg-K
		ρ = 1.24 gm/cm^3
		μ = 0.31
Abrasive Particles	Alumina	K = 11 W/m-K
		ρ = 3949 kg/m^3
		C_p = 451 J/kg-K
Type of Media	Polyborosiloxane	ρ = 985 kg/m^3
		C_p = 2025 J/kg-K
		K = 0.22 W/m-K

$$x_1^2 + x_2^2 + x_3^2 + \ldots + x_k^2 = m^2 = \text{constant}$$

Percentage improvements in R_a and micro-hardness have been taken into consideration in the current experimental study to assess finishing performance. The CFAAFM configuration, depicted in Fig. 5, was used for the tests. Table 7 lists the variables that were chosen, along with their upper and lower bounds: CFG rod rotation, cycle count, abrasive concentration, extrusion pressure, and abrasive particle size. PLA material was utilized to create the test object. Thirty-two trials were conducted in accordance with the central composite design. Optical surface profilometer (MicroXam-100) was used to measure surface roughness. Surface roughness value was observed at two different positions of the work piece, and mean of those was considered as the final roughness.

$$\% \text{ improvement in surface roughness} = [(R_a \text{ value before finishing} - R_a \text{ value after finishing})/(R_a \text{ value before finishing})] * 100$$

Micro-hardness of the workpiece was measured by (DURAMIN-40 M1) hardness tester. Micro-hardness of the surface was taken at three places, and average of these value were considered as the final micro-hardness of the surface.

3. Results and Discussion

3.1. Analysis of Computational Model

The final analysis of model was carried out using the Fluent® programme. The analysis was done to determine the crucial circumstances for the chosen parameters (Pressure, Media velocity and temperature distribution). Geometrically, the effects of the analysis were applied to the planes (0,0,0) and (1,0,0). The impact on the media's velocity along the Z axis is depicted in Fig. 8. The graph demonstrates that the media was passing and continuously slowing down due to frictional losses as indicated in the literature [26].

From Fig. 9, it is obvious that the media flow velocity was at its highest in the center where the abrasive particles impacted the work piece surface [27]. The media's velocity increased as the abrasive in it made contact with the workpiece because of the energy gained from the impact [27]. The figure also demonstrates that media velocity was decreasing close to the wall [28]. Significant friction forces are applied during the media flow. The media must move by overcoming the frictional losses, which has been shown by continuously loss of energy [29]. The results also reveal that the media temperature would rise, as previously described [30], when the media began to flow from one to another media cylinder, as seen in Fig. 10. The impact of media temperature on the effectiveness of the procedure is quite interpretive. Hot spots can be seen all over the surface in Fig. 10. Due to the increase in media temperature, these hot spots over the surface were seen. The frictional resistance provided by the media particle and exterior boundary caused this rise in media temperature. With the number of cycles, the temperature of the medium steadily rises [31].

This could be the cause of the hot spots that are visible on the surface. It raises the media temperature, which reduces the media viscosity and results in a poorer capacity to contain abrasives. When the temperature of the media is increased further, the wear rate of the abrasive particles actually decreases at first. The fluctuation of the media flow pressure in the setup for CFAAFM is shown in Fig. 11 from intake to exit. Static and dynamic pressure are added together to provide the overall media pressure during the flow. The total media pressure during the flow is summation of static pressure as well as dynamic pressure. The reason for the application of static pressure is the arbitrary movement of media molecules in the fluid, while due to movement of fluid, dynamic pressure is exerted. In the traditional AFM method, abrasive-laden media passes through the workpiece by the virtue of pressure difference. Figure 11 demonstrates how frictional losses cause the pressure to continuously decline when the media starts flowing from the higher pressurized zone to the lower pressurized zone.

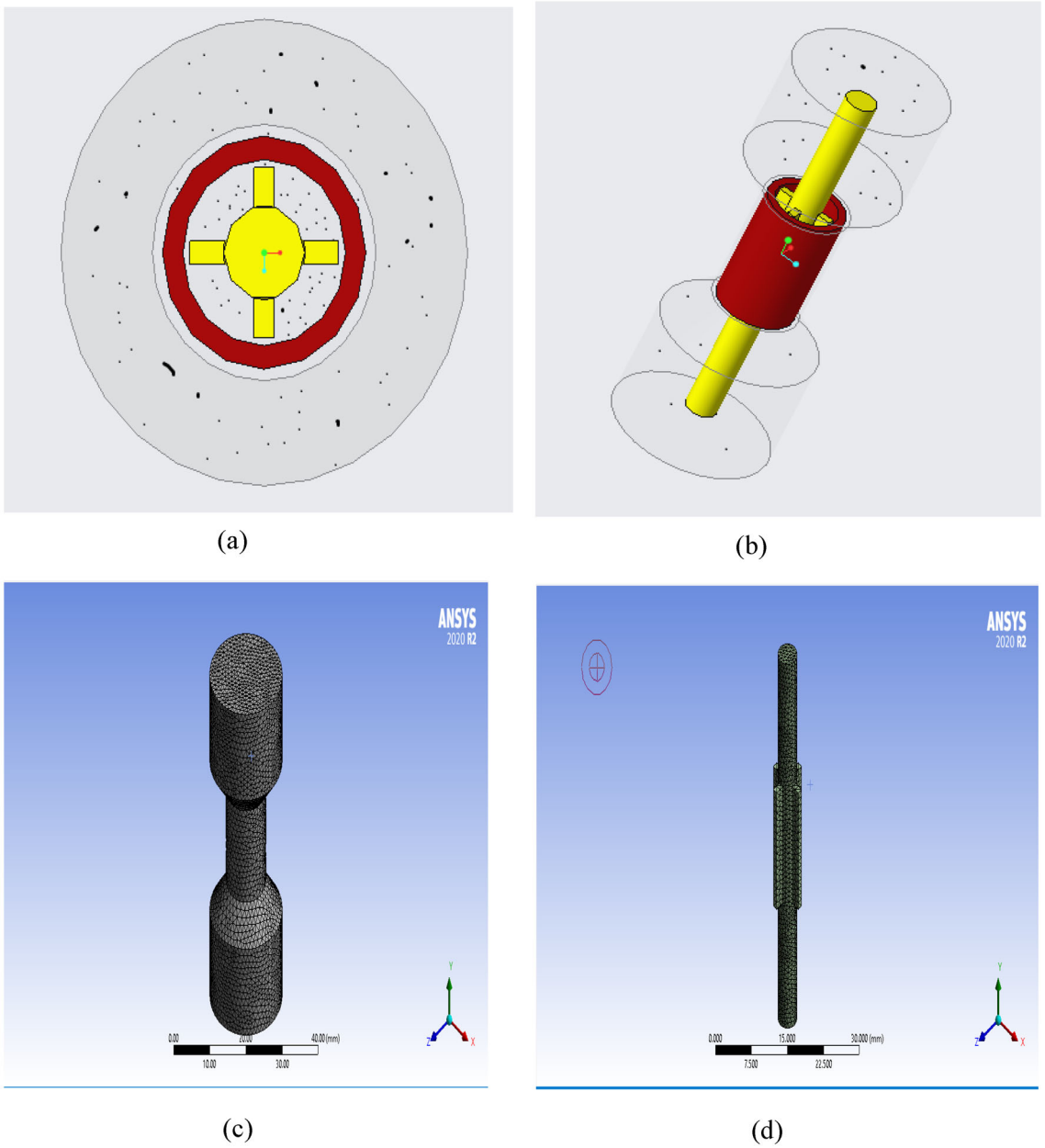


Fig. 6 Model of components **a** top view **b** side view **c** mesh file **d** CFG rod

Table 5 Meshing details

Sr. no.	Components	Material	Elements	Nodes
1	Workpiece	PLA	3963	5742
2	CFG rod	Copper	19,113	4287
3	Fluid	Polyborosiloxane	62,045	12,838
4	Assembly	–	85,094	22,867

Table 6 Boundary conditions

Sr.no	Boundary condition	Description
1	Inlet	Pressure Inlet 40 MPa
2	Outlet	Pressure Outlet 20 MPa
3	Fluid boundary	Stationary wall
4	CFG Rod	Rotation 100 rpm
5	Workpiece	Stationary

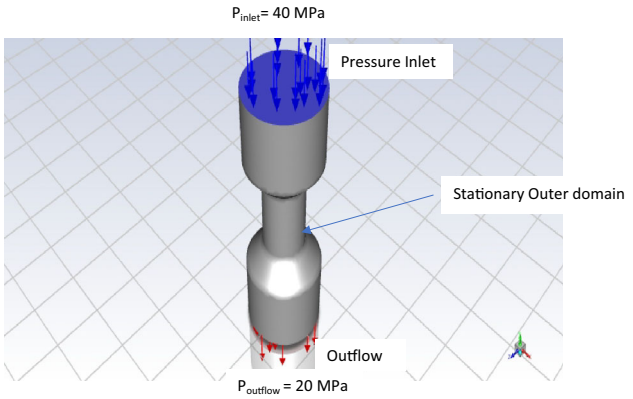


Fig. 7 Boundary conditions

3.2. Calculation of Wear Using CFD Model

Hardness of PLA work piece $= 47.74$ MPa

Pressure applied on the surface by the abrasive particles $= 2.343 \times 10^7$ Pa

$a = 25$ micron

$C = 0.6$, $d_{\text{media}} = 1219$ kg/m^3, $V_{\text{media}} = 290 \times e^{-6}$ m^3,

$d_{\text{abrasive}} = 1340$ kg/m^3

after entering the values,

$V_a = 159.37 \times 10^{-6}$ m^3,

$V_o = 0.8181 \times 10^{-14}$ m^3,

Overall abrasive particles in media's entire volume, $(N) = 194.81 \times 10^8$,

$N_a = 2.42 \times 10^8$,

$$K = \frac{N_a}{N} = 0.0124,$$

$W = 11.133 \times 10^6$ N,

Total amount of wear debris $= 1.4$ mg

Table 8 displays the measured response values, expressed as a percentage improvement in R_a and micro-hardness, that match to the CCD matrix. Response surface and contour plots have been used to analyze how variables affect the response.

Table 7 Process parameters with their limits

Factors	Levels		
	− 1	0	+ 1
A—CFG rod rotation (rpm)	60	80	100
B—Number of Cycle (Number)	3	6	9
C—Concentration of Abrasive (Fraction)	0.4	0.5	0.6
D—Extrusion Pressure (MPa)	10	20	30
E—Abrasive Particle size (microns)	300	400	500

3.3. Effects of Process Parameters on % Improved R_a

According to the fitted model, the following three-dimensional response surface and contour plots have been established for the effect of variables on % Improved R_a:

$$\begin{aligned}\%\,\textbf{Improved } R_a &= +64.31670 - 0.274250 \\ &\quad * \text{CGF rod rotation} - 4.51313 * \text{No. of cycle} - \\ &123.68269 * \text{Abrasive Concentration} + 0.327269 \\ &\quad * \text{Extrusion Pressure} + 0.073121 * \text{Abrasive} \\ &\text{Particle size} + 0.010521 * \text{CGF rod rotation} \\ &\quad * \text{No. of cycle} + 0.577500 * \text{CGF rod rotation} * \\ &\text{Abrasive Concentration} + 0.003600 \\ &\quad * \text{CGF rod rotation} * \text{Extrusion Pressure} \\ &\quad + 6.98750 * \text{No. of cycle} * \text{Abrasive Concentration} \\ &\quad - 0.002483 * \text{No. of cycle} * \text{Abrasive Particle size} \\ &\quad - 0.727500 \\ &* \text{Abrasive Concentration} * \text{Extrusion Pressure} \\ &\quad - 0.123625 * \text{Abrasive Concentration} * \\ &\text{Abrasive Particle size} + 0.105620 * \text{No. of cycle}^2 \\ &\quad + 102.18269 * \text{Abrasive Concentration}^2 - \\ &0.006632 * \text{Extrusion Pressure}^2\end{aligned}$$

ANOVA table for % Improved R_a with model F value as 39.87 is shown in Table 9, representing the model as significant. The likelihood of getting the greater F value, which may have been caused by noise, was 0.01%. The value of P below 0.05 represents that the model terms are significant. The significant model terms for % Improved R_a were A, AB, AC, AD, BC, BE, CD, CE, B^2, C^2 and D^2 with corresponding percentage contributions of 48.48%, 1.44%, 4.8%, 1.87%, 15.86%, 2%, 1.9%, 6.06%, 7%, and 2.95%. This contribution was obtained through dividing the sum of squares of the individual parameter with sum of squares of the model. Values larger than 0.1000 indicate that the model terms were non-significant. If there are a lot of unnecessary model terms, the model can be improved by applying the model reduction approach. A F value of 0.5 indicates that the lack of fit was not statistically significant. In this instance, there is an 84.19% likelihood that noise

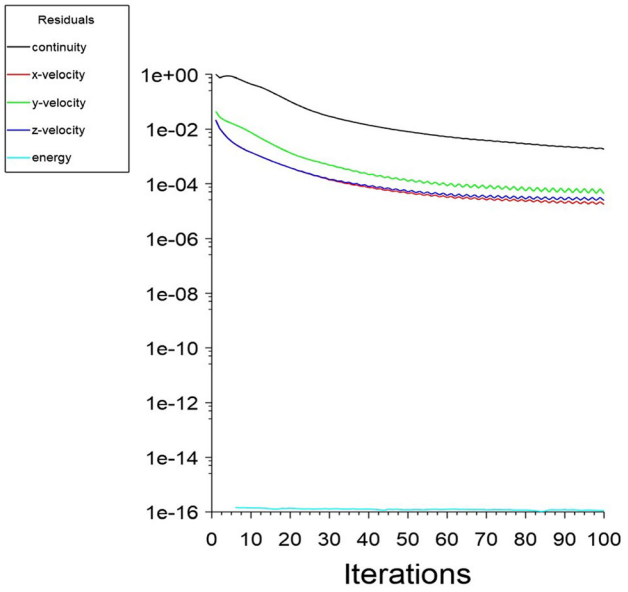


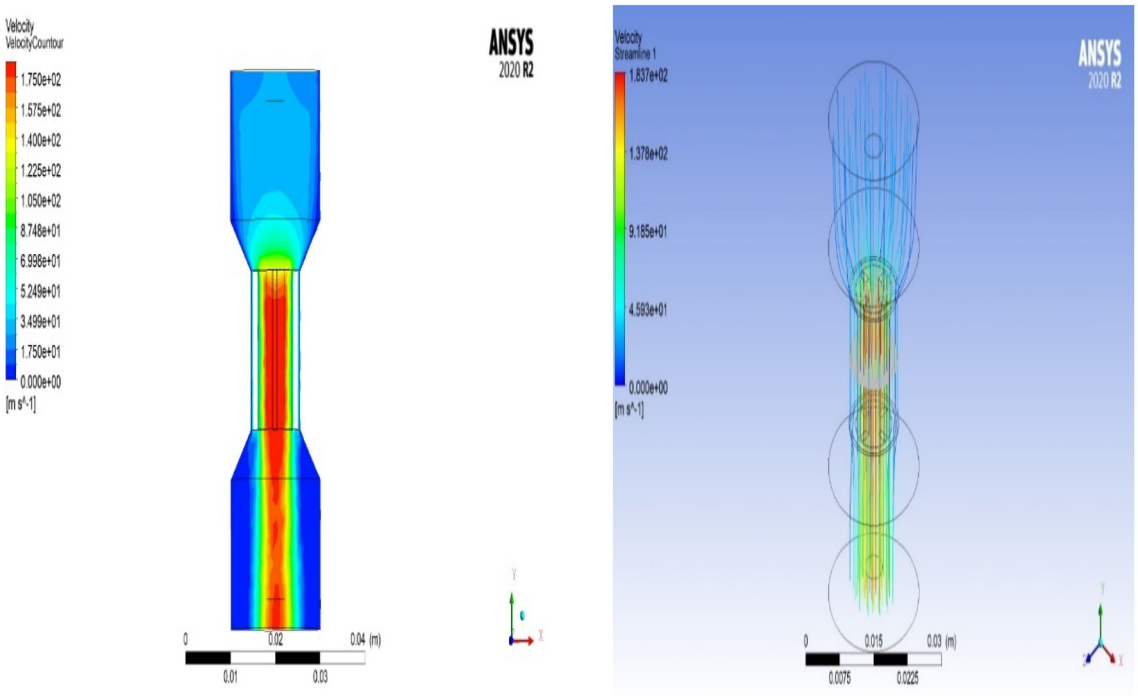
Fig. 8 Residual effects of the media during flow

will contribute to a greater Lack of Fit F value. The Adjusted R^2 of 0.9495 was in reasonable proportion to the Predicted R^2 of 0.8930, indicating that the difference was less than 0.2. The *S/N* ratio is estimated by Adeq Precision. It is advised to have a ratio greater than 4 to minimize the signal to noise ratio. The ratio of 29.549 shows an adequate signal which can be further used to navigate the design space.

Figure 12 demonstrates the regression model's outstanding acceptability. Each observed value may be compared to the predicted value that the model could produce, as illustrated in Fig. 12. The predicted line and the actual data have a respectable level of agreement, as seen in Fig. 12. The experimental results are consistent with expectations because the actual value points were closer to the predicted line.

The impression of varying process variables on the % Improved R_a is depicted in Fig. 13a, (i.e., perturbation graph). The design expert displays a base point for each parameter at the mid, denoted as (value 0). The sharp slope of the curves for the parameters CFG rod rotation (A), number of cycles (B), abrasive concentration (C), extrusion pressure (D), and abrasive particle size (E) demonstrates how strongly dependent these parameters are on improved R_a. While the interaction effects of the parameters were being shown, the explanation for this situation was examined.

The effect of CFG rod rotation and number of cycles is depicted in Fig. 13b. The surface plot shows that with CFG rod rotation, surface integrity got improved. During the flow of media through the restrictive passage, some of the abrasive particles participate in abrasion but some of the abrasive particles passes without contacting the surface



(a) (b)

Fig. 9 Effect on the media velocity

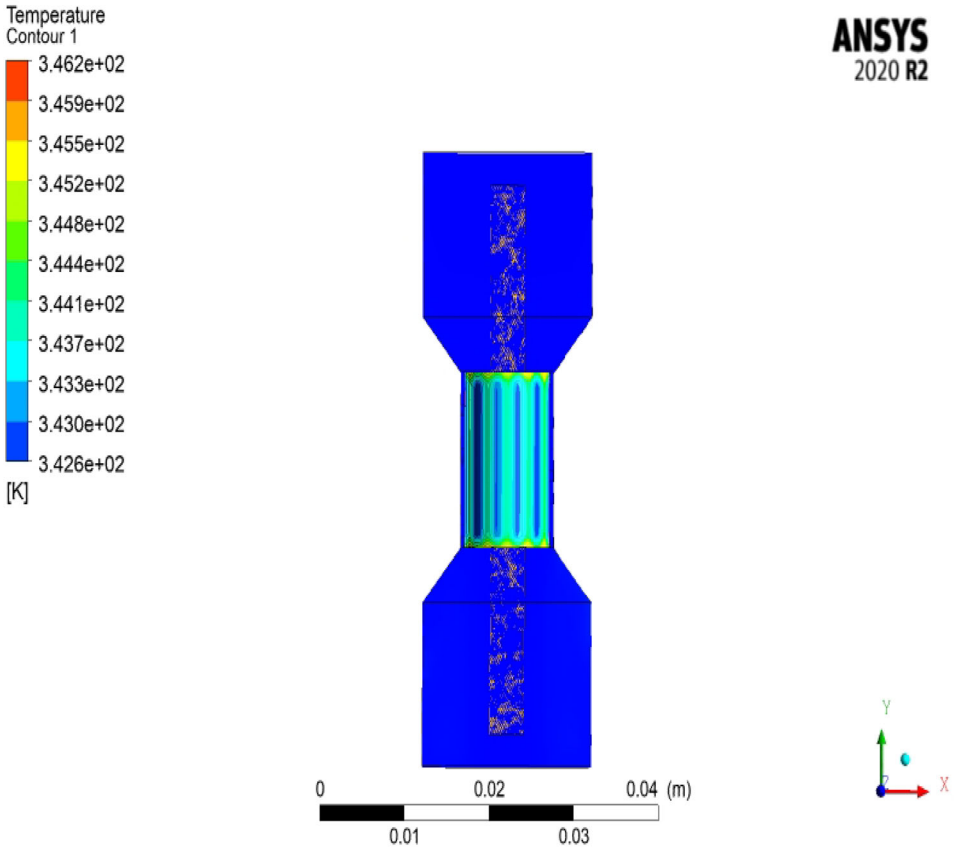


Fig. 10 Effect on the media temperature

which is called as non-active abrasive particle. These abrasive particles do not participate in abrasion. So, there is a chance that some of the abrasives may not engage with the workpiece as the media flows through the constrained passage, but as the CFG rod is rotated inside the media flow channel, it develops a centrifugal force in the flow path. As a result, the non-active abrasive particles which were not taking participation in abrasion previously are thrown

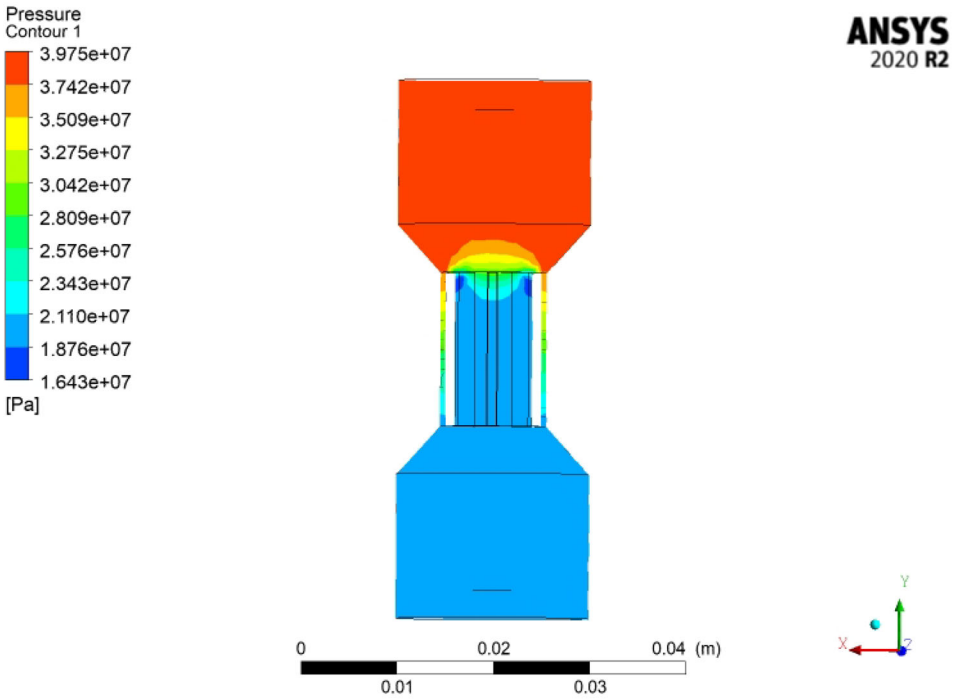


Fig. 11 Effect on the media pressure

Table 8 CCD for the observed responses and actual factors

Input							Output										
Std	Run	CFG rod rotation (rpm)	Number of cycle (number)	Abrasive concentration (fraction)	Extrusion pressure (MPa)	Abrasive particle size (Mesh)	R_a before machining (μm)	R_a after machining (μm)	% Improved R_a (%)	Weight before machining (g)	Weight after machining (g)	Material removal (mg)	Micro-hardness 1 (HV)	Micro-hardness 2 (HV)	Micro-hardness 3 (HV)	Average micro-hardness (HV)	Random uncertainty in micro-hardness
13	1	60	3	0.6	30	500	14.22	10.59	25.54	2.3554	2.3535	1.94	34.46	35.59	35.97	35.34	± 0.50
2	2	100	3	0.4	10	300	13.24	8.69	34.33	2.8642	2.8592	4.98	46.64	48.23	48.35	47.74	± 0.57
11	3	60	9	0.4	30	500	14.54	10.13	30.32	3.1034	3.0997	3.67	38.45	37.74	36.16	37.45	± 0.76
21	4	80	6	0.3	20	400	12.22	7.81	36.1	3.2365	3.2322	4.35	40.23	38.36	40.3	39.63	± 0.65
26	5	80	6	0.5	20	600	13.75	9.29	32.47	2.9134	2.9098	3.56	38.1	38.79	36.57	37.46	± 0.74
15	6	60	9	0.6	30	300	14.26	9.51	33.34	2.9964	2.9923	4.14	37.36	39.26	38.13	38.25	± 0.63
27	7	80	6	0.5	20	400	12.88	8.70	32.46	3.2645	3.2605	4.03	37.43	38.72	36.77	37.64	± 0.65
6	8	100	3	0.6	10	500	13.06	8.46	35.22	3.1264	3.1222	4.21	41.36	43.36	41.91	42.21	± 0.67
17	9	40	6	0.5	20	400	13.85	10.06	27.33	3.3256	3.3235	2.13	35.12	34.94	37.19	35.75	± 0.75
7	10	60	9	0.6	10	500	12.34	8.50	31.11	2.9623	2.9589	3.42	35.64	37.33	35.45	36.14	± 0.63
4	11	100	9	0.4	10	500	14.14	9.43	33.29	3.3534	3.3500	3.36	38.68	41.12	38.91	39.57	± 0.81
9	12	60	3	0.4	30	300	14.53	9.84	32.26	3.5325	3.5294	3.12	35.26	36.25	37.87	36.46	± 0.87
18	13	120	6	0.5	20	400	13.65	8.27	39.44	2.4684	2.4630	5.43	48.85	48.84	50.36	49.35	± 0.51
19	14	80	0	0.5	20	400	12.87	8.17	36.54	2.9965	2.9919	4.63	40.52	41.52	41.74	41.26	± 0.41
16	15	100	9	0.6	30	500	13.04	7.87	39.67	3.2433	3.2376	5.71	39.22	41.23	39.64	40.03	± 0.67
22	16	80	6	0.7	20	400	12.64	7.89	37.55	3.4254	3.4208	4.57	37.43	37.64	38.75	37.94	± 0.44
1	17	60	3	0.4	10	500	14.13	8.94	36.75	2.7699	2.7654	4.55	33.41	35.55	33.79	34.25	± 0.71
24	18	80	6	0.5	40	400	13.96	9.78	29.93	3.1433	3.1402	3.13	38.25	38.74	39.59	38.86	± 0.45
12	19	100	9	0.4	30	300	14.62	9.31	36.35	3.5546	3.5505	4.15	41.36	43.25	40.97	41.86	± 0.76
10	20	100	3	0.4	30	500	13.46	8.16	39.37	3.0232	3.0179	5.32	44.24	43.24	42.51	43.33	± 0.58
31	21	80	6	0.5	20	400	14.02	9.19	34.46	2.6754	2.6711	4.33	39.73	40.76	40.59	40.36	± 0.34
20	22	80	12	0.5	20	400	12.97	8.23	36.54	3.4235	3.4187	4.81	41.13	41.35	39.14	40.54	± 0.74
8	23	100	9	0.6	10	300	12.77	7.18	43.76	2.9845	2.9789	5.63	43.77	45.35	45.13	44.75	± 0.53
5	24	60	3	0.6	10	300	13.38	9.20	31.25	3.5342	3.5310	3.22	33.6	35.13	34.98	34.57	± 0.51
25	25	80	6	0.5	20	200	13.86	9.12	34.23	2.6755	2.6720	3.54	40.03	41.27	40.62	40.64	± 0.41
3	26	60	9	0.4	10	300	14.65	10.34	29.45	2.6674	2.6650	2.43	35.86	36.64	35.98	36.16	± 0.26
32	27	80	6	0.5	20	400	12.33	8.36	32.23	3.4324	3.4291	3.31	39.68	40.35	40.03	40.02	± 0.22
29	28	80	6	0.5	20	400	12.98	8.80	32.24	2.3442	2.3412	3.04	39.14	40.46	38.39	39.33	± 0.69
28	29	80	6	0.5	20	400	13.05	8.97	31.24	3.3423	3.3394	2.95	38.67	39.25	38.9	38.94	± 0.19
30	30	80	6	0.5	20	400	14.55	9.86	32.24	3.2453	3.2419	3.41	41.35	38.98	39.25	39.86	± 0.79
14	31	100	3	0.6	30	300	12.78	8.23	35.63	3.0023	2.9983	4.05	42.87	43.14	40.95	42.32	± 0.73
23	32	80	6	0.5	0	400	13.76	9.60	30.24	3.2246	3.2214	3.22	37.98	37.03	38.6	37.87	± 0.52

Table 9 Pooled ANOVA table for % improved R_a

Source	C. V	R. V	S.S	D.F	M.S	*F* value	Prob. > *F*
Model	+ 32.86	+ 64.31670	443.23	15	29.55	39.87	< 0.0001
A	+ 2.99	− 0.274250	214.92	1	214.92	290.03	< 0.0001
B	+ 0.2892	− 4.51313	2.01	1	2.01	2.71	0.1193
C	+ 0.2625	− 123.68269	1.65	1	1.65	2.23	0.1547
D	− 0.1375	+ 0.327269	0.4538	1	0.4538	0.6123	0.4453
E	− 0.3592	+ 0.073121	3.10	1	3.10	4.18	0.0578
AB	+ 0.6312	+ 0.010521	6.38	1	6.38	8.60	0.0097
AC	+ 1.15	+ 0.577500	21.34	1	21.34	28.80	< 0.0001
AD	+ 0.7200	+ 0.003600	8.29	1	8.29	11.19	0.0041
BC	+ 2.10	+ 6.98750	70.31	1	70.31	94.88	< 0.0001
BE	− 0.7450	− 0.002483	8.88	1	8.88	11.98	0.0032
CD	− 0.7275	− 0.727500	8.47	1	8.47	11.43	0.0038
CE	− 1.24	− 0.123625	24.45	1	24.45	33.00	< 0.0001
B^2	+ 0.9506	+ 0.105620	26.85	1	26.85	36.23	< 0.0001
C^2	+ 1.02	+ 102.18269	31.03	1	31.03	41.87	< 0.0001
D^2	− 0.6632	− 0.006632	13.07	1	13.07	17.63	0.0007
Residual			11.86	16	0.7410		
Lack of Fit			6.22	11	0.5655	0.5017	0.8419
Pure Error			5.64	5	1.13		
Cor Total			455.09	31	**C.V- Coded Value**		
Std. Dev	0.8608		R^2	0.9739	**R.V- Real Value**		
Mean	33.84		**Adjust. R^2**	0.9495	**S.S- Sum of Squares**		
C.V. %	2.54		**Pred. R^2**	0.8930	**D.F- Degree of freedom**		
Press	48.68		**Adeq. Precision**	29.5491	**M.S- Mean square**		
					Adjust. R^2- Adjusted R^2		
					Pred. R^2- Predicted R^2		
					Adeq. Precision- Adequate Precision		

Coded value: It means the assessment of the fit; the different follow-up analyses that are used depending on what type of model is fitted as well as the outcome of the analysis; and the importance of visualizing the response surface

Real Value: It is related to the Coded value

Sum of Squares: The sum of squares is a statistical measure of deviation from the mean

Degree of Freedom: Degrees of freedom are the maximum number of logically independent values, which may vary in a data sample. Degrees of freedom are calculated by subtracting one from the number of items within the data sample

Mean Square: Mean square represents an estimate of population variance. It is calculated by dividing the corresponding sum of squares by the degree of Freedom

F Value: An F statistic is a value you get when you run an ANOVA test or a regression analysis to find out if the means between two populations are significantly different

toward workpiece surface due to centrifugal action. This enhanced the material removal which corresponded better % R_a [32]. This can be clearly observed through the surface plots shown in Fig. 14.

Figure 13b also shows that, on increasing the number of cycles, the surface integrity of the workpiece was enhanced. The reason may be that as the amount of cycle was increased more and more abrasive particles engaged with the workpiece surface which increased the dynamic number of abrasive particles. Abrasive particles removed a small bit of surface material each time, and as the cycles increased, layer by layer more material was removed from the surface, improving the surface finishing [26]. The effect of number of cycles over the surface integrity can be clearly understood through the SEM images as shown in Fig. 15.

The effect of CFG rod rotation and abrasive concentration on the surface integrity is depicted in Fig. 13c. By enhancing the concentration of abrasives, the surface integrity was improved. More cutting edges came into contact with the workpiece surface as the abrasive concentration was increased, which resulted in more material removal. Moreover, this improves the surface's integrity [26]. Figure 13d demonstrates that while a superior surface

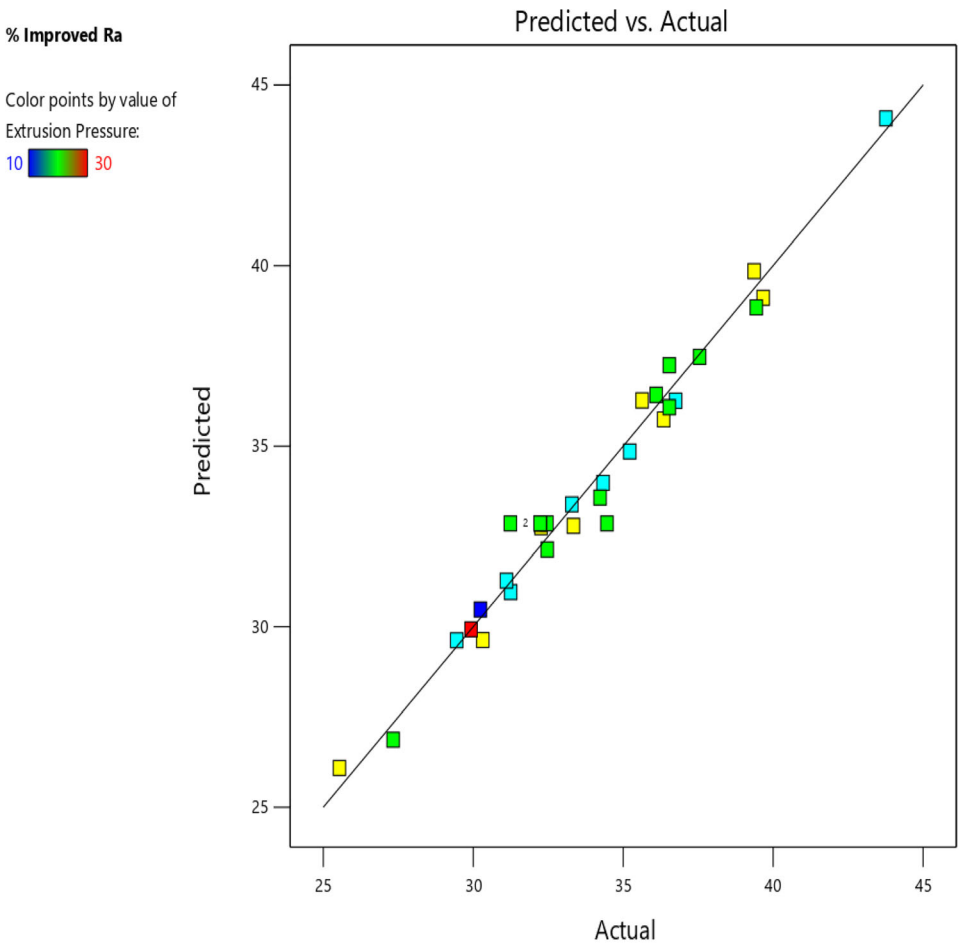


Fig. 12 Plot of actual versus predicted response of % improved R_a

finish was initially achieved as the extrusion pressure was raised, the surface finish declined as the extrusion pressure was raised further. This could be because as the extrusion pressure was raised, the force with which the abrasives struck the workpiece surface increased, leading to more material removal. However, as the extrusion pressure was raised further, the surface integrity was damaged as a result of the large amount of material removal, which led to a decrease in the surface integrity [33]. The impact of number of cycles and abrasive concentration on surface integrity is depicted in Fig. 13e. Surface integrity was improved in both instances as a result of an increase in dynamic abrasive particles [26]. Surface finish declines as abrasive mesh size is increased, as shown in Fig. 13f. Because the size of the abrasive is larger with lower mesh size abrasives, more material may be removed, and thus, better surface finish is achieved. This is the reason why abrasives with a greater mesh size provide inferior surface finishing to those with a smaller mesh size [34]. Figure 13g illustrates the relationship between extrusion pressure and abrasive concentration, whereas Fig. 13h illustrates the relationship between abrasive mesh size and abrasive concentration, as was previously discussed.

3.4. Effects of Process Parameters on Micro-hardness

According to the fitted model, the following three-dimensional response surface and contour plots have been established for the effect of variables on micro-hardness:

$$\begin{aligned}\textbf{Micro hardness} = {} & +27.52673 + 0.193997 \\ & * \text{ CGF rod rotation } + 0.154792 * \text{ No. of cycle } - \\ & 17.08333 * \text{ Abrasive Concentration } + 0.269185 \\ & * \text{ Extrusion Pressure } + 0.004392 * \text{ Abrasive} \\ & \text{Particle size } - 0.017469 * \text{ CGF rod rotation} \\ & * \text{ No. of cycle } - 0.004097 * \text{ CGF rod rotation} * \\ & \text{Extrusion Pressure } - 0.000290 * \text{ CGF rod rotation} \\ & * \text{ Abrasive Particle size } + 2.38958 * \text{ No. of} \\ & \text{cycle } * \text{ Abrasive Concentration } + 0.000519 \\ & * \text{ Extrusion Pressure } * \text{ Abrasive Particle size } + \\ & 0.001726 * \text{ CGF rod rotation}^2 - 0.003560 \\ & * \text{ Extrusion Pressure}^2\end{aligned}$$

ANOVA table for micro-hardness with model F value as 37.20, is shown in Table 10, represents the model as

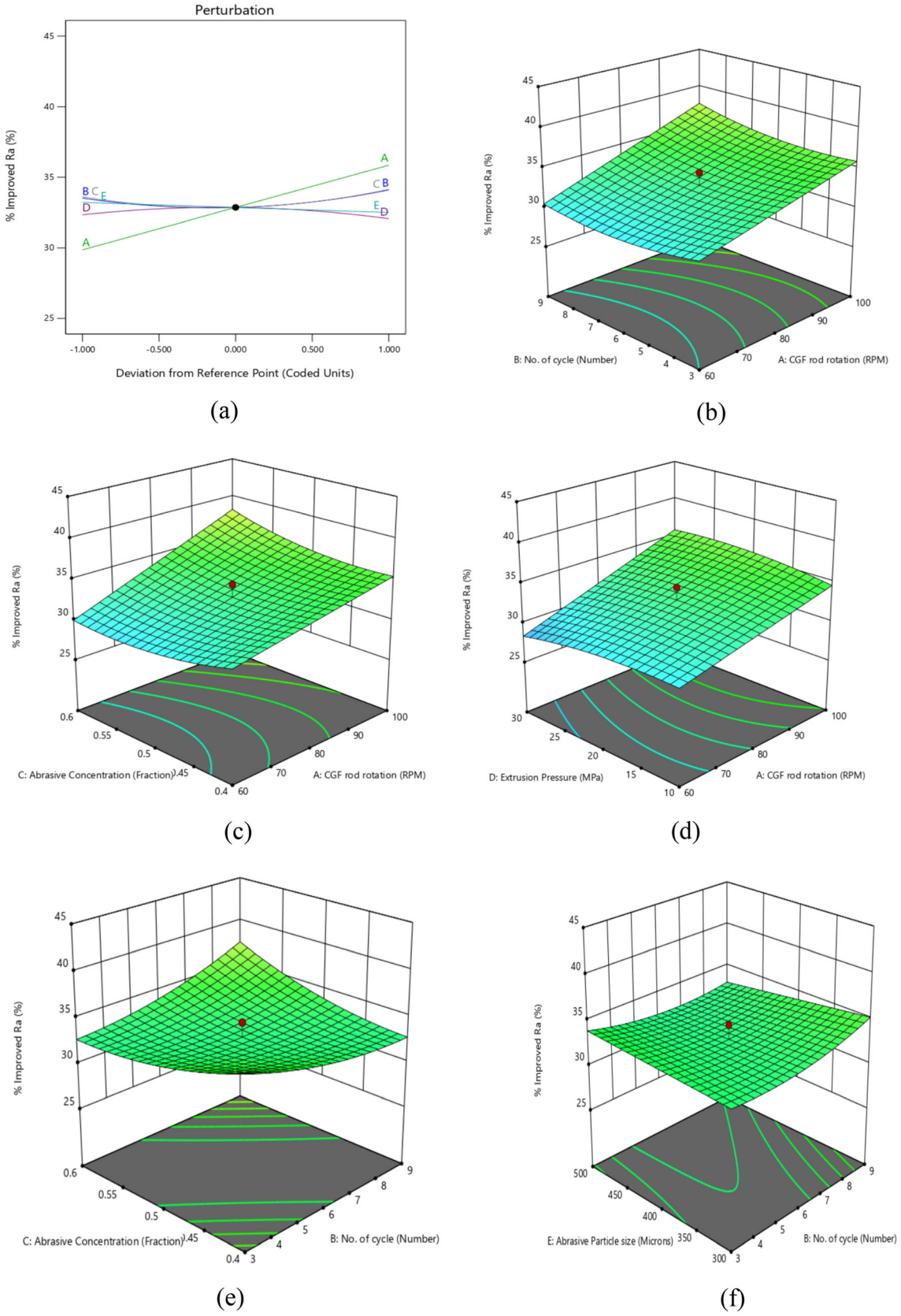


Fig. 13 Effect of significant parameters on the response (% Improved R_a)

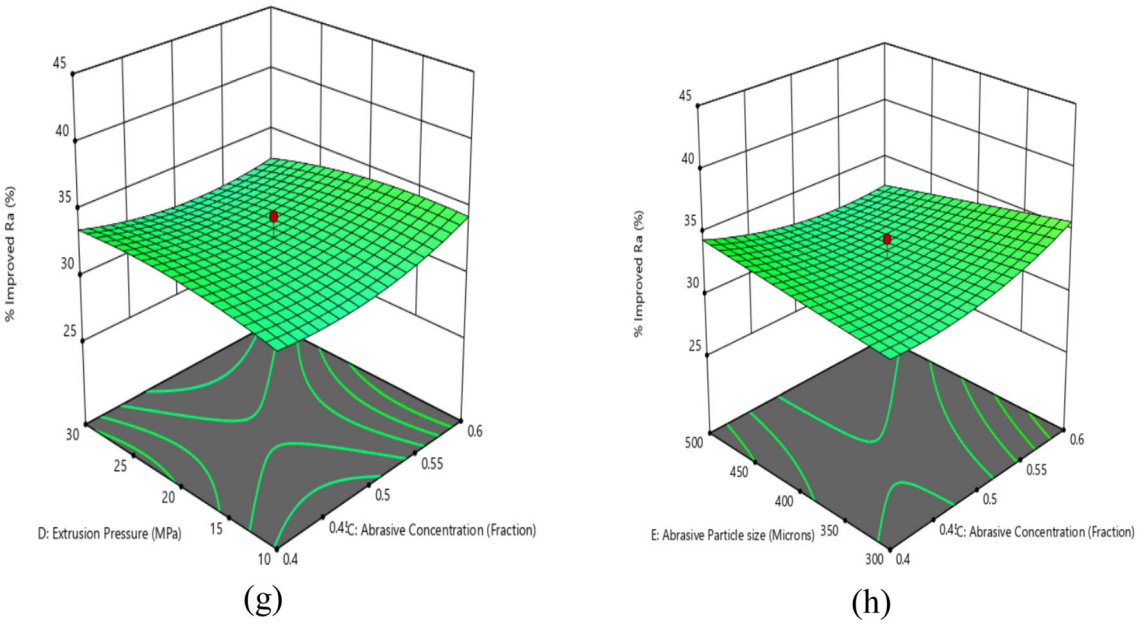


Fig. 13 continued

significant. The likelihood of getting the greater F value, which may have been caused by noise, was 0.01%. The value of P below 0.05 represents the model terms as significant. The major model terms for micro-hardness were A, E, AB, AD, AE, BC, DE, A^2, D^2 with corresponding percentage contributions of 76.08%, 4.78%, 4.97%, 3.03%, 1.52%, 2.32%, 1.22%, 4.02% and 1.07%. Values larger than 0.1000 represented the model terms as non-significant. The model can be strengthened by using the model reduction strategy if there are many extraneous model terms. The F value of 0.76 indicates that there was no substantial Lack of Fit. There is a 68.77% chance that a greater Lack of Fit F value, which could be caused by noise, will occur in this situation. The Adjusted R^2 of 0.9334 was in reasonable proportion to the Predicted R^2 of 0.8978, indicating that the difference was less than 0.2. The *S/N* ratio is estimated by Adeq Precision. It is advised to have a ratio greater than 4 to minimize the signal to noise ratio. The ratio of 25.511 shows an adequate signal which can be further used to navigate the design space.

Figure 16 demonstrates the regression model's outstanding acceptability. Each observed value may be compared to the predicted value that the model could produce, as illustrated in Fig. 16. A reasonable match between the predicted line and the actual data is seen in Fig. 16. The experimental results are consistent with expectations because the actual value points were closer to the predicted line.

The effect of varying process parameters on the micro-hardness is depicted in Fig. 17a (i.e., perturbation graph). The design expert displays a base point for each parameter at the mid, denoted as (value 0). The sharp slope of the curves for the parameters CFG rod rotation (A), number of cycles (B), abrasive concentration (C), extrusion pressure (D) and abrasive particle size (E) demonstrates how strongly dependent these parameters are on micro-hardness. While the interaction effects of the parameters were being shown, the explanation for this situation was examined.

Figure 17b depicts the effect of CFG rod rotation and number of cycles over the micro-hardness of the surface. Figure 17b illustrates how the surface's micro-hardness rose as rotating speed was increased. A combination of the axial forces as well as radial forces acting in CFAAFM process during the finishing stage might be liable for the material removal. The radial force grows as the rotational speed rises, enhancing the net force. Due to the use of centrifugal force in the media flow path, abrasive particles were flung at the surface with more force. It raised the micro-hardness of the finishing surface and corresponded work hardening of the surface [35]. The effect of CFG rod rotation is shown in Fig. 18. Fig. 18 clearly shows increase in micro-hardness of the surface as the CFG rod rotation was increased.

On the other side as the number of cycles are increased larger number of abrasive particles impacted over the surface and developed a stress over it. This is the reason of increasing the micro-hardness with the number of cycles [35]. Figure 17c shows the enhancement of micro-hardness on increasing the extrusion pressure of the abrasive laden

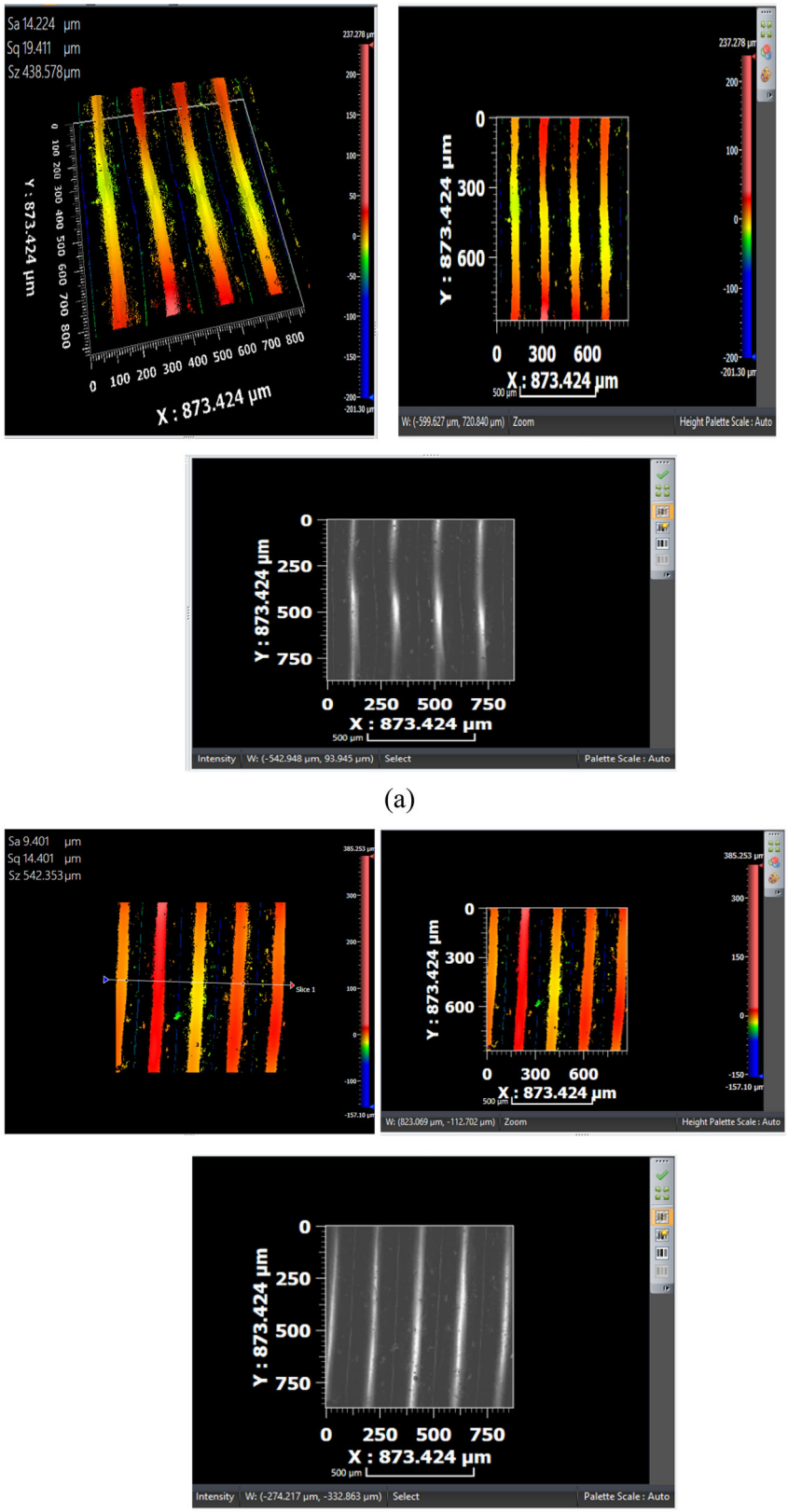


Fig. 14 Surface roughness plots **a** without rotation **b** CFG rod rotation at 80 rpm

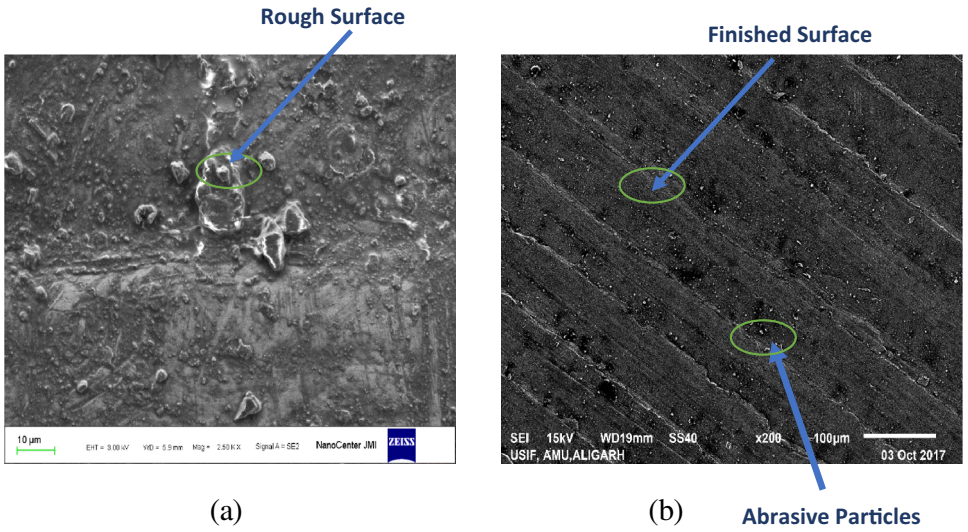


Fig.15 SEM images of the surface **a** before finishing **b** After 6 No. of cycles

Table 10 Pooled ANOVA table for micro-hardness

Source	C. V	R. V	S.S	D.F	M.S	*F* value	Prob. > *F*
Model	+ 39.31	+ 27.52673	353.91	12	29.49	37.20	< 0.0001
A	+ 3.35	+ 0.193997	269.27	1	269.27	339.60	< 0.0001
B	− 0.1437	+ 0.154792	0.4959	1	0.4959	0.6255	0.4388
C	− 0.2746	− 17.08333	1.81	1	1.81	2.28	0.1473
D	+ 0.0679	+ 0.269185	0.1107	1	0.1107	0.1396	0.7128
E	− 0.8396	+ 0.004392	16.92	1	16.92	21.34	0.0002
AB	− 1.05	− 0.017469	17.58	1	17.58	22.17	0.0002
AD	− 0.8194	− 0.004097	10.74	1	10.74	13.55	0.0016
AE	− 0.5794	− 0.000290	5.37	1	5.37	6.77	0.0175
BC	+ 0.7169	+ 2.38958	8.22	1	8.22	10.37	0.0045
DE	+ 0.5194	+ 0.000519	4.32	1	4.32	5.44	0.0308
A^2	+ 0.6903	+ 0.001726	14.23	1	14.23	17.95	0.0004
D^2	− 0.3560	− 0.003560	3.78	1	3.78	4.77	0.0416
Residual			15.07	19	0.7929		
Lack of Fit			10.24	14	0.7317	0.7589	0.6877
Pure Error			4.82	5	0.9643		
Cor total			368.98	31	**C.V- Coded Value**		
Std. Dev	0.8905		R^2	0.9592	**R.V- Real Value**		
Mean	39.56		**Adjust.** R^2	0.9334	**S.S- Sum of Squares**		
C.V. %	2.25		**Pred.** R^2	0.8978	**D.F- Degree of freedom**		
Press	48.68		**Adeq. Precision**	29.5491	**M.S- Mean square**		
					Adjust. R^2**- Adjusted** R^2		
					Pred. R^2**- Predicted** R^2		
					Adeq. Precision- Adequate Precision		

Coded value: It means the assessment of the fit; the different follow-up analyses that are used depending on what type of model is fitted as well as the outcome of the analysis; and the importance of visualizing the response surface

Real Value: It is related to the Coded value

Sum of Squares: The sum of squares is a statistical measure of deviation from the mean

Degree of Freedom: Degrees of freedom are the maximum number of logically independent values, which may vary in a data sample. Degrees of freedom are calculated by subtracting one from the number of items within the data sample

Mean Square: Mean square represents an estimate of population variance. It is calculated by dividing the corresponding sum of squares by the degree of Freedom

F Value: An F statistic is a value you get when you run an ANOVA test or a regression analysis to find out if the means between two populations are significantly different

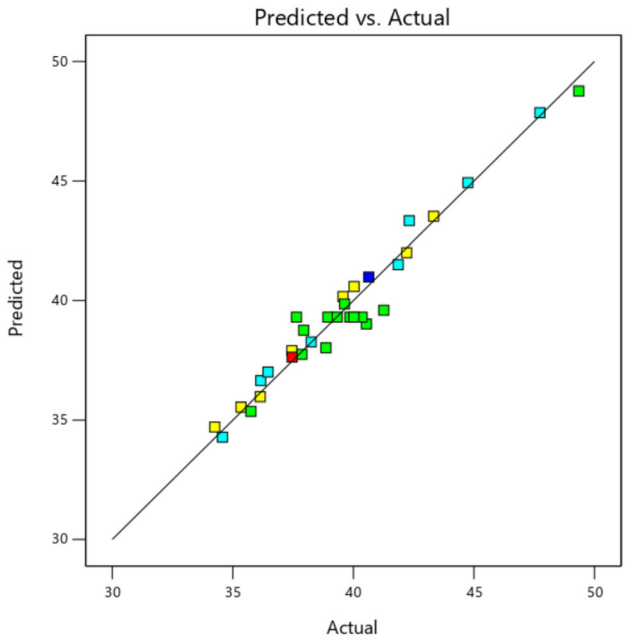


Fig. 16 Plot of actual versus predicted response of micro-hardness

media. Only axial force, acting in media flow direction, is responsible for material removal in a conventional AFM operation. However, the CFAAFM procedure combines centrifugal force with conventional AFM. Both axial as well as radial force are used in the currently designed CFAAFM to remove material. The direction normal over the surface is where radial force operates. A combination of axial and radial forces will cause the abrasive particles to exert their force. The initial resultant force is lower at lower pressure, and abrasives are unable to shear the surface's peaks. Yet, when the pressure rises, the resulting force rises as well and eliminates more debris, improving the surface smoothness as well as micro-hardness [35]. Micro-hardness decreases as abrasive mesh size increases, as shown in Fig. 17d. Abrasive particles with a larger mesh size are smaller and remove much less material from the surface, exerting less force on the surface and lowering the micro-hardness [35]. Figure 17e demonstrates the surface's micro-hardness was marginally increased as a result of the increase in concentration of abrasive. As the concentration of abrasive was raised, greater number of abrasives made contact with the workpiece's surface, causing residual stress to accumulate there and raising the surface's micro-hardness [35]. Figure 17f shows increase in micro-hardness with extrusion pressure while the micro-hardness of the surface was decreased on increasing the abrasive mesh size as discussed earlier.

The optimization was performed by using Design expert software. The desirability factor ranged from 0 to 1, with the smallest value indicating a low factor. The process variables that have the highest attractiveness have the best variable settings. The software evaluated the best values for the input parameters and the corresponding replies, and the results are detailed in Table 11. Some response variables were ignored during single factor optimization; however, they were taken into account and given equivalent importance during multivariable optimization. Confirmation tests were carried out in order to validate the optimized replies, as indicated in Table 11. The experimental validation revealed that the anticipated values and experimental values were quite similar.

4. Contribution of the Developed Process on Surface Integrity and Micro-hardness of the Surface

AFM is one of the most well-liked non-traditional finishing techniques, that is used in a variety of contemporary industries to replace the numerous manual finishing processes. One of the significant drawbacks of AFM is the reduced material removal, though it is evident from this paper that only a small portion of abrasive particles actually take part in the finishing process. The remaining abrasive particles only continue to flow through the work piece passage, increasing finishing time and cost [32]. The following is a summary of the proposed hypothesis regarding the causes of the improved surface integrity and micro-hardness with the application of centrifugal force to AFM:

(a) Modification of the abrasive particle distribution pattern at the inner surface of the hollow work piece, resulting in more abrasive particles participating in abrasion. The total number of active abrasive particles in conventional AFM was found as 1.392×10^8, which increased up to 2.42×10^8 in case of CFAAFM process. This shows 73.85% improvement in dynamic abrasive particles.

(b) Abrasive particles that are flowing axially may be deflected so that they strike the workpiece surface at an incident angle.

(c) The percentage improvement in surface finish for the CFAAFM process was determined to be 39.27%, which was somewhat less than the Thermal additive Centrifugal Abrasive Flow Machining (TACAFM) process (i.e., 39.74%). Although TACAFM offers slightly superior finishing, it also creates a recast layer over the surface and requires more energy input than CFAAFM does. But this process gives better result in comparison with Electrochemical-assisted Abrasive flow machining.

(d) Although micro-hardness in case of CFAAFM is increased but is very less in comparison with other hybrid forms of AFM. The shear yield strength of the

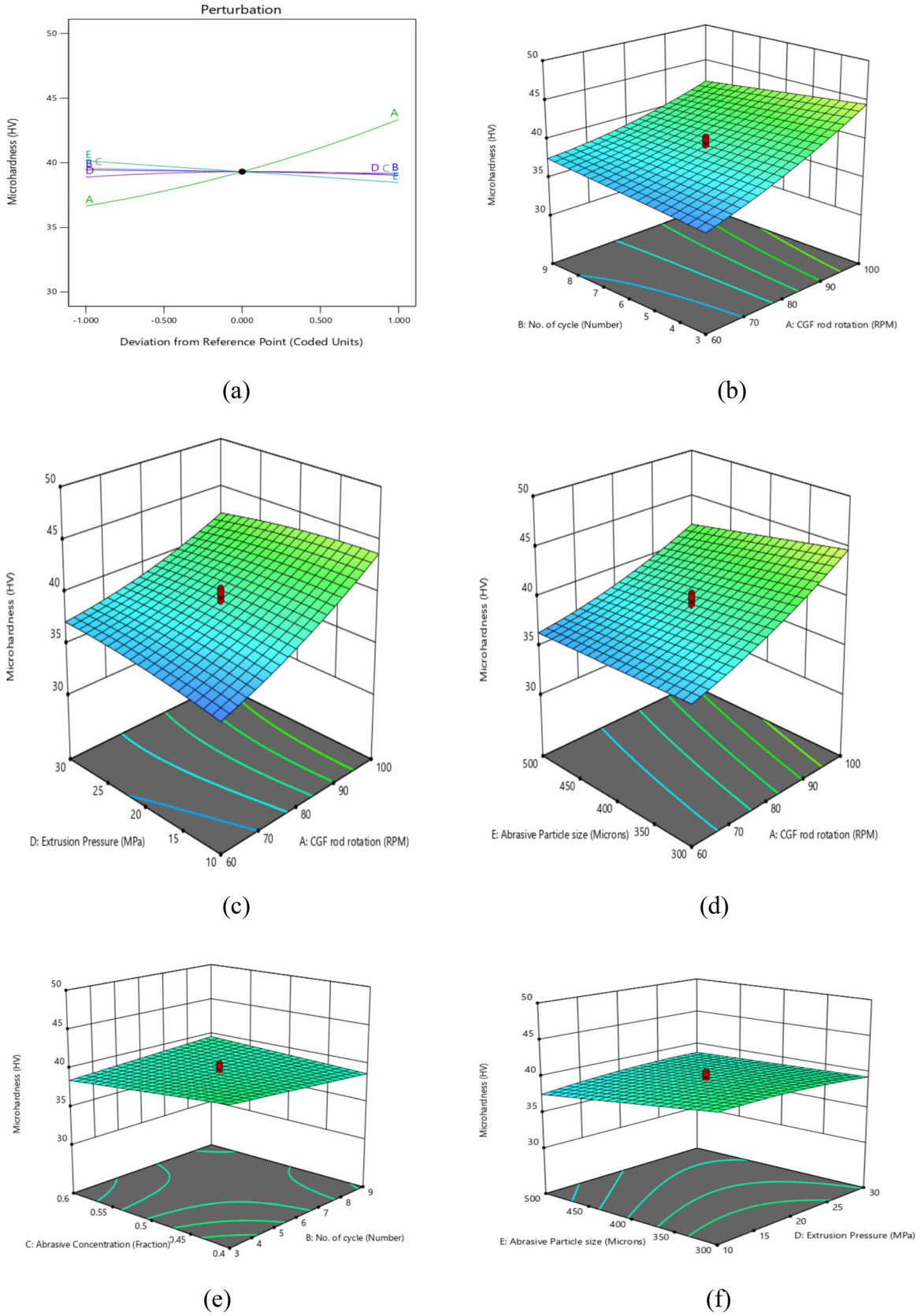


Fig. 17 Effect of significant variables over the micro-hardness

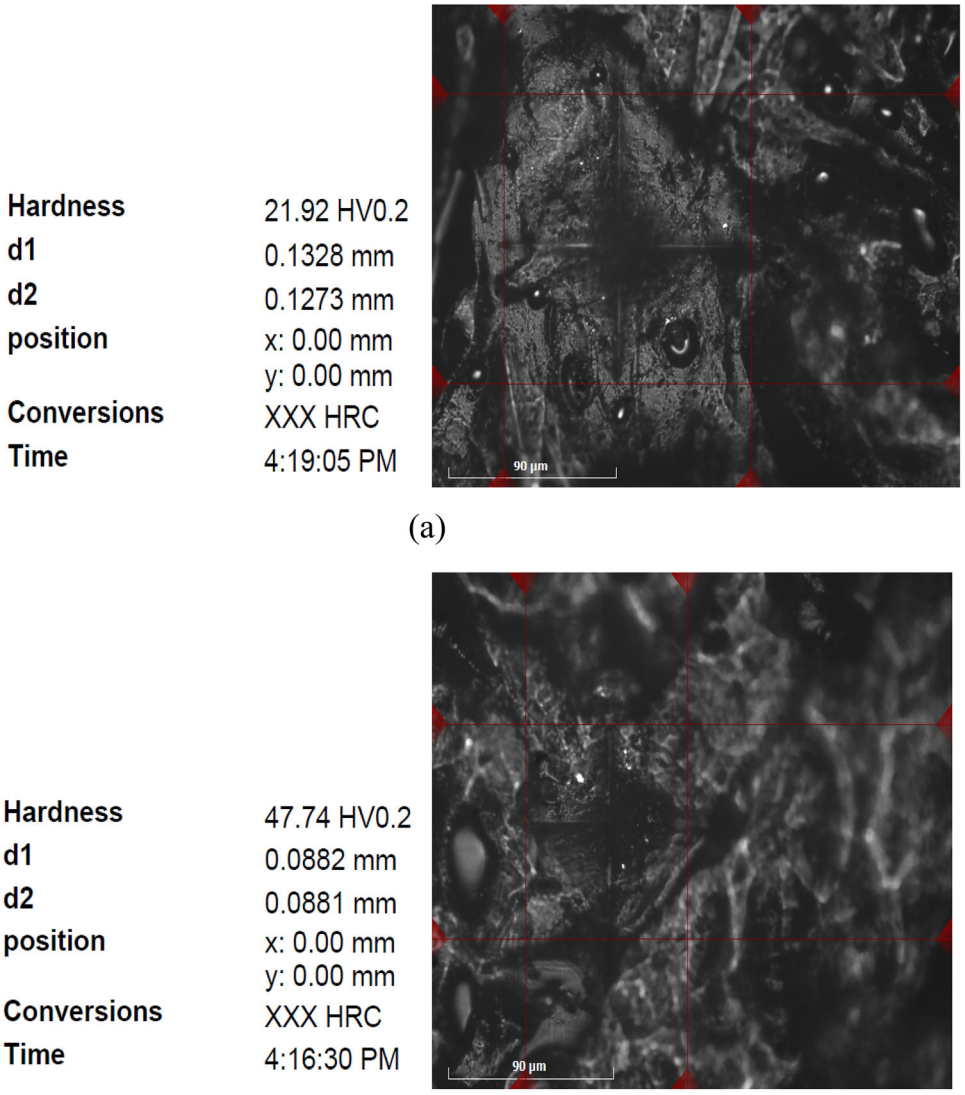


Fig. 18 Micro-hardness of the surface **a** without rotation **b** CFG rotation at 80 rpm

workpiece is strongly dependent on the surface texture. On increasing the surface hardness, strength also increases which can be attributed to the strengthening of the material and the rise of stress concentrators which may cause failure of the component.

5. Conclusions

Following statements can be carried out through experimental and modeling study -

- In comparison with the standard AFM method, the experimental findings of the CFAAFM technique demonstrate an average improvement in surface quality of 39.27%.
- CFG rod rotational speed has the largest contribution of 48.48% and 76.08% against % Improved R_a and micro-hardness, respectively.
- With the addition of Centrifugal force in media flow path, dynamic number of abrasive particles are increased by 73.85% in comparison with convention AFM process.
- This process increases the interaction between the abrasive particles and workpiece surface which reduces the finishing time in comparison with conventional AFM.
- All the experimental results were validated through the previously available literatures, which shows a good agreement between them.
- From computational model, it can be concluded that total wear debris was found 1.4 mg, which was quite nearer to the experimental results.

Table 11 Single and multifactor optimization outcomes

Optimization type	Objective	Optimized process variables					Response (predicted)	Response (experimental)	Desirability
		CFG rod rotation (rpm)	Number of cycles (number)	Abrasive concentration (fraction)	Extrusion pressure (MPa)	Abrasive particle size (Mesh)			
Single response	To maximize % Improved R_a	80	6	0.5	20	400	32.86%	30.25%	0.860
Single response	To maximize Micro-hardness	60	4	0.5	20	300	39.31 HV	38.80 HV	0.891
Multi response	To maximize Micro-hardness and maximize % Improved R_a	80	6	0.5	30	400	41.29 HV and 35.34%	39.27 HV and 33.47%	0.868

- The computational results show that the extrusion pressure of the media was continuously decreasing during the media flow due to frictional losses while the velocity was found maximum while the abrasive particles impacted over the workpiece surface.
- The optimum values of percentage improvement in surface finish and micro-hardness are found as 35.34% and 41.29 HV, respectively, which were obtained at 80 rpm of CFG rod rotation, 6 number of cycles, 0.5 fraction of abrasive concentration, 30 MPa of extrusion pressure and 400 mesh size of abrasive particles, respectively.

Author Contribution The design and conceptualization were implemented by Dr. PA The data collection and analysis were performed by AB. The draft of the manuscript and all editing were done by Dr. PA. Visualization and supervision were done by Dr. PA. All the authors have read and approved the final manuscript.

Declarations

Conflict of interest The author declares no competing interests.

Ethics Approval Not applicable.

References

[1] T.S. Petrovskaya, N.E. Toropkov, E.G. Mironov and F. Azarmi, 3D printed biocompatible polylactide-hydroxyapatite based material for bone implants. Mater. Manuf. Process., **33** (2018) 1899–1904. https://doi.org/10.1080/10426914.2018.1476764.

[2] S. Chaitanya and I. Singh, Processing of PLA/sisal fiber biocomposites using direct- and extrusion-injection molding. Mater. Manuf. Process., **32** (2017) 468–474. https://doi.org/10.1080/10426914.2016.1198034.

[3] M. Rimašauskas, T. Kuncius and R. Rimašauskienė, Processing of carbon fiber for 3D printed continuous composite structures. Mater. Manuf. Process., **34** (2019) 1528–1536. https://doi.org/10.1080/10426914.2019.1655152.

[4] Z. Liu, Y. Weng, Z. Huang, Y. Jin and J. Hu, Manufacture of a hydrophobic CaO/polylactic acid composite. Mater. Manuf. Process., **34** (2019) 303–311. https://doi.org/10.1080/10426914.2018.1512113.

[5] M. Harris, J. Potgieter, R. Archer and K.M. Arif, In-process thermal treatment of polylactic acid in fused deposition modelling. Mater. Manuf. Process., **34** (2019) 701–713. https://doi.org/10.1080/10426914.2019.1566611.

[6] M. Pastorek and A. Kovalcik, Effects of thermal annealing as polymer processing step on polylactic acid. Mater. Manuf. Process., **33** (2018) 1674–1680. https://doi.org/10.1080/10426914.2018.1453153.

[7] P. Ali, R.S. Walia, Q. Murtaza and M.S. Ranganath, Modeling and analysis of developed thermal additive centrifugal abrasive flow machining process. Surf. Topogr. Metrol. Prop., **8** (2020) 035013.

[8] A. Bhardwaj, K. Srinivas and R. Chaudhary, Modelling of wind turbine magnet for magnetic abrasive finishing and magnetic force assisted abrasive flow machining process. J. Eng. Res. ICARI Special Issue (2021). https://doi.org/10.36909/jer.ICARI.15291.

[9] A. Bhardwaj, K. Srinivas and R. Chaudhary, Analysis of shapes of centrifugal force- generating rod in centrifugal force-assisted abrasive flow machining process. MAPAN, **38**(1) (2023) 459–479. https://doi.org/10.1007/s12647-023-00626-0.

[10] P. Ali, S. Dhull, R.S. Walia, Q. Murtaza and M. Tyagi, Hybrid abrasive flow machining for nano finishing—a review. International Conference on Advancements in Aeromechanical materials for manufacturing. Mater. Today Proc., **4**(8) (2017) 7208–7218.

[11] P. Ali, R.S. Walia, Q. Murtaza and M.S. Ranganath, Enhancement of surface integrity using spark assisted abrasive flow machining process. Int. J. Mech. Prod. Eng. Res. Dev. (IJMPERD), **8**(6) (2018) 635–642.

[12] A. Parvesh, M.S. Ranganath, R.S. Walia and Q. Murtaza optimization of micro hardness of finished surface in spark assisted abrasive flow machining. Int. J. Mech. Prod. Eng. Res. Dev. **9** (2) (2019) 415–424. (IJMPERD) ISSN (P): 2249–6890; ISSN (E): 2249–8001.

[13] J.D. Kechagias, K. Ninikas, M. Petousis, N. Vidakis and N. Vaxevanidis, An investigation of surface quality characteristics of 3D printed PLA plates cut by CO_2 laser using experimental design. Mater. Manuf. Process., **36** (2021) 1544–1553. https://doi.org/10.1080/10426914.2021.1906892.

[14] C.C. Kuo and R.C. Mao, Development of a precision surface polishing system for parts fabricated by fused deposition

modeling. Mater. Manuf. Process, **31**(8) (2016) 1113–1118. https://doi.org/10.1080/10426914.2015.1090594.
[15] A. Kumar, Z. Alam, D.A. Khan and S. Jha, Nano finishing of FDM-fabricated components using ball end magnetorheological finishing process. Mater. Manuf. Process., **34** (2019) 232–242. https://doi.org/10.1080/10426914.2018.1512136.
[16] J. Kechagias and D. Chaidas, Fused filament fabrication parameter adjustments for sustainable 3D printing. Mater. Manuf. Process., **38**(8) (2023) 933–940. https://doi.org/10.1080/10426914.2023.2176872.
[17] J.D. Kechagias, Stephanos and P. Zaoutsos, Optimising fused filament fabrication surface roughness for a dental implant. Mater. Manuf. Process., **38**(8) (2023) 954–959. https://doi.org/10.1080/10426914.2023.2176870.
[18] N.A. Fountas, I. Papantoniou, J.D. Kechagias, D.E. Manolakos and N.M. Vaxevanidis, Modeling and optimization of flexural properties of FDM-processed PET-G specimens using RSM and GWO algorithm. Eng. Fail. Anal., **138** (2022) 106340. https://doi.org/10.1016/j.engfailanal.2022.106340.
[19] N.A. Fountas, K. Kitsakis, K.E. Aslani, J.D. Kechagias and N.M. Vaxevanidis, An experimental investigation of surface roughness in 3D-printed PLA items using design of experiments. Proc. Inst. Mech. Eng. Part J J. Eng. Tribol., **236**(10) (2022) 1979–1984. https://doi.org/10.1177/13506501211059306.
[20] K.E. Aslani, D. Chaidas, J. Kechagias, P. Kyratsis and K. Salonitis, Quality performance evaluation of thin-walled PLA 3D printed parts using the Taguchi method and grey relational analysis. J. Manuf. Mater. Process., **4**(2) (2020) 47. https://doi.org/10.3390/jmmp4020047.
[21] A. Bhardwaj, P. Ali, R.S. Walia, Q. Murtaza and S.M. Pandey, Development of hybrid forms of AFM process: a review. In: K. Shanker, R. Shankar and R. Sindhwani (eds) Advances in industrial and production engineering, Springer Singapore (2019), pp. 41–67. https://doi.org/10.1007/978-981-13-6412-9_5 .
[22] R.S. Walia, H.S. Shan and P. Kumar, Parametric optimization of centrifugal force-assisted abrasive flow machining (CFAAFM) by the Taguchi method. Mater. Manuf. Process, **21** (2006) 375–382. https://doi.org/10.1080/10426910500411645.
[23] https://en.wikipedia.org/wiki/Archard_equation. (Cited on 24th March 2023).
[24] R.S. Walia, H.S. Shan and P.K. Kumar, Determining dynamically active abrasive particles in the media used in centrifugal force assisted abrasive flow machining process. Int. J. Adv. Manuf. Technol., **38** (2008) 1157–1164.
[25] N.R. Draper, Center points in second order response surface designs. Techno Metr., **24** (1982) 127–133.
[26] P. Ali, S.M. Pandey, M.S. Ranganath, R.S. Walia and Q. Murtaza, Experimentation and modelling of CNT additive abrasive media for micro finishing. Measurement, **151** (2020) 107–133. https://doi.org/10.1016/j.measurement.2019.107133.
[27] T. Wang, D. Chen and L. Zhang, An Study on key parameters of a new Abrasive Flow Machining process for surface finishing. Int. J. Adv. Manuf. Technol., **101** (2019) 39–54.
[28] M.A. Manjunath, A. Kumar, V. Prakash and N. Balashnmugam, Simulation of force generated and material removal in Abrasive flow finishing for Aluminium material. Proceedings of 10th International conference on Precision. Meso, Micro and nano engineering, IIT Madras (2017). https://api.semanticscholar.org/CorpusID:162173146.
[29] H. Tzeng, B. Yan, R. Hsu and Y. Lin, Self-modulating abrasive medium and its application to abrasive flow machining for finishing micro channel surfaces. Int. J. Adv. Manuf. Technol., **32** (2007) 1163–1169.
[30] P. Ali, R.S. Walia, Q. Murtaza and R.M. Singari, Material removal analysis of hybrid EDM-assisted centrifugal abrasive flow machining process for performance enhancement. J. Braz. Soc. Mech. Sci. Eng., **42** (2020) 302. https://doi.org/10.1007/s40430-020-02375-6.
[31] S. Singh, M.R. Sankar, V.K. Jain and J. Ramkumar, Modelling of finishing forces and surface roughness in Abrasive flow finishing process using rheological properties. 5th International & 26th All India Manufacturing Technology Design and Research Conference (AIMTDR) December 12th–14th, (2014) IIT Guwahati, Assam, India.
[32] R.S. Walia, H.S. Shan and P. Kumar, Abrasive flow machining with additional centrifugal force applied to the media. Mach. Sci. Technol., **10** (2006) 1–14.
[33] R.K. Jain and V.K. Jain, Simulation of surface generated in abrasive flow machining process. Robot. Comput. Integr. Manuf., **15** (1999) 403–412.
[34] D. Amir, F. Behzad, M. Behrouz and N. Navid, Abrasive flow finishing of stainless steel 304 biomedical devices. Res. Dev. Mater. Sci., **8**(2) (2018) 000683. https://doi.org/10.31031/RDMS.2018.08.000683.
[35] P. Ali, R.S. Walia, Q. Murtaza and R.M. Singari, Characterisation of finished surface through thermal additive centrifugal abrasive flow machining for better surface integrity. Indian J. Eng. Mater. Sci., **29** (2022) 29–44.

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

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

Fusion of Transformer Model and Skeleton Detection Model for Abnormal Human Activity Detection with Transfer Learning

Akhilesh Kumar Verma*
Computer Science and Engineering
Delhi Technological University
Delhi, India
emailakhilesh.verma@gmail.com

Abhishek Soren
Computer Science and Engineering
Delhi Technological University
Delhi, India
abhisheksoren95@gmail.com

Ankit Kumar Shrivastav
Computer Science and Engineering
Delhi Technological University
Delhi, India
akumar33951@gmail.com

Sanjay Kumar
Computer Science and Engineering
Delhi Technological University
Delhi, India
sanjay.kumar@dtu.ac.in

2023 IEEE World Conference on Applied Intelligence and Computing (AIC) | 979-8-3503-1006-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/AIC57670.2023.10263965

***Abstract*—Abnormal activity detection is an area of study that focuses on finding abnormal activities like theft, vandalism, abuse, etc in video and CCTV footage. In this paper, a combination of skeleton based and transformer-based approach is used. The skeletons are used to enrich the motion of human bodies in the video. The transformer is used to model dynamic convolutions which helps to overcome the problems associated with fixed size kernels. To preserve the background information the architecture also uses feature extraction. The methodology used in the paper, also proposed a way to use transfer learning in the transformer network. The proposed framework provides excellent results outperforming many of existing methodologies. ROC (AUC score) attained by the proposed method for some classes of actions are up to 97.71%, which is even higher than current state of art approaches. The average ROC (AUC score) of our method is better than many of the approaches which are considered in the paper.**

***Index Terms*—Skeleton Detection, Abnormality Detection, Surveillance, UCF-Crime, Transformer, Encoder**

I. Introduction

Citizens are always concerned about public safety in today's world where crimes happen often. Crime rates drastically rise, particularly at night. A significant step to eliminate these and further contribute to improving public safety is Automatic Abnormal Human Activity detection. This solution can help to facilitate police operations and benefit in security surveillance system. It is remarkable how much video content is freely available on the Internet, and how easy it can be to analyze it. Abnormal Human Activity Detection is one such area where studying and analyzing videos can help prevent various crimes. This field, however, is still in its infancy and does not yet have a concrete solution that is reliable when it comes to surveillance systems. This is an area of open research that has the potential for a great deal of improvement and dedicated investigation. The method developed has applications in monitoring the videos of CCTV cameras, detecting the kind of action that may or may not be considered abnormal. This is needed to prevent accidents and track criminals more easily. The common criminal activities captured in public datasets include abuse, snatching, fights, burglary, robbery, explosion, road rage accidents, etc.

We have evaluated a few of the many articles that have been published in this field to determine gaps in the research [1] and the possible strategies we can use to fill those gaps and get the best results. In order to identify the anomalies, the current techniques employ fixed-size networks that collect local spatial or local temporal information from the video.

The major contributions of this research work are listed under:

1) Fusion of the Skeleton extraction and the Transformer encoders: In this paper, we annotated the skeletons detected in each frame on the frame. This way we preserved the skeleton information and using the same encoders to perform the feature extraction as the one used for RGB frames.
2) Using transfer learning to reuse the existing transformer encoders: The final layers are modified to reuse the existing transformer encoders. In our proposed methodology, we explained a suitable way to modify the dimensions of final layers to reduce the training required.
3) Using CNN and Transformer encoder Framework with Residual Connections: We have proposed a Transformer Encoder Network which uses CNN and Encoders alternately to perform feature extraction each layer uses a residual connection. There are 5 such layers in our proposed methodology. After this proposed network is a Sparse Object Interaction Transformer. [2]

979-8-3503-1006-1/23/$31.00 ©2023 IEEE
DOI: 10.1109/AIC.2023.93

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 06:04:10 UTC from IEEE Xplore. Restrictions apply.

II. Literature Review

This is an open field of active study so lots of work has been done in this field. The works can be categorized for easy understanding.

A. Deep Action Recognition

C3D used deep 3D ConvNets trained on large supervised video dataset for spatiotemporal learning [3]. Authors in [3] finds the following conclusions:

1) 3D ConvNets outperforms 2D ConvNets for spatiotemporal feature learning
2) homogeneous architecture with $3 \times 3 \times 3$ kernels in all layers are among best performing for 3D ConvNets.

Even with simple architecture [3] achieved 52.8% accuracy on UCF-101 dataset which was comparable to existing state-of-art work.

I3D re-examines the state-of-art architectures on Kinetics Human Action dataset [4]. Authors proposed a classification model that converts 2D into 3D by processing multiple frames simultaneously. Models after pre-training on Kinetics achieved appreciable performance improvement, approaching 80.90% on HMDB-51 and 98.00% on UCF-101.

ResNet [5] proposed 3D CNNs derived from ResNets regarding better action recognition. Authors examined 3D ResNet on Kinetics and ActivityNet datasets. This model outperforms C3D, achieved 68.9% accuracy.

In recent years, Deep Learning based techniques have become pivotal in solving many real-life problems like various problems related to computer vision, medical imaging [6], [7], social networks [8], and others (table 1).

B. Multi-Scale CNNS and Dynamic Convolution

For the video tasks, [9] proposed various timescale inference pooling to observe videos across various timescales. TEA extends the Res2Net block with temporal modelling capability. The concept of a dynamic filter was proposed in [10]. Later, several works [1], [11] in image tasks generated weights for averaging the convolutional kernels. [12] generalizes the idea to predict the averaging weights for video data. It is like the attention to be given to each component. As the time dependent kernel is sent across all different channels it further limits its efficiency and performance.

C. Object Interaction Models

LSTM and Graph Convolutional Networks (GCNs) are used to deduce relationship of space and time. Later sparse graphs connecting the entities in the time and space. [13] used the graph based LSTM on social networks. There can be local Neural Networks which make use of object detectors to locate the parts of interest. The other is Non-Local Neural Network which study pairwise interaction in all features.

D. Transformers

Many neural networks appoint transformers for processing vector sequences, encoding and decoding them as well. Recently Transformers are being used in deep action recognition. Action Transformer [14] proposed an architecture using space and time dependencies of a person to classify their actions. AcT [15] proposed a simple, fully self-attentional architecture for deep action recognition that outperforms many existing advanced architectures.

LSTR presents a temporal modeling approach for online action recognition [16]. To model prolonged sequence data [16] used long-short term memory mechanism. On various benchmarks it yielded efficient and much better performance [16]. These researches prove that transformers can be deemed effective in recognizing and classifying human actions from CCTV footage. The auto encoders are finding application in all fields like, [17] used to detect community of a person.

E. Skeleton Detection

Skeletal information illustrating the human movement in the videos is privacy-protecting and can conquer some of the problems constituted by appearance-based features. The study of skeletal graphs for human action recognition is ongoing and an active field of interest of many researchers.

MPED-RNN [18] proposed method using dynamic skeleton features for anomaly detection in surveillance video. They decoupled the skeletal features into two parts: global body and local body movement, which synergistically interact with spatiotemporal models to precisely identify irregular human activity. CycleGAN [19] proposed a framework using a cyclic consistent generative adversarial network for anomaly detection in which they used dynamic skeleton features as network constraints to attenuate the imprecision of feature extraction methods of generative adversarial network.

III. Methodology

This section discusses the proposed methodology for Abnormal Human Activity Detection in detail. The basic architecture is a combination of two techniques the skeleton based approach and the transformer based encoders. Similar to Siamese Network, fusion computes the encoding for both types of the frames simultaneously; however unlike Siamese Network the weights are not shared between the two branches. The model is similar because the Skeleton based and the RGB network have similar characteristics but the skeleton network has more features so different set of weights allows to give more attention to the features. The RBG frames compensate for the background information which is not addressed in the skeleton based frames.

The model can be divided into 4 modules, viz, Segmented Inputs, Parallel Architecture, Network of Transformer Encoders, Classifier function and Loss function. The following sub-sections describe each of the four components of the model in detail.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 06:04:10 UTC from IEEE Xplore. Restrictions apply.

TABLE I
COMPARATIVE TABLE OF EXISTING RESEARCH TECHNIQUES

Ref. No.	Author(s)	Highlights of the technique used
[20]	Guo et al. (2014)	This paper provides a model with space and time focal points with the help of joint positions of skeletons and RGB in videos.
[21]	Boussaid et al. (2017)	This paper proposes a novel approach for transforming skeleton graphs. For improving the accuracy of the model multiple learning tasks are employed.
[22]	Lin et al. (2018)	This paper talks about the use of space and time skeleton graph configurations that are used for identifying various actions in real time.
[23]	Heeseung et al. (2020)	This paper introduces a newer model using MS module for video motion detection which can be easily trained on any video architecture with back-propagation.
[24]	Lianyu et al. (2020)	This discusses creating skeleton nodes from images and establishing connections between them for recognizing various actions.
[25]	Gao et al. (2022)	This paper proposes a CNN model capturing motion sequences in videos and modeling the space-time dependencies.
[26]	Tian et al. (2022)	An analysis of video adaptive kernels generated from video datasets is presented in this paper, along with a model of the interactions between objects with the help of a transformer.

A. Segmented Input

For better memory management and efficiency instead of using all the frames, the frames are sub divided into groups of fixed number. This provides dual advantages first it consumes lesser memory allowing bigger batches to be processed and at the same time, it guarantees fixed input dimensions. In this implementation, 8 segments per video are used. Despite these advantages, to preserve the motion details of the video four additional frames are used per segment these are calculated as the difference from the corresponding segment to encode the motion details. This gives 8 times 5, i.e., 40 images per video. Each image has 3 channels hence the input sequence contains 120 channels in total. In addition, each frame is resized to size 224×224 to ensure uniformity. Some more processing is performed on each segment to get the most of the features into the input tensor.

This can be optimized since the 4 frames per segment are mostly sparse matrices, which can be compressed to fit in smaller sizes and create a more detailed representation. The benefit of this is that parameters will reduce and eliminate the problem of over- fitting associated with sparse inputs. This is typically done in the embedding layer in natural language processing.

B. Parallel Architecture

"Fig. 1" shows the broad architecture of the network. The two branches go through identical steps except in one branch there is an additional skeleton extraction step. The block diagram indicates that the skeleton extraction is done in the model, however, since it is the first step in the branch it can be performed separately and the model can be defined identically for both branches. This helps in decreasing the training time as the same processing needs to be repeated every time the same video is used. After the operation on the two branches is complete the two branches are merged back using concatenation and then a classifier is used in the end.

As an implementation note, the model used in two branches can be a pre-trained model. The original classifier is detached to serve as the classifier in the end. The classifier is replaced with another layer to reduce the dimension to half of the original network. This allows concatenation of outputs of the two branches so that the dimensions match again and the model requires a lesser end to end training. This hack is also incorporated in this research to speed up training to acceptable results however there is a downside to this in the case of a branch with skeletons as they are different from the RGB frames so might boil down to local maxima instead of a global minimum for the loss.

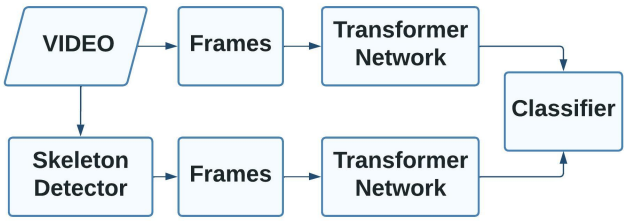


Fig. 1. Flow diagram for the network. This is the complete architecture of our proposed methodology. Every component is shown as block diagram. The Skeleton detector block is applied as pre-processing to save computation.

C. Network of Transformers

There are a few transformer networks that are put in between the backbone CNN. In the transformer, CNN of various kernel sizes are performed and the attention mechanism is used to get the final output. This method fuses the power of multiple convolutions, unlike the inception network we need not stack everything but can still get the power of the same with lesser memory requirements. Many such attention networks are stacked together to get different mechanisms. "Fig. 2" shows the architecture of the transformer network used by us in this research paper. The CNN and transformer encoders are used alternatively. Five such layers are used. The CNN also have residual connections in between to ensure the gradients don't vanish.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 06:04:10 UTC from IEEE Xplore. Restrictions apply.

There is another transformer at the very end which extracts features from interactions of the segments. This is done by applying some self-attention to the segments. This is the last transformer in our network. The numbers of the two types of transformers used in the model is a hyper-parameter to tune. This transformer takes the channels as the input and uses global-attention on them to get the next layer of output.

D. Classifier and Loss Function

A neural network classifier is used for the integrated feature space and finally to train the model categorical cross entropy loss is calculated and employed. The amalgamation of experiments and trials are stated below.

IV. Datasets and Metrics

A. Datasets

The datasets used are listed below:

1) UCF-Crime
   It is a large-scale video dataset of 128 hours consisting various human abnormal behaviors distributed in thirteen different classes including fights, thefts, blasts, etc. The recordings include data from various aspects making it more genuine and representative in real life.
2) Kinetics 400, 700
   This large-scale dataset focuses on a wide range of human actions linked from 650000 high resolution YouTube videos. It classifies them into 400 (kinetics 400) or 700 (kinetics 700) different classes where each clip is around 10s long. The actions can range from hugging, shaking hands to playing games, eating and many more.

B. Metrics

There are two types of task – anomaly detections and action detections. The anomaly task used ROC score in most of the literature while action detection uses top1 and top5 scores. Some state of the art results of the time are listed in the "Table II". I3D outperforms all the approaches including ours. But our method whereas our method used transfer learning which also beats four of the other methods.

TABLE II
ACCURACY OF ACTION DETECTION ON UCF 101 AND HMDB51 DATASETS COMPARED WITH OUR PROPOSED METHODOLOGY.

Technique	UCF 101 (%)	HMDB51(%)
IDT	85.9	57.2
Two Stream [27]	88.0	59.4
TSN RGB	85.7	51.0
C3D (3 Net) [5]	85.2	NA
I3D (RGB) [4]	95.1	74.8
3D Resnet 50	89.3	61.0
MSRF Net [28]	91.0	59.9
Our Method	88.1	NA

ROC is used more in comparison to the Accuracy in the Abnormal Activity Detection so we have compared the results of our approach in "Table III". Apart from the overall ROC AUC "Table IV" provides the classwise ROC AUC score for each of the class. We observe that for some classes ROC AUC is as high as 97.71%. On the other hand, for Vandalism the ROC AUC turned out to be very low i.e., 32.83%. "Fig. 3 show the ROC of all the classes.

TABLE III
TABLE SHOWING ROC (AUC SCORES) FOR VARIOUS TECHNIQUES ON UCF CRIME DATA SET COMPARED WITH OUR PROPOSED METHODOLOGY.

Technique	ROC AUC Score (%)
MGFN [29]	86.98
S3R [30]	85.99
WSAL [31]	85.38
MIST [32]	82.30
MILR [33]	76.67
Our Method	83.61

TABLE IV
ROC SCORES OF INDIVIDUAL CLASSES AND THE AGGREGATE SCORE OBTAINED USING OUR PROPOSED METHODOLOGY. MICRO AVERAGE IS USED TO CALCULATE THE AGGREGATE SCORE FOR OUR METHODOLOGY.

Class	ROC AUC Score (%)
Abuse	97.712
Arrest	45.792
Arson	90.083
Assault	95.500
Burglary	90.196
Explosion	62.000
Fighting	88.042
Normal	94.605
Road Accident	72.848
Robbery	83.106
Shooting	66.083
Shoplifting	79.667
Stealing	89.435
Vandalism	32.833
Micro Average	83.613

V. Result Analysis

We have implemented the proposed methodology using PyTorch framework. The code divided into modules as explained in the Methodology Section. The skeleton annotation of every video was performed as a pre-processing step because that part used a pre-trained skeleton detector. This helped us train rapidly as need not to extract skeleton from same video in every epoch of training. We have seen that the proposed techniques perform decently. The model was tuned for 80 epochs on the data set. The optimizer used for training was Adam optimizer with weight decay of 3×10^{-4} and momentum of 0.9, maximum learning rate of 0.01.

The "Table IV" shows that for class "Abuse", the ROC (AUC Score) obtained by our proposed methodology is close to 97.7%. However, for class "Vandalism" is 32.833%. This is expected because act of Vandalism is similar to act of Shooting and Explosion as they need similar human actions that explains lower AUC scores of the three classes compared to other classes. However, this needs further investigation. The overall aggregate ROC (AUC) is calculated as micro average of all the points. The scikit metric implementation is used with

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 06:04:10 UTC from IEEE Xplore. Restrictions apply.

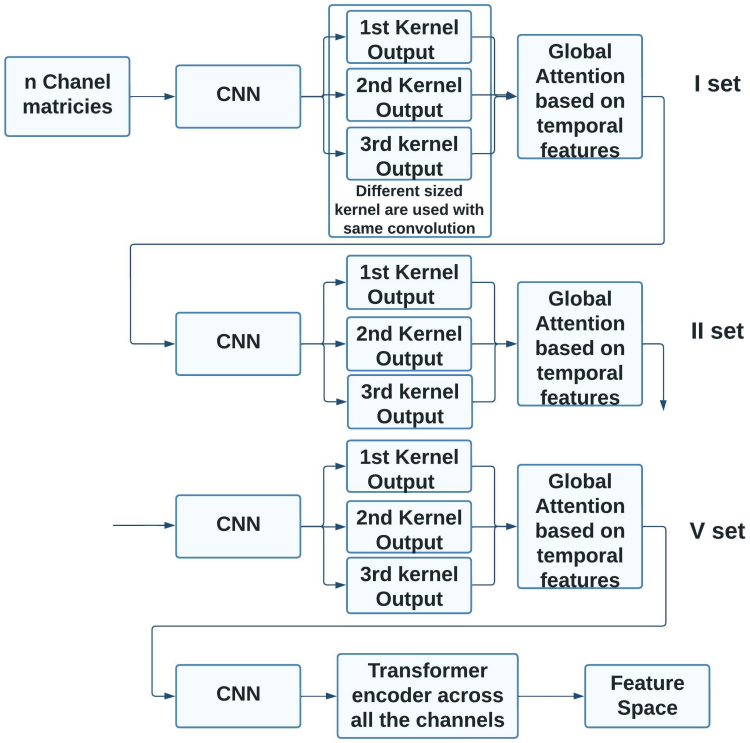


Fig. 2. Figure showing the architecture of the two transformer networks. This is our proposed transformer network which we used in this paper. The residual connections are not shown in this block diagram but are present in the implementation.

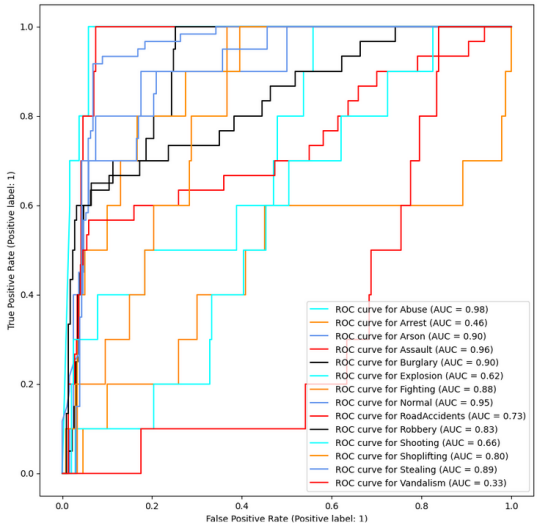


Fig. 3. ROC plots for each of the fourteen classes using our proposed methodology.

micro averaging and one-vs-rest option for multi class ROC. "Table II" compares our methodology in action classification task with multiple approaches and it demonstrates that our approach is better than many of the existing approaches.

VI. Conclusion

In this research paper, we presented a way of fusing the transformer based architecture with skeleton detection architecture. We have also highlighted to take advantage of the transfer learning as much as possible. Using our proposed methodology, we are able to get fairly good results which outperform some state of the art techniques. This paper provides the comparison of results among various classes as well. The reason of disparity of results among the classes is to be explored. The system presented in the paper is mature enough to be deployed to analyze the CCTV footage and alert the human operator only when it detects the abnormality in the video. This will help security personnel. The paper also highlights the modifications needed to incorporate the transfer learning, by explaining how the feature space is to be modified to use the same classifiers and transformer weights. The low ROC for certain classes pose a challenge for us and requires more analysis. This paper demonstrates an end to end model that takes advantage from transfer learning hence suitable for devices with lower computation power however that may be a source of bias apart from the similarity of classes highlighted in the Result analysis section.

References

[1] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 030–11 039.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 06:04:10 UTC from IEEE Xplore. Restrictions apply.

[2] R. Morais, V. Le, S. Venkatesh, and T. Tran, "Learning asynchronous and sparse human-object interaction in videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 041–16 050.

[3] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[4] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[5] K. Hara, H. Kataoka, and Y. Satoh, "Learning spatio-temporal features with 3d residual networks for action recognition," in *Proceedings of the IEEE international conference on computer vision workshops*, 2017, pp. 3154–3160.

[6] S. Kumar and A. Mallik, "Covid-19 detection from chest x-rays using trained output based transfer learning approach," *Neural processing letters*, pp. 1–24, 2022.

[7] S. Kumar, A. Mallik, A. Kumar, J. Del Ser, and G. Yang, "Fuzz-clustnet: Coupled fuzzy clustering and deep neural networks for arrhythmia detection from ecg signals," *Computers in Biology and Medicine*, p. 106511, 2023.

[8] S. Anand, A. Mallik, and S. Kumar, "Integrating node centralities, similarity measures, and machine learning classifiers for link prediction," *Multimedia Tools and Applications*, vol. 81, no. 27, pp. 38 593–38 621, 2022.

[9] C. Zhang, Y. Zou, G. Chen, and L. Gan, "Pan: Towards fast action recognition via learning persistence of appearance," *arXiv preprint arXiv:2008.03462*, 2020.

[10] X. Jia, B. De Brabandere, T. Tuytelaars, and L. V. Gool, "Dynamic filter networks," *Advances in neural information processing systems*, vol. 29, 2016.

[11] B. Yang, G. Bender, Q. V. Le, and J. Ngiam, "Condconv: Conditionally parameterized convolutions for efficient inference," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[12] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[13] S. Kumar, A. Mallik, and B. Panda, "Influence maximization in social networks using transfer learning via graph-based lstm," *Expert Systems with Applications*, vol. 212, p. 118770, 2023.

[14] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman, "Video action transformer network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 244–253.

[15] V. Mazzia, S. Angarano, F. Salvetti, F. Angelini, and M. Chiaberge, "Action transformer: A self-attention model for short-time pose-based human action recognition," *Pattern Recognition*, vol. 124, p. 108487, 2022.

[16] M. Xu, Y. Xiong, H. Chen, X. Li, W. Xia, Z. Tu, and S. Soatto, "Long short-term transformer for online action detection," *Advances in Neural Information Processing Systems*, vol. 34, pp. 1086–1099, 2021.

[17] S. Kumar, A. Mallik, and S. S. Sengar, "Community detection in complex networks using stacked autoencoders and crow search algorithm," *The Journal of Supercomputing*, vol. 79, no. 3, pp. 3329–3356, 2023.

[18] R. Morais, V. Le, T. Tran, B. Saha, M. Mansour, and S. Venkatesh, "Learning regularity in skeleton trajectories for anomaly detection in videos," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 11 996–12 004.

[19] Z. Fan, S. Yi, D. Wu, Y. Song, M. Cui, and Z. Liu, "Video anomaly detection using cyclegan based on skeleton features," *Journal of Visual Communication and Image Representation*, vol. 85, p. 103508, 2022.

[20] Y. Zhu, W. Chen, and G. Guo, "Evaluating spatiotemporal interest point features for depth-based action recognition," *Image and Vision computing*, vol. 32, no. 8, pp. 453–464, 2014.

[21] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3288–3297.

[22] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[23] H. Kwon, M. Kim, S. Kwak, and M. Cho, "Motionsqueeze: Neural motion feature learning for video understanding," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*. Springer, 2020, pp. 345–362.

[24] L. Hu, L. Feng, and S. Liu, "Hfnet: a novel model for human focused sports action recognition," in *Proceedings of the 3rd International Workshop on Multimedia Content Analysis in Sports*, 2020, pp. 35–43.

[25] B. Zhang, Q. Wang, Z. Gao, R. Zeng, and P. Li, "Temporal grafter network: Rethinking lstm for effective video recognition," *Neurocomputing*, vol. 505, pp. 276–288, 2022.

[26] Y. Tian, Y. Yan, G. Zhai, G. Guo, and Z. Gao, "Ean: event adaptive network for enhanced action recognition," *International Journal of Computer Vision*, vol. 130, no. 10, pp. 2453–2471, 2022.

[27] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Two stream lstm: A deep fusion framework for human action recognition," in *2017 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2017, pp. 177–186.

[28] A. Srivastava, D. Jha, S. Chanda, U. Pal, H. D. Johansen, D. Johansen, M. A. Riegler, S. Ali, and P. Halvorsen, "Msrf-net: a multi-scale residual fusion network for biomedical image segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 5, pp. 2252–2263, 2021.

[29] Y. Chen, Z. Liu, B. Zhang, W. Fok, X. Qi, and Y.-C. Wu, "Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 387–395.

[30] J.-C. Wu, H.-Y. Hsieh, D.-J. Chen, C.-S. Fuh, and T.-L. Liu, "Self-supervised sparse representation for video anomaly detection," in *European Conference on Computer Vision*. Springer, 2022, pp. 729–745.

[31] H. Lv, C. Zhou, Z. Cui, C. Xu, Y. Li, and J. Yang, "Localizing anomalies from weakly-labeled videos," *IEEE transactions on image processing*, vol. 30, pp. 4505–4515, 2021.

[32] J.-C. Feng, F.-T. Hong, and W.-S. Zheng, "Mist: Multiple instance self-training framework for video anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 009–14 018.

[33] S. Dubey, A. Boragule, and M. Jeon, "3d resnet with ranking loss function for abnormal activity detection in videos," in *2019 International Conference on Control, Automation and Information Sciences (ICCAIS)*. IEEE, 2019, pp. 1–6.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 06:04:10 UTC from IEEE Xplore. Restrictions apply.

Optical and Quantum Electronics (2023) 55:1174
https://doi.org/10.1007/s11082-023-05462-9

Highly sensitive twin core photonic crystal fiber for hazardous cancer cell detection in THz frequency regime

Vishal Chaudhary[1] · **Sonal Singh**[1]

Received: 22 July 2023 / Accepted: 18 September 2023 / Published online: 17 October 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract
This article proposes a twin-core photonic crystal fiber (TC-PCF) with a hexagonal lattice that is designed to identify different cancers at an early stage using the refractive index of cancerous and normal cells in the terahertz (THz) frequency regime. The central elliptical hole of the TC-PCF, which is comparatively smaller than the rest of the air holes, is used in the proposed method to infiltrate liquid samples of cancer and normal cells at 80% and 30–70%, respectively. Subsequently, the model undergoes simulation across the frequency range of 0.7–0.8 THz in order to assess its optical characteristics. The finite element method (FEM) is used to evaluate the birefringence, transmission curve and coupling length of the TC-PCF to identify cancerous and normal cells. The transmitted spectral shift demonstrates that the proposed design provides a high optimal sensitivity of 1,016,666.67 nm/RIU, 1,521,428.57 nm/RIU, 1,578,571.42 nm/RIU, 1,370,000 nm/RIU, 1,671,428.57 nm/RIU and 1,507,142.85 nm/RIU for Hela, Jurkat, PC12, Basal, MDA-MB-231 and MCF-7 cancerous cells, respectively. These sensitivities are the highest that have been reported thus far, to the best of our knowledge. Additionally, the impact of changing air holes diameter on the proposed model's sensitivity is examined. Thus, the proposed TC-PCF sensor may be used to identify various kinds of cancer using this low-cost and simple detection technique.

Keywords Sensitivity · Twin core PCF · Cancerous cell · Terahertz · Wavelength dip · Transmission curve

1 Introduction

Photonic crystal fiber (PCF) has emerged as a promising potential for detection of gas and chemical due to its flexible architecture, low loss, large effective mode area, high birefringence and low dispersion (Leon and Disha 2021). PCF sensors are appropriate for biosensing applications since they do not require a lot of mechanical components (Mahfuz et al. 2019; Sarker et al. 2023). They have recently been used in the detection of diseases such as

✉ Sonal Singh
sonalsingh@dtu.ac.in

[1] Department of Electronics and Communication Engineering, Delhi Technological University, Shahbad Daulatpur, Main Bawana Road, New Delhi, Delhi 110042, India

oncovirus (Mollah et al. 2020), cholesterol (Soylemez et al. 2015), DNA (Sun et al. 2014) and glucose (An et al. 2017; Rahaman et al. 2022a, b; Rahaman et al. 2019). More than 9.6 million people died worldwide in 2018 as a result of oncovirus (World health organization, Cancer 2018), a virus that has been related to the emergence of cancer. The smaller size of the virus and how it diverges around the cell region define the effect of cancer existence. Early detection is essential for lowering mortality rates because cancer is one of the main causes of death around the globe. As a result, researchers are concentrating on finding new methods for detecting cancer cells before they spread, and PCF technology has emerged as an effective approach. Leland C. Clark used a biosensor to monitor blood glucose levels for the first time in 1962, as reported (Clark 1987). Various methods to detect cancerous cells have been developed throughout time, including microfluidic devices (Hajba and Guttman 2014), immunocytochemistry (Li et al. 2013), electrochemical methods (Li et al. 2010), and Raman spectroscopy (Shupeng Liu et al. 2014). Particularly, spark fused tapers have been proposed for breast cancer detection using Raman spectroscopy (Neugebauer et al. 2010), allowing early detection of cancerous cells at the molecular level. Using a plasmonic optical fiber, Ribaut et al. (2017) presented an immunosensor for lung cancer detection in 2017, while Ayyanar et al. presented a novel PCF based sensor for detection of cancerous cell in 2018. The PCF sensor required cancerous cells to be injected into 0.6 µm air holes, but in fact, this was found to be a challenging operation (Ayyanar et al. 2018). A nanocomposite material-coated photonic crystal based sensor was also presented that year, but it had a low sensitivity of just 43 nm/RIU (Ramanujam et al. 2019).

A number of applications, including astrophysics, genetics, biomedical research, spectroscopy, sensing and telecommunications, are finding the terahertz (THz) band to be more attractive as a suitable frequency range (Ahmed et al. 2019; Bulbul et al. 2020a, b; Islam et al. 2019a, b; Kaur and Singh 2019; Mou et al. 2020; Podder et al. 2019; Poonam and Preeta 2015; Rifat et al. 2015). This frequency band, which covers a broad range from 0.1 to 10 THz, offers a unique range of opportunities for different fields of study and technological advancements. Researchers have successfully developed chemical and biosensors that operate in the THz range using PCF technology. These sensors can detect on a wide range of chemical and biological elements, including salinity (Vigneswaran et al. 2018), cholesterol (Rahman et al. 2020), blood components (Ahmed et al. 2019), and adulterated fuel (De et al. 2019). PCF sensors, as compared to traditional fiber sensors, provide greater diversity and standard values for optical parameters, faster detection systems, and adjustable optical parameters with careful modification of the cladding and core hole of the PCF. PCF sensors also provide a number of advantageous characteristics, including improved sensitivity, low confinement loss, light weight and durability. The potential applications in chemical and biological sensing of PCF-based sensors have expanded due to the broad range of capabilities they provide. These include using surface plasmon resonance-based biosensors (Esfahani Monfared 2020; Monfared et al. 2021; Monfared 2020), investigating RNA and DNA (Fischer et al. 2005, 2002; Roitberg et al. 2000), identifying cancerous cells (Cheon et al. 2017; Rahman et al. 2016; Yu et al. 2012), categorizing and detecting alcohol (Arik et al. 2013; Eid et al. 2021; Paul et al. 2020a, b; Paul et al. 2017; Rahman et al. 2019; Sultana et al. 2018; Iqbal et al. 2020), and detecting formalin (Bulbul et al. 2020a, b; Huq Arif et al. 2019). The potential applications in chemical and biosensing of PCF-based sensors have increased as a result of these characteristics.

The proposed work demonstrates the design of a rapid cancer detection method based on a TC-PCF architecture in the terahertz (THz) frequency regime. Among the cancer cell types identified are those that cause blood cancer (Jurkat), cervical cancer (HeLa), cancer of the adrenal glands (PC12), skin cancer (Basal) and two forms of breast cancer

(MDA-MB-231 and MCF-7). The FEM-based detection method assumes that the concentration of cancerous cell is to be 80% in the form of liquid. The finite element approach is used in a specifically designed PCF to examine the mode coupling of twin cores as part of the proposed method for early detection of different cancers. A significant difference was observed by comparing the changes in birefringence, coupling properties, and transmission between cancerous and normal cells using different refractive indices. The shift of the transmitted spectrum's largest wavelength dip is used to evaluate the sensitivity of the proposed TC-PCF sensor. This shows that a simple TC-PCF-based sensor could be used for the early detection of different cancerous cells.

2 PCF sensor design

The cancerous and normal cells, which is inserted within the PCF elliptical central hole, separates the two solid cores that make up the proposed TC-PCF, which is illustrated in Fig. 1a. On the base material, the circular air holes are arranged in a hexagonal pattern with a pointed orientation. The air holes have a diameter (D) of 470 µm and a pitch (Λ) of 500 µm between any two adjacent air holes, which maintains the PCF over all air filling fraction (AFF) of 0.94. Various polymer materials, including TOPAS, ZEONEX and PMMA, have found applications as fiber materials in operations within the THz regime (Kanmani et al. 2019; Rahaman et al. 2022a, b; Jibon et al. 2021). Among these options, TOPAS stands out as the most prominent choice. Notably, TOPAS exhibits a material absorption coefficient of 0.2 cm^{-1} (Islam et al. 2018), which is lower compared to other polymer materials. Another distinctive feature of TOPAS glass is its constant refractive index of 1.53 across the frequency range of 0.1–1.5 THz (Markos et al. 2013; Paul and Ahmed 2019; Rahaman et al. 2021). This unique refractive index behavior influenced the selection of the operating frequency range (0.7–0.8 THz) for our proposed TC-PCF, ensuring that alterations in frequency do not disrupt the refractive index of TOPAS glass. Furthermore, TOPAS offers favorable attributes such as heightened heat resistance, a larger glass transition temperature coefficient, and insensitivity to humidity, underscoring its suitability for such applications (Topas advanced polymer 2021). The PCF is suitable for operation at THz frequencies with low confinement loss owing to these dimensions and a constant refractive index of 1.53 for the base material (TOPAS),

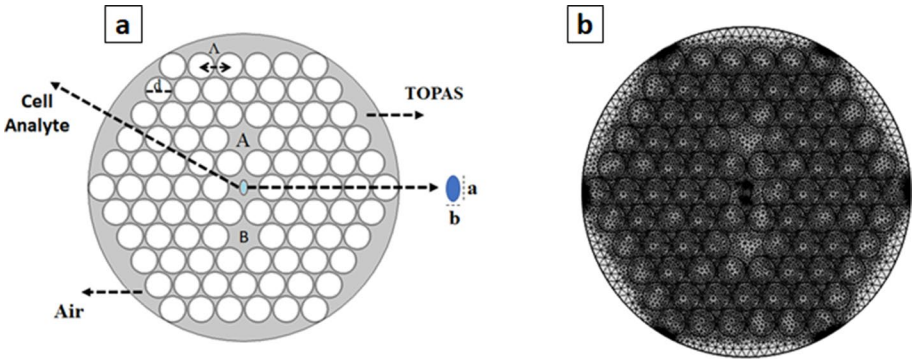


Fig. 1 **a** Cross-sectional view and **b** geometrical meshing statistics of proposed twin-core PCF sensor

offering flexibility in fiber manufacture (Islam et al. 2019a, b; Paul et al. 2019, 2018). FEM based COMSOL Multiphysics software is used to simulate and design the proposed PCF construction. There are two cores in the PCF named as Core A and Core B. As shown in Fig. 1a, these cores are created by vertically eliminating two air holes that are located in the middle of the PCF. The elliptical hole in the center of the PCF is used for inserting a normal and cancerous cell to be detected. The major-axis of the elliptical hole in the proposed configuration is a = 260 µm, the minor axis, b = 140 µm is considered. According to Fig. 1b, the geometrical meshing statistics required 3173 edge elements, 360 vertex elements and 24,222 triangular elements for the proposed PCF. The R.I. of the surroundings has a significant impact on the way light propagates in the core region of PCF. Therefore, both normal and cancerous cells exhibit a noticeable shift in the coupling length, transmission curve and birefringence. The R.I. and liquid percentage (%) of both normal and cancerous cells are obtained from (Ayyanar et al. 2018; Jabin et al. 2019). Numerous fabrication methods are available, including extrusion, 3D printing, capillary stacking, stack and draw, and sol–gel. Circular and elliptical air holes are typically manufactured using capillary stacking, or stack and draw methods (Cubillas et al. 2013; Lataoui et al. 2023), sol–gel (Bise and Trevor 2005), while 3D printing (Ebendorff-Heidepriem et al. 2014) and extrusion (Ghazanfari et al. 2017) are more commonly employed for creating TC-PCFs with rectangular air holes. Therefore, for the proposed TC-PCF design that incorporates sol–gel, either capillary stacking or stack and draw techniques could be recommended for the fabrication process.

A general experimental setup to detect different cancerous cells is shown in Fig. 2. A single mode fiber (SMF) is used in the setup to transmit light into the TC-PCF from a light source like laser or LED. An optical spectrum analyzer (OSA) can be used to see the output signal via another SMF. Between the TC-PCF and the SMF, a splice-free connecting approach can be used to reduce coupling losses. The OSA can be used to examine the shift in wavelength of the transmission spectra to detect the presence of cancerous cells. This procedure offers a simple and effective way to detect different cancerous cells.

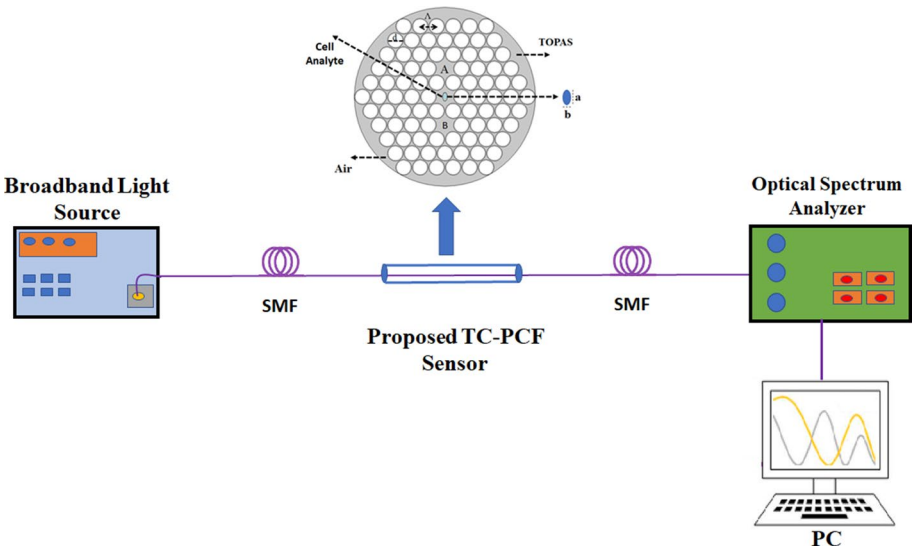


Fig. 2 General representation of the PCF sensing setup for cancer cell detection

According to coupling theory, the guided mode of a twin-core PCF contains four supermodes polarized in the X and Y directions that stand in for the even and odd modes at a frequency of 0.75 THz. The distribution of these four supermodes electric fields in the proposed PCF is illustrated in Fig. 3.

3 Results simulation and discussion

The variation of effective refractive index difference between even and odd mode (Δn_{eo}) for the x-polarized and y-polarized direction in Fig. 4 show how the effective refractive index difference for different normal (30–70%) and cancerous cell (80%) decreases as the frequency increases. A higher confinement of light within the core is implied by the decrease in Δn_{eo} with frequency seen in the proposed TC-PCF.

A birefringent material can cause light to be divided into two separate rays that travel in different directions when it is subjected to polarization. Birefringence is the main cause of the optical phenomenon, known as double refraction. To calculate the TC-PCF birefringence parameter, B, Eq. (1) is used.

$$B = \left| Re\left(n_{eff}^{x}\right) - Re\left(n_{eff}^{y}\right) \right| \tag{1}$$

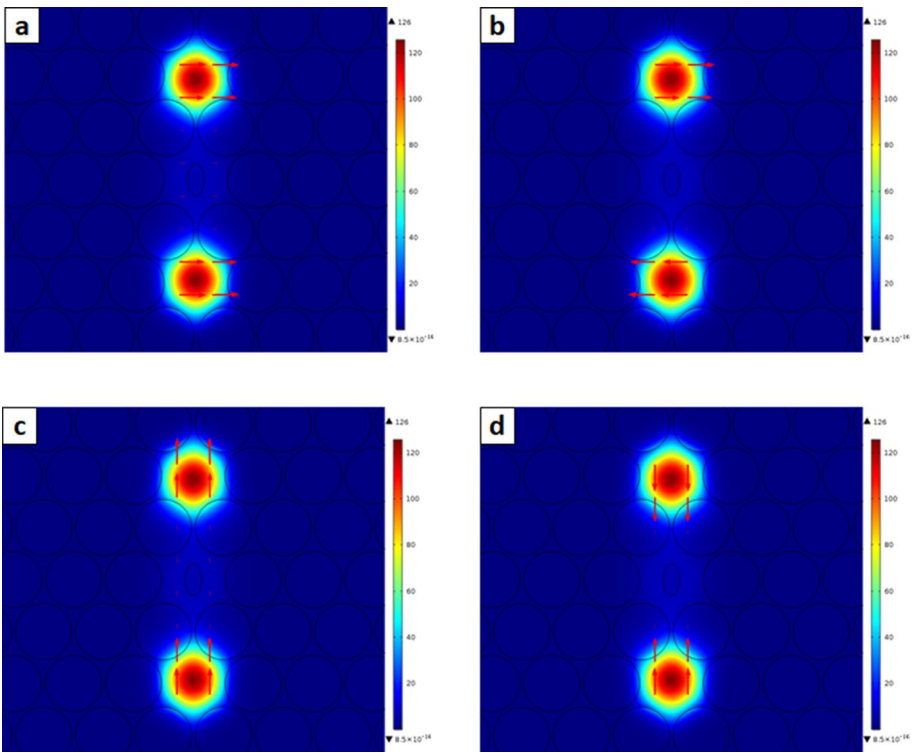


Fig. 3 Electric field distribution of **a** *x*-even, **b** *x*-odd, **c** *y*-even and **d** *y*-odd mode at a frequency of 0.75 THz

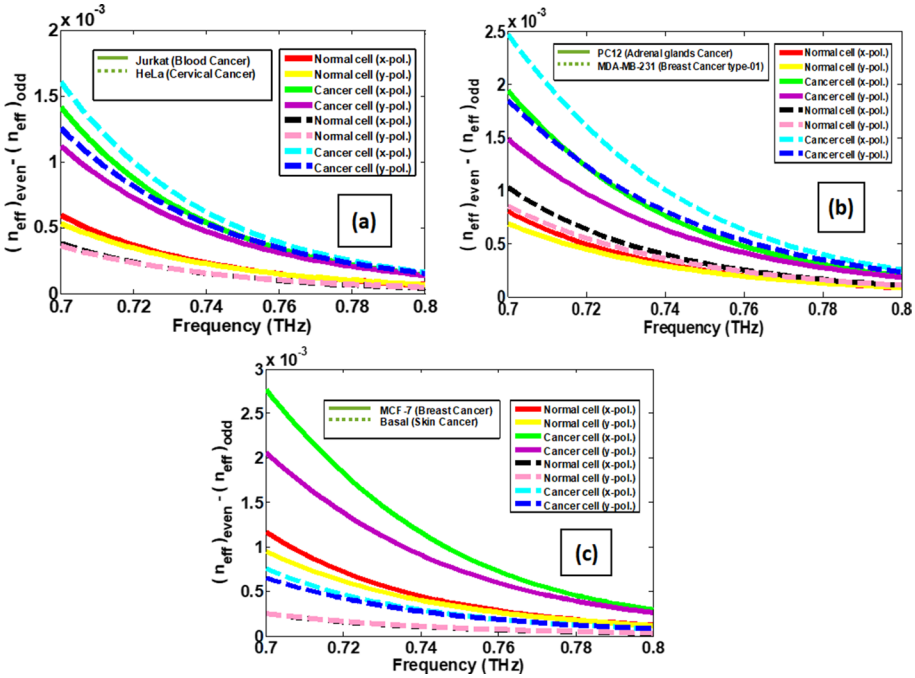


Fig. 4 Variation of Δn_{eo} with frequency for normal and cancerous cell **a** Hela and Jurkat **b** MDA-MB-231 and PC12 **c** Basal and MCF-7

where the n_{eff}^{x} and n_{eff}^{y} symbols stand for the effective refractive indices of the x-polarized and y-polarized modes, respectively (Jabin et al. 2019). The total amount of variations in birefringence with changes in frequency is shown in Fig. 5 for both cancerous (80%) and normal cell (30–70%). The maximum range of birefringence, which exhibits a downward trend, is between 0.0026 and 0.0015. The sensitivity and coupling length of the proposed geometry are directly correlated with its birefringence. The proposed sensor birefringence starts at 0.0026 and decreases linearly as the frequency increases until it achieves a low birefringence value of 0.0015.

The smallest length at which the maximum transmission of light can efficiently pass through the core for a PCF is known as the coupling length. This is applicable to both X and Y polarization and can be calculated by using Eq. (2).

$$L_c = \frac{\pi}{|\beta_e^i - \beta_o^i|} = \frac{\lambda}{2|n_e^i - n_o^i|} \tag{2}$$

where β_e^i and β_o^i respectively, are used to denotes the propagation constants of the even and odd supermodes with $i = x, y$ polarization, symbol λ represents the wavelength and the symbols n_e^i and n_o^i, respectively, denotes the effective refractive indices for the even and odd mode (Paul et al. 2020a, b). The coupling length gradually varies as the frequency changes for both normal cells (30–70%) and cancerous cells (80%), as seen in Fig. 6. The graph indicates that the maximum coupling length that can be achieved lies between 0.8435 and 6.013 m, with an upward curving trend.

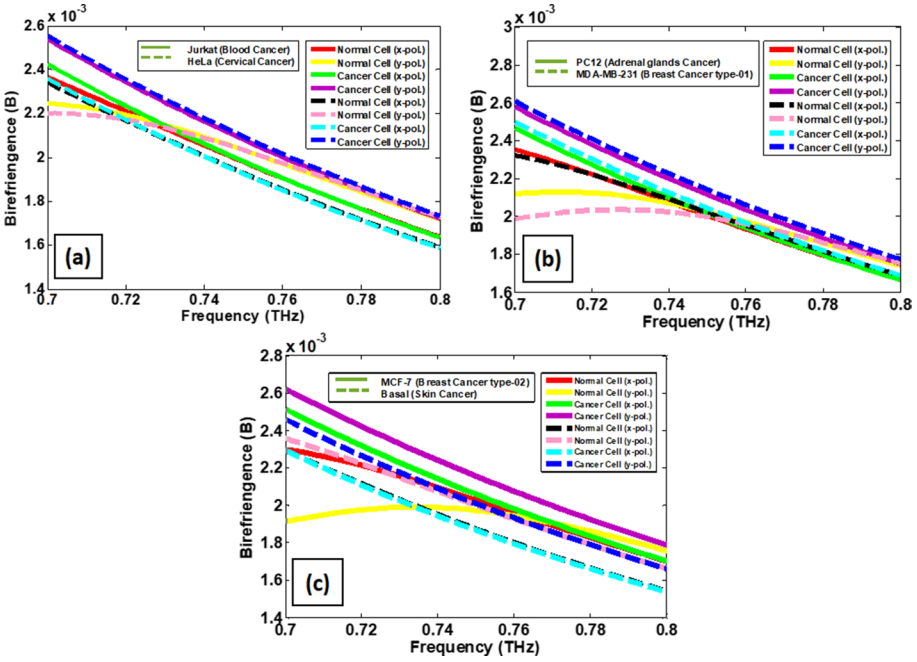


Fig. 5 Variation of birefringence with frequency for normal and cancerous cell **a** Hela and Jurkat **b** MDA-MB-231 and PC12 **c** Basal and MCF-7

The amount of optical power that is transferred between cores in a TC-PCF varies with the PCF length (L), according to the mode-coupling theory. The following expression can be used to describe this:

$$I_i(\lambda) = \sin^2\left(\frac{\pi}{\lambda}\Delta n_{i,eo}L\right), with i = x, y \tag{3}$$

where symbol $\Delta n_{i,eo}$, denotes the effective refractive index difference for the even and odd modes (Chaudhary and Singh 2023).

By incorporating both x and y polarized light, the TC-PCF displays the total intensity of the operating light as output. The total output can be determined by taking into consideration the unique properties of the orthogonal x and y polarized light (Aslam Mollah et al. 2020).

$$I(\lambda) = I_x + I_y \tag{4}$$

Equations (3) and (4) can be used to determine the total output intensity.

$$I(\lambda) = 1 - COS\left[\frac{\pi\left(\Delta n_x + \Delta n_y\right)L}{\lambda}\right].COS\left[\frac{\pi\left(\Delta n_x - \Delta n_y\right)L}{\lambda}\right] \tag{5}$$

The output intensity spectra for both normal and cancerous cells is illustrated in Fig. 7 using fibers with varied lengths and frequency between 0.7 and 0.8 THz. The maximum dips observed for different cancerous cells (80%) were at wavelengths of

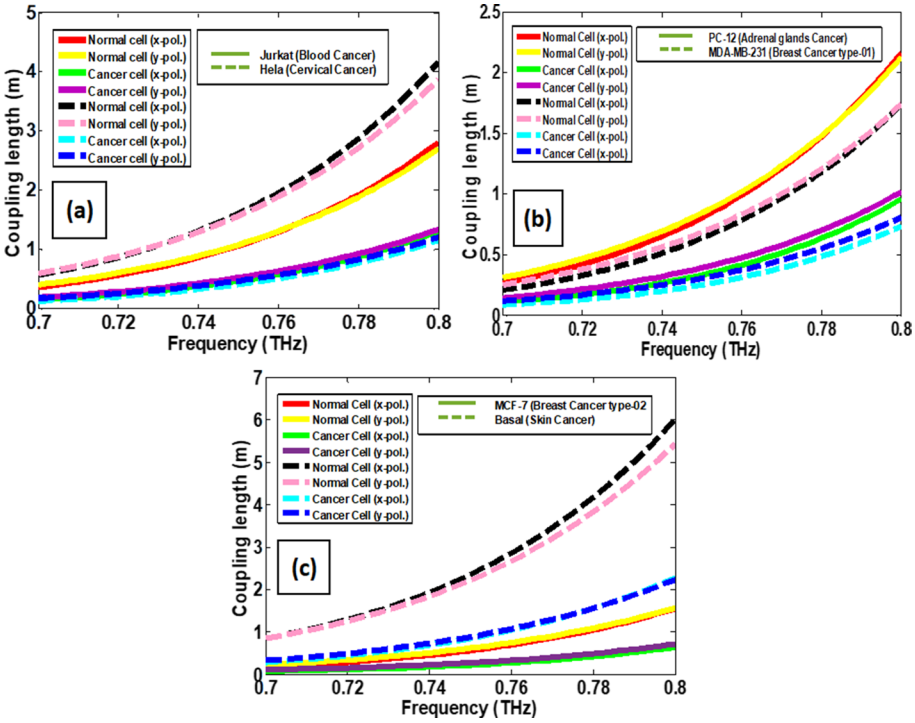


Fig. 6 Variation of coupling length with frequency for normal and cancerous cell **a** Hela and Jurkat **b** MDA-MB-231 and PC12 **c** Basal and MCF-7

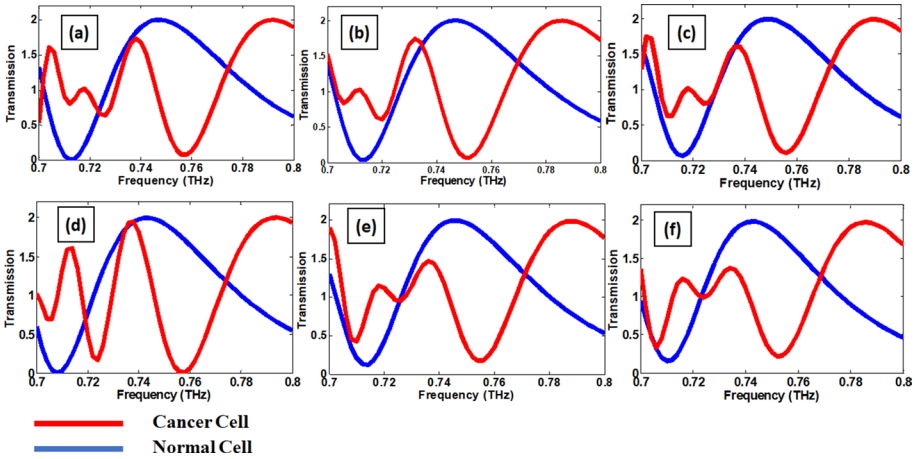


Fig. 7 Transmission curve with frequency for normal and cancerous cell **a** Hela **b** Jurkat **c** PC12 **d** Basal **e** MDA-MB-231 and **f** MCF-7

3.963×10^5 nm, 3.994×10^5 nm, 3.968×10^5 nm, 3.963×10^5 nm, 3.973×10^5 nm and 3.984×10^5 nm for Hela, Jurkat, PC12, Basal, MDA-MB-231 and MCF-7, respectively with corresponding fiber lengths of 1.0 m, 1.0 m, 0.8 m, 2.0 m, 0.6 m, 0.5 m,

respectively. In contrast, different normal cells (30–70%) showed their maximum dips at 4.207×10^5 nm, 4.207×10^5 nm, 4.189×10^5 nm, 4.237×10^5 nm, 4.207×10^5 nm and 4.195×10^5 nm for Hela, Jurkat, PC12, Basal, MDA-MB-231 and MCF-7, respectively with corresponding fiber lengths of 1.0 m, 1.0 m, 0.8 m, 2.0 m, 0.6 m, 0.5 m, respectively.

The wavelength shift in the transmission spectra caused by change in cancerous and normal cells is used to determine the proposed sensor's sensitivity. The proposed TC-PCF sensitivity for cancerous cell is defined as

$$S_\lambda\left(nmRIU^{-1}\right) = \Delta\lambda_{Dip}/\Delta n_i \tag{6}$$

where $\Delta\lambda_{Dip}$ represents the overall shift in maximum wavelength dip, Δn_i is the difference in R.I. for normal and cancerous cell (Jabin et al. 2019) and $\left(\lambda_{max\text{dip}} = \frac{\text{speed of light(c)}}{f_{max\text{dip}}}\right)$. The proposed sensor has a sensitivity of 1,016,666.67 nm/RIU, 1,521,428.57 nm/RIU, 1,578,571.42 nm/RIU, 1,370,000 nm/RIU, 1,671,428.57 nm/RIU and 1,507,142.85 nm/RIU for Hela, Jurkat, PC12, Basal, MDA-MB-231 and MCF-7 cancerous cells, respectively, as obtained by using Eq. (6). A comparison of the sensitivity between the proposed sensor and the sensor reported in the reference (Ayyanar et al. 2018; Jabin et al. 2019; Mollah et al. 2020; Mishra et al. 2020; Singh and Prajapati 2023; Yasli 2021) is shown in Table 1. The findings show that our proposed TC-PCF sensor has higher sensitivity over the currently available sensors, to the best of our knowledge. Previous models had a drawback in that they only evaluated the performance of individual cancer cells in relation to that of nearby cancer cells. They didn't compare cancerous cells to normal cells while making their evaluation. The proposed method aims to get over this limitation by detecting various cancerous cell bio-samples in contrast to normal cells. This method is considered to be the most effective one for detecting cancer.

4 Impact of diameter variation on sensitivity

Figure 8, 9, 10, 11, 12 and 13 illustrates the TC-PCF transmission curve with different diameters for Hela, Jurkat, PC 12, Basal, MDA-MB-231 and MCF-7 cell, respectively with corresponding fiber lengths of 1.0 m, 1.0 m, 0.8 m, 2.0 m, 0.6 m, 0.5 m, respectively. Figure 8a–c shows the proposed TC-PCF can obtain a sensitivity of 1,008,323.56 nm/RIU at D = 466 µm, 1,012,334.34 nm/RIU at D = 468 µm and 1,016,666.67 nm/RIU at D = 470 µm, respectively for Hela cell. Figure 9a–c indicates the proposed TC-PCF can obtain a sensitivity of 1,507,142.85 nm/RIU at D = 466 µm, 1,514,285.71 nm/RIU at D = 468 µm and 1,521,428.57 nm/RIU at D = 470 µm, respectively for Jurkat cell. Figure 10a–c illustrates the proposed TC-PCF can obtain a sensitivity of 1,576,757.37 nm/RIU at D = 466 µm, 1,578,486.50 nm/RIU at D = 468 µm and 1,578,571.42 nm/RIU nm/RIU at D = 470 µm, respectively for PC 12 cell. Figure 11a–c shows the proposed TC-PCF can obtain a sensitivity of 1,360,157.89 nm/RIU at D = 466 µm, 1,363,654.15 nm/RIU at D = 468 µm and 1,370,000 nm/RIU nm/RIU at D = 470 µm, respectively for Basal cell. Figure 12a–c indicates the proposed TC-PCF can obtain a sensitivity of 1,662,471.65 nm/RIU at D = 466 µm, 1,666,757.37 nm/RIU at D = 468 µm and 1,671,428.57 nm/RIU nm/RIU at D = 470 µm, respectively for MDA-MB-231 cell. Figure 13a–c illustrates the proposed TC-PCF can obtain a sensitivity of 1,498,258.96 nm/RIU at D = 466 µm, 1,502,324.65 nm/RIU at D = 468 µm and 1,507,142.85 nm/RIU nm/RIU at D = 470 µm, respectively for MCF-7

Table 1 Comparison between the sensor sensitivity and structure of the proposed and previous reported sensors

Reference	PCF Structure	Cell name	Sensitivity (nm/RIU)	Background material
Jabin et al. (2019)	Side polished PCF	Cervical cancer (HeLa)	10,208.33	Silica
		Blood cancer (Jurkat)	10,714.28	
		Adrenal glands cancer (PC12)	10,000	
		Skin cancer (Basal)	17,500	
		Breast cancer (MDA-MB-231)	17,857.14	
		Breast cancer (MCF-7)	18,071.42	
Ayyanar et al. (2018)	Dual-core PCF	Cervical cancer (HeLa)	7083.33	Silica
		Skin cancer (basal)	6000	
		Breast cancer (MDA-MB-231)	7857.14	
Mollah et al. (2020)	Triangular shape PCF	Blood cancer (Jurkat)	8571.43	Silica
Singh and Prajapati (2023)	Bottom-Side Polished PCF	Cervical cancer (HeLa)	20,000	Silica
		Blood cancer (Jurkat)	22,857	
		Adrenal glands cancer (PC12)	20,714	
		Skin cancer (Basal)	20,000	
		Breast cancer (MDA-MB-231)	21,428	
		Breast cancer (MCF-7)	25,000	
Mishra et al. (2020)	Heart-shaped PCF	Cervical cancer (HeLa)	7916.67	Silica
		Blood cancer (Jurkat)	8571.43	
		Adrenal glands cancer (PC12)	9285.71	
		Breast cancer (MDA-MB-231)	10,000	
		Breast cancer (MCF-7)	10,000	

Table 1 (continued)

Reference	PCF Structure	Cell name	Sensitivity (nm/RIU)	Background material
Yasli (2021)	Side tapered PCF	Cervical cancer (HeLa)	4333.33	Silica
		Blood cancer (Jurkat)	4642.86	
		Adrenal glands cancer (PC12)	5500	
		Skin cancer (Basal)	3150	
		Breast cancer (MDA-MB-231)	6428.57	
		Breast cancer (MCF-7)	7142.86	
Proposed structure	Twin-CorePCF	Cervical cancer (HeLa)	1,016,666.67	TOPAS (cyclic olefin copolymer)
		Blood cancer (Jurkat)	1,521,428.57	
		Adrenal glands cancer (PC12)	1,578,571.42	
		Skin cancer (Basal)	1,370,000	
		Breast cancer (MDA-MB-231)	1,671,428.57	
		Breast cancer (MCF-7)	1,507,142.85	

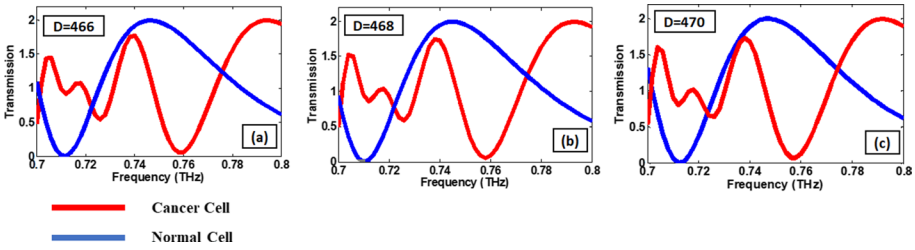


Fig. 8 Transmission curve with different diameter of D, for Hela cell **a** 466 µm **b** 468 µm and **c** 470 µm

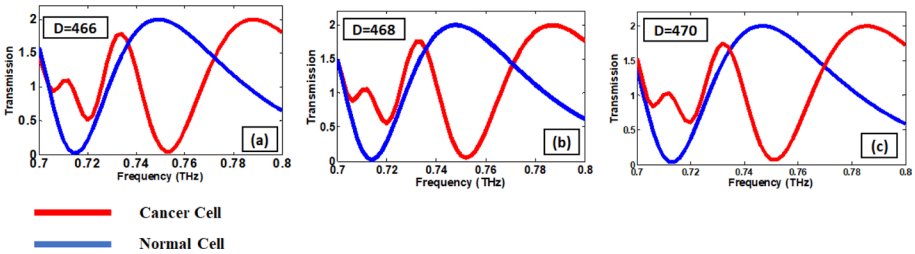


Fig. 9 Transmission curve with different diameter of D, for Jurkat cell **a** 466 µm **b** 468 µm and **c** 470 µm

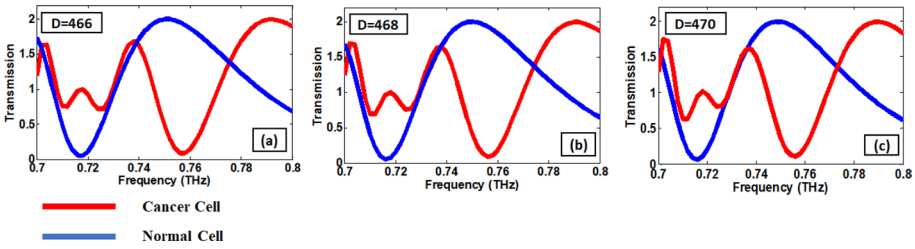


Fig. 10 Transmission curve with different diameter of D, for PC 12 cell **a** 466 µm **b** 468 µm and **c** 470 µm

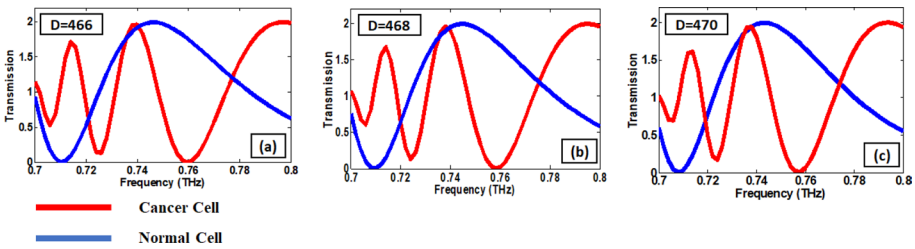


Fig. 11 Transmission curve with different diameter of D, for Basal cell **a** 466 µm **b** 468 µm and **c** 470 µm

cell. Diameter optimization indicates the proposed TC-PCF sensor has superior sensitivity at D = 470 µm.

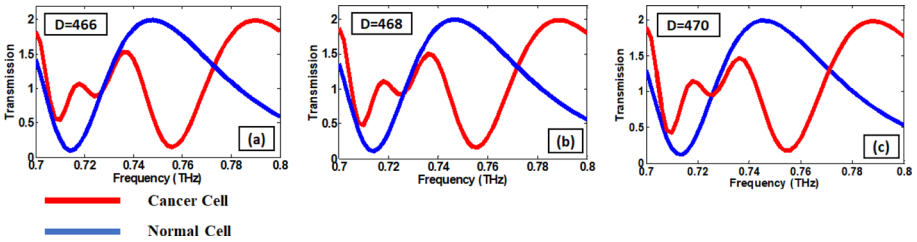


Fig. 12 Transmission curve with different diameter of D, for MDA-MB-231 cell **a** 466 µm **b** 468 µm and **c** 470 µm

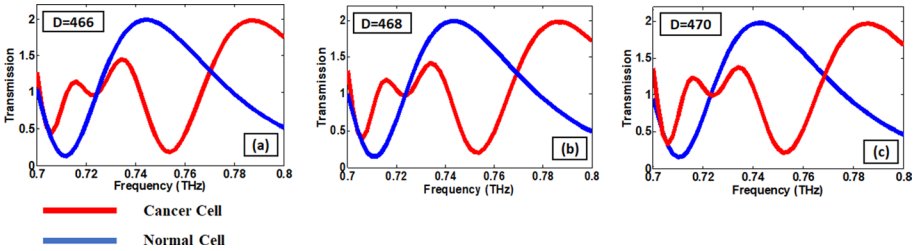


Fig. 13 Transmission curve with different diameter of D, for MCF-7 cell **a** 466 µm **b** 468 µm and **c** 470 µm

5 Conclusion

In this work, for the early detection of different cancerous cells, a novel type of sensor based on TC-PCF technology has been introduced by using the Finite Element Method and the COMSOL software. Our proposed sensor comprises injecting the sample cell (normal or cancerous) in the form of fluid, which interacts with the electric field during the propagation of light between the two cores of the TC-PCF, into the central elliptical hole of the structure. The guiding characteristics of transmission, are strongly influenced by the R.I. of the sample cell. As a result of the different R.I. values between normal and cancer cells, the proposed TC-PCF displays distinct optical properties for both types of cells. The proposed sensor can achieve a high sensitivity of 1,016,666.67 nm/RIU, 1,521,428.57 nm/RIU, 1,578,571.42 nm/RIU, 1,370,000 nm/RIU 1671428.57 nm/RIU and 1,507,142.85 nm/RIU for Hela, Jurkat, PC12, Basal, MDA-MB-231 and MCF-7 cancerous cells, respectively based on numerical analysis. Our proposed sensor exhibits great potential for the detection of different cancerous cells in the future because of its simple design and sensing technique.

Acknowledgements The authors are grateful to the participants who contributed to this research. The authors have not received any funding for this research.

Author contributions VC: Methodology, Writing—review & editing. SS: Supervision, Writing—original draft, review & editing.

Funding The authors have not received any funding for this research.

Availability of data and materials Not applicable.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal ties that would have influenced the work presented in this study.

Consent for publication All authors provide consent for publication.

Consent to participate (Ethics) All author is agreed to submit the paper in this journal.

Human and animal rights No bad Impact on animal.

References

Ahmed, K., Ahmed, F., Roy, S., Paul, B.K., Aktar, M.N., Vigneswaran, D., Islam, M.S.: Refractive index-based blood components sensing in Terahertz Spectrum. IEEE Sens. J. **19**, 3368–3375 (2019). https://doi.org/10.1109/JSEN.2019.2895166

Al Mahfuz, M., Mollah, M.A., Momota, M.R., Paul, A.K., Masud, A., Akter, S., Hasan, M.R.: Highly sensitive photonic crystal fiber plasmonic biosensor: design and analysis. Opt. Mater. (amst) **90**, 315–321 (2019). https://doi.org/10.1016/j.optmat.2019.02.012

An, G., Li, S., An, Y., Wang, H., Zhang, X.: Glucose sensor realized with photonic crystal fiber-based Sagnac interferometer. Opt. Commun. **405**, 143–146 (2017). https://doi.org/10.1016/j.optcom.2017.08.003

Arik, E., Koral, C., Altan, H., Esentürk, O.: A new method for alcohol content determination of fuel oils by terahertz spectroscopy.In: International Conference Infrared, Millimeter, Terahertz Waves, IRMMW-THz. (2013). https://doi.org/10.1109/IRMMW-THz.2013.6665885

Aslam Mollah, M., Yousufali, M., Faysal, M.R.B.A., Hasan, M.R., Hossain, M.B., Amiri, I.S.: Highly sensitive photonic crystal fiber salinity sensor based on Sagnac interferometer. Results Phys. (2020). https://doi.org/10.1016/j.rinp.2020.103022

Ayyanar, N., Thavasi Raja, G., Sharma, M., Sriram Kumar, D.: Photonic crystal fiber-based refractive index sensor for early detection of cancer. IEEE Sens. J. **18**, 7093–7099 (2018). https://doi.org/10.1109/JSEN.2018.2854375

Bise, R.T., Trevor, D.J.: Sol-gel derived microstructured fiber: fabrication and characterization. In: Conference Optical Fiber Communications Techical Digest Series 3, 269–271 (2005). https://doi.org/10.1109/ofc.2005.192772

Bulbul, A.A.M., Imam, F., Awal, M.A., Mahmud, M.A.P.: A novel ultra-low loss rectangle-based porous-core pcf for efficient THz waveguidance: design and numerical analysis. Sensors (switzerland). **20**, 1–18 (2020a). https://doi.org/10.3390/s20226500

Bulbul, A.A.M., Jibon, R.H., Das, S.K., Roy, T., Saha, A., Hossain, M.B.: PCF based formalin detection by exploring the optical properties in THz regime. Nanosci. Nanotechnol.-Asia. **11**, 314–321 (2020b). https://doi.org/10.2174/2210681210999200525171303

Chaudhary, V., Singh, S.: Twin core photonic crystal fiber based temperature sensor with improved sensitivity over a wide range of temperature. Opt. Quant. Electron. (2023). https://doi.org/10.1007/s11082-022-04304-4

Cheon, H., Yang, H.J., Son, J.H.: Toward clinical cancer imaging using Terahertz spectroscopy. IEEE J. Sel. Top. Quant. Electron. (2017). https://doi.org/10.1109/JSTQE.2017.2704905

Clark, L.C.: Implantable gas-containing biosensor and method for measuring an analyte such as glucose. Biotechnol. Adv. **5**, 343 (1987). https://doi.org/10.1016/0734-9750(87)90513-1

Cubillas, A.M., Unterkofler, S., Euser, T.G., Etzold, B.J.M., Jones, A.C., Sadler, P.J., Wasserscheid, P., Russell, P.S.J.: Photonic crystal fibres for chemical sensing and photochemistry. Chem. Soc. Rev. **42**, 8629–8648 (2013). https://doi.org/10.1039/c3cs60128e

De, M., Pathak, A.K., Singh, V.K.: Single channel photonic crystal fiber based high sensitive petrol adulteration detection sensor. Optik (stuttg). **183**, 539–546 (2019). https://doi.org/10.1016/j.ijleo.2019.03.001

Ebendorff-Heidepriem, H., Schuppich, J., Dowler, A., Lima-Marques, L., Monro, T.M.: 3D-printed extrusion dies: a versatile approach to optical material processing. Opt. Mater. Express **4**, 1494 (2014). https://doi.org/10.1364/ome.4.001494

Eid, M.M.A., Rashed, A.N.Z., Bulbul, A.A.M., Podder, E.: Mono-rectangular core photonic crystal fiber (MRC-PCF) for skin and blood cancer detection. Plasmonics **16**, 717–727 (2021). https://doi.org/10.1007/s11468-020-01334-0

Esfahani Monfared, Y.: Refractive index sensor based on surface plasmon resonance excitation in a D-shaped photonic crystal fiber coated by titanium nitride. Plasmonics **15**, 535–542 (2020). https://doi.org/10.1007/s11468-019-01072-y

Fischer, B.M., Walther, M., Jepsen, P.U.: Far-infrared vibrational modes of DNA components studied by terahertz time-domain spectroscopy. Phys. Med. Biol. **47**, 3807–3814 (2002). https://doi.org/10.1088/0031-9155/47/21/319

Fischer, B.M., Hoffmann, M., Helm, H., Wilk, R., Rutz, F., Kleine-Ostmann, T., Koch, M., Jepsen, P.U.: Terahertz time-domain spectroscopy and imaging of artificial RNA. Opt. Express **13**, 5205–5215 (2005). https://doi.org/10.1364/opex.13.005205

Ghazanfari, A., Li, W., Leu, M.C., Hilmas, G.E.: A novel freeform extrusion fabrication process for producing solid ceramic components with uniform layered radiation drying. Addit. Manuf. **15**, 102–112 (2017). https://doi.org/10.1016/j.addma.2017.04.001

Hajba, L., Guttman, A.: Circulating tumor-cell detection and capture using microfluidic devices. TrAC-Trends Anal. Chem. **59**, 9–16 (2014). https://doi.org/10.1016/j.trac.2014.02.017

Huq Arif, M.F., Mobarak Hossain, M., Rahman, M.M., Mithun Kumar, P.K., Khaled, S.M.: Photonic crystal based liquid sensor: a theoretical analysis for formalin detection. In: 2018 Joint 7th International Conference Informatics, Electronics Vision. 2nd International Conference on Imaging, Vision Pattern Recognition, ICIEV-IVPR 2018. 295–299 (2019). https://doi.org/10.1109/ICIEV.2018.8640998

Iqbal, F., Biswas, S., Bulbul, A.A.M., Rahaman, H., Hossain, M.B., Rahaman, M.E., Awal, M.A.: Alcohol sensing and classification using PCF-based sensor. Sens. Bio-Sens. Res. **30**, 100384 (2020). https://doi.org/10.1016/j.sbsr.2020.100384

Islam, M.S., Sultana, J., Ahmed, K., Islam, M.R., Dinovitser, A., Ng, B.W.H., Abbott, D.: A novel approach for spectroscopic chemical identification using photonic crystal fiber in the terahertz regime. IEEE Sens. J. **18**, 575–582 (2018b). https://doi.org/10.1109/JSEN.2017.2775642

Islam, M.S., Faisal, M., Razzak, S.M.A.: Dispersion flattened extremely high-birefringent kagome lattice elliptic core photonic crystal fiber in THz regime. Opt. Quant. Electron. (2019a). https://doi.org/10.1007/s11082-019-1744-9

Islam, M.S., Sultana, J., Cordeiro, C.M.B., Cruz, A.L.S., DInovitser, A., Ng, B.W.H., Abbott, D.: Broadband characterization of glass and polymer materials using THz-TDS. In: International Conference on Infrared, Millimeter, Terahertz Waves, IRMMW-THz. 2019-Septe, (2019). https://doi.org/10.1109/IRMMW-THz.2019.8874013

Jabin, M.A., Ahmed, K., Rana, M.J., Paul, B.K., Islam, M., Vigneswaran, D., Uddin, M.S.: Surface plasmon resonance based titanium coated biosensor for cancer cell detection. IEEE Photon. J. (2019). https://doi.org/10.1109/JPHOT.2019.2924825

Jibon, R.H., Rahaman, M.E., Alahe, M.A.: Detection of primary chemical analytes in the THz regime with photonic crystal fiber. Sens. Bio-Sens. Res. **33**, 100427 (2021). https://doi.org/10.1016/j.sbsr.2021.100427

Kanmani, R., Ahmed, K., Roy, S., Ahmed, F., Kumar Paul, B., Mani Rajan, M.S.: The performance of hosting and core materials for slotted core Q-PCF in Terahertz spectrum. Optik (stuttg). (2019). https://doi.org/10.1016/j.ijleo.2019.163084

Kaur, V., Singh, S.: Design approach of solid-core photonic crystal fiber sensor with sensing ring for blood component detection. J. Nanophotonics **13**, 1 (2019). https://doi.org/10.1117/1.jnp.13.026011

Lataoui, J., Rjeb, A., Jaba, N., Fathallah, H., Machhout, M.: Design of elliptical photonic crystal fiber (E-PCF) for the transmission of 116 OAM channels across the S, C, L and U bands. Results Opt. **11**, 100420 (2023). https://doi.org/10.1016/j.rio.2023.100420

Leon, M.J.B.M., Disha, A.S.: A simple structure of PCF based sensor for sensing sulfur dioxide gas with high sensitivity and better birefringence. Sens. Int. (2021). https://doi.org/10.1016/j.sintl.2021.100115

Li, T., Fan, Q., Liu, T., Zhu, X., Zhao, J., Li, G.: Detection of breast cancer cells specially and accurately by an electrochemical method. Biosens. Bioelectron. **25**, 2686–2689 (2010). https://doi.org/10.1016/j.bios.2010.05.004

Li, F.R., Li, Q., Zhou, H.X., Qi, H., Deng, C.Y.: Detection of circulating tumor cells in breast cancer with a refined immunomagnetic nanoparticle enriched assay and nested-RT-PCR. Nanomed. Nanotechnol. Biol. Med. **9**, 1106–1113 (2013). https://doi.org/10.1016/j.nano.2013.03.002

Markos, C., Stefani, A., Nielsen, K., Rasmussen, H.K., Yuan, W., Bang, O.: High-Tg TOPAS microstructured polymer optical fiber for fiber Bragg grating strain sensing at 110 degrees. Opt. Express **21**, 4758–4765 (2013). https://doi.org/10.1364/OE.21.004758

Mishra, G.P., Kumar, D., Chaudhary, V.S., Murmu, G.: Cancer cell detection by a heart-shaped dual-core photonic crystal fiber sensor. Appl. Opt. **59**, 10321 (2020). https://doi.org/10.1364/ao.409221

Mollah, M.A., Yousufali, M., Ankan, I.M., Rahman, M.M., Sarker, H., Chakrabarti, K.: Twin core photonic crystal fiber refractive index sensor for early detection of blood cancer. Sens. Bio-Sens. Res. **29**, 100344 (2020). https://doi.org/10.1016/j.sbsr.2020.100344

Monfared, Y.E.: Overview of recent advances in the design of plasmonic fiber-optic biosensors. Biosensors (2020). https://doi.org/10.3390/BIOS10070077

Monfared, Y.E., Hajati, M., Liang, C., Yang, S., Qasymeh, M.: Quasi-D-shaped fiber optic plasmonic biosensor for high-index analyte detection. IEEE Sens. J. **21**, 17–23 (2021). https://doi.org/10.1109/JSEN.2019.2945003

Mou, F.A., Rahman, M.M., Islam, M.R., Bhuiyan, M.I.H.: Development of a photonic crystal fiber for THz wave guidance and environmental pollutants detection. Sens. Bio-Sens. Res (2020). https://doi.org/10.1016/j.sbsr.2020.100346

Neugebauer, U., Clement, J.H., Bocklitz, T., Krafft, C., Popp, J.: Identification and differentiation of single cells from peripheral blood by Raman spectroscopic imaging. J. Biophoton. **3**, 579–587 (2010). https://doi.org/10.1002/jbio.201000020

Paul, B.K., Ahmed, K.: Highly birefringent TOPAS based single mode photonic crystal fiber with ultra-low material loss for Terahertz applications. Opt. Fiber Technol. (2019). https://doi.org/10.1016/j.yofte.2019.102031

Paul, B.K., Islam, M.S., Ahmed, K., Asaduzzaman, S.: Alcohol sensing over O+E+S+C+L+U transmission band based on porous cored octagonal photonic crystal fiber. Photon. Sens. **7**, 123–130 (2017). https://doi.org/10.1007/s13320-017-0376-6

Paul, B.K., Ahmed, K., Vigneswaran, D., Ahmed, F., Roy, S., Abbott, D.: Quasi-photonic crystal fiber-based spectroscopic chemical sensor in the terahertz spectrum: design and analysis. IEEE Sens. J. **18**, 9948–9954 (2018). https://doi.org/10.1109/JSEN.2018.2872892

Paul, B.K., Ahmed, K., Vigneswaran, D., Sen, S., Islam, M.S.: Quasi photonic crystal fiber for chemical sensing purpose in the terahertz regime: design and analysis. Opt. Quant. Electron. **51**, 1–12 (2019). https://doi.org/10.1007/s11082-019-1956-z

Paul, B.K., Ahmed, K., Dhasarathan, V., Nguyen, T.K.: Oligoporous-core Quasi cladding photonic crystal fiber based micro-sensor for alcohol detection. Phys. B Condens. Matter. (2020a). https://doi.org/10.1016/j.physb.2020.412104

Paul, B.K., Ahmed, K., El-Khozondar, H.J., Pobre, R.F., Peña, J.S.G., Merciales, M.C., Zainuddin, N.A.M., Zakaria, R., Dhasarathan, V.: The design and analysis of a dual-diamond-ring PCF-based sensor. J. Comput. Electron. **19**, 1288–1294 (2020b). https://doi.org/10.1007/s10825-020-01509-2

Podder, E., Hossain, M.B., Jibon, R.H., Bulbul, A.A.M., Mondal, H.S.: Chemical sensing through photonic crystal fiber: sulfuric acid detection. Front. Optoelectron. **12**, 372–381 (2019). https://doi.org/10.1007/s12200-019-0903-8

Poonam, S., Preeta, S.: Design of photonic crystal-based biosensor for detection of glucose concentration in urine. IEEE Sens. J. **15**, 1035–1042 (2015). https://doi.org/10.1109/JSEN.2014.2359799

Rahaman, M.E., Saha, R., Ahsan, M.S., Sohn, I.B.: Design and performance analysis of a D-shaped PCF and surface plasmon resonance based glucose sensor. In: 4th International Conference on Electrical Engineering and Information and Commununication Technololgy. iCEEiCT 2018. 325–329 (2019). https://doi.org/10.1109/CEEICT.2018.8628080

Rahaman, M.E., Jibon, R.H., Mondal, H.S., Hossain, M.B., Bulbul, A.A.M., Saha, R.: Design and optimization of a PCF-based chemical sensor in THz regime. Sens. Bio-Sens. Res. **32**, 100422 (2021). https://doi.org/10.1016/j.sbsr.2021.100422

Rahaman, M.E., Hossain, M.B., Mondal, H.S.: Effect of background materials in photonic crystal fiber sensor. Opt. Rev. **29**, 1–6 (2022a). https://doi.org/10.1007/s10043-021-00712-1

Rahaman, M.E., Jibon, R.H., Ahsan, M.S., Ahmed, F., Sohn, I.B.: Glucose level measurement using photonic crystal fiber–based plasmonic sensor. Plasmonics **17**, 1–11 (2022b). https://doi.org/10.1007/s11468-021-01497-4

Rahman, A., Rahman, A.K., Rao, B.: Early detection of skin cancer via terahertz spectral profiling and 3D imaging. Biosens. Bioelectron. **82**, 64–70 (2016). https://doi.org/10.1016/j.bios.2016.03.051

Rahman, M.M., Mou, F.A., Al Mahmud, A., Bhuiyan, M.I.H., Islam, M.R.: Photonic Crystal Fiber based Terahertz Sensor for Alcohol Detection in Beverages: Design and Analysis. In: 3rd IEEE International Conference Telecommunication Photonics, ICTP 2019. (2019). https://doi.org/10.1109/ICTP48844.2019.9041767

Rahman, M.M., Mou, F.A., Bhuiyan, M.I.H., Islam, M.R.: Photonic crystal fiber based terahertz sensor for cholesterol detection in human blood and liquid foodstuffs. Sens. Bio-Sens. Res. (2020). https://doi.org/10.1016/j.sbsr.2020.100356

Ramanujam, N.R., Amiri, I.S., Taya, S.A., Olyaee, S., Udaiyakumar, R., Pasumpon Pandian, A., Joseph Wilson, K.S., Mahalakshmi, P., Yupapin, P.P.: Enhanced sensitivity of cancer cell using one dimensional nano composite material coated photonic crystal. Microsyst. Technol. **25**, 189–196 (2019). https://doi.org/10.1007/s00542-018-3947-6

Ribaut, C., Loyez, M., Larrieu, J.C., Chevineau, S., Lambert, P., Remmelink, M., Wattiez, R., Caucheteur, C.: Cancer biomarker sensing using packaged plasmonic optical fiber gratings: towards in vivo diagnosis. Biosens. Bioelectron. **92**, 449–456 (2017). https://doi.org/10.1016/j.bios.2016.10.081

Rifat, A.A., Mahdiraji, G.A., Sua, Y.M., Shee, Y.G., Ahmed, R., Chow, D.M., Adikan, F.R.M.: Surface plasmon resonance photonic crystal fiber biosensor: a practical sensing approach. IEEE Photon. Technol. Lett. **27**, 1628–1631 (2015). https://doi.org/10.1109/LPT.2015.2432812

Roitberg, A., Heilweil, E.J., Markelz, A.G.: Pulsed terahertz spectroscopy of DNA, bovine serum albumin and collagen between 0.1 and 2.0 THz. Chem. Phys. Lett. **320**, 42–48 (2000). https://doi.org/10.1016/S0009-2614(00)00227-X

Sarker, S., Vigneswaran, D., Studnička, F.: Design of a nanoscale gold-coated photonic crystal fiber biosensor. Front. Phys. (2023). https://doi.org/10.3389/fphy.2023.1164255

Shupeng Liu, S.L., Lianxin Li, L.L., Zhenyi Chen, Z.C., Na Chen, N.C., Zhangmin Dai, Z.D., Jing Huang, J.H., Bo Lu, B.L.: Surface-enhanced Raman spectroscopy measurement of cancerous cells with optical fiber sensor. Chin. Opt. Lett. **12**, S13001-313003 (2014). https://doi.org/10.3788/col201412.s13001

Singh, S., Prajapati, Y.K.: Novel bottom-side polished PCF-based plasmonic biosensor for early detection of hazardous cancerous cells. IEEE Trans. Nanobiosci. (2023). https://doi.org/10.1109/tnb.2023.3233990

Soylemez, S., Udum, Y.A., Kesik, M., Gündoıdu Hizliateş, C., Ergun, Y., Toppare, L.: Electrochemical and optical properties of a conducting polymer and its use in a novel biosensor for the detection of cholesterol. Sens. Actuators, B Chem. **212**, 425–433 (2015). https://doi.org/10.1016/j.snb.2015.02.045

Sultana, J., Islam, M.S., Ahmed, K., Dinovitser, A., Ng, B.W.-H., Abbott, D.: Terahertz detection of alcohol using a photonic crystal fiber sensor. Appl. Opt. **57**, 2426 (2018). https://doi.org/10.1364/ao.57.002426

Sun, D., Guo, T., Ran, Y., Huang, Y., Guan, B.O.: In-situ DNA hybridization detection with a reflective microfiber grating biosensor. Biosens. Bioelectron. **61**, 541–546 (2014). https://doi.org/10.1016/j.bios.2014.05.065

Topas advanced polymer. [online]. available: https://topas.com/markets/electronics/more-electronics-apps. Accessed 15 Mar 2021

Vigneswaran, D., Ayyanar, N., Sharma, M., Sumathi, M., Mani, M.R., Porsezian, K.: Salinity sensor using photonic crystal fiber. Sens. Actuators, A Phys. **269**, 22–28 (2018). https://doi.org/10.1016/j.sna.2017.10.052

World health organization, Cancer: https://www.who.int/news-room/fact-sheets/ detail/cancer, (2018).

Yasli, A.: Cancer detection with surface plasmon resonance-based photonic crystal fiber biosensor. Plasmonics **16**, 1605–1612 (2021). https://doi.org/10.1007/s11468-021-01425-6

Yu, C., Fan, S., Sun, Y., Pickwell-Macpherson, E.: The potential of terahertz imaging for cancer diagnosis: a review of investigations to date. Quant. Imaging Med. Surg. **2**, 33–45 (2012). https://doi.org/10.3978/j.issn.2223-4292.2012.01.04

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

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

HindiPersonalityNet: Personality Detection in Hindi Conversational Data using Deep Learning with Static Embedding

AKSHI KUMAR

Department of Computing, Goldsmiths, University of London, United Kingdom

DIPIKA JAIN*

Department of Computer Science & Engineering, Delhi Technological University, New Delhi, India

ROHIT BENIWAL

Department of Computer Science & Engineering, Delhi Technological University, New Delhi, India

Personality detection along with other behavioural and cognitive assessment can essentially explain why people act the way they do and can be useful to various online applications such as recommender systems, job screening, matchmaking, and counselling. Additionally, psychometric NLP relying on textual cues and distinctive markers in writing style within conversational utterances reveal signs of individual personalities. This work demonstrates a text-based deep neural model, HindiPersonalityNet of classifying conversations into three personality categories {ambivert, extrovert, introvert} for detecting personality in Hindi conversational data. The model utilizes GRU with BioWordVec embeddings for text classification and is trained/tested on a novel dataset, शख्सियत (pronounced as Shakhsiyat) curated using dialogues from an Indian crime-thriller drama series, Aarya. The model achieves an F1-score of 0.701 and shows the potential for leveraging conversational data from various sources to understand and predict a person's personality traits. It exhibits the ability to capture semantic as well as long-distance dependencies in conversations and establishes the effectiveness of our dataset as a benchmark for personality detection in Hindi dialogue data. Further, a comprehensive comparison of various static and dynamic word embedding is done on our standardized dataset to ascertain the most suitable embedding method for personality detection.

Keywords: Personality, low-resource, deep learning, word embeddings, NLP

CCS CONCEPTS • Computing methodologies → Natural language processing; Machine learning

Additional Keywords and Phrases: Personality Psychology, Natural language, Conversational data

Taxonomy used in Paper:

Abbreviation	Definition
NLP	Natural Language Processing
MBTI	Myers Briggs Type Indicator
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
CNN	Convolutional Neural Network
BiLSTM	Bidirectional Long Short-Term Memory
GRU	Gated Recurrent Unit
GPT	Generative Pre-trained Transformer
RoBERTa	Robustly Optimized BERT Pretraining Approach
ELECTRA	Efficiently Learning an Encoder that Classifies Token Replacements Accurately
T5	Text-to-Text Transfer Transformer
XLM	Cross-lingual Language Model

* Corresponding author: *dipikajain_2k20phdco505@dtu.ac.in*

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
2375-4699/2023/1-ART1 $15.00
http://dx.doi.org/10.1145/3625228

Just Accepted

1 INTRODUCTION

NLP (Natural Language Processing) can be used to analyse sentiments and emotions in real-time language data and give insights to an individual's communication style in real-time, such as the use of assertive or passive language, the frequency of interruptions, or the level of engagement in the conversation. Psychometric NLP is a relatively new area of research that has gained popularity in recent years due to the increasing availability of large-scale textual data and the advancement of natural language processing techniques. Application of psychometric NLP to conversational data such as chat logs or transcripts of spoken conversations, has the potential to measure psychological constructs such as personality traits, emotions, and mental health conditions.

Personality traits constitute the most tangible psychometrics for psychological explanation of human behaviour. The study of personality traits using NLP provides a non-invasive, valuable tool for understanding individual differences and developing interventions to promote psychological well-being in various settings. There are multiple personality dimension theories in literature that offer distinct frameworks to comprehend personality and behaviour. Examples of such theories include the Big Five [1,2], MBTI [3-7], HEXACO [8,9] Personality Inventory, Cattell's Sixteen Personality Factor [10], Eysenck's Personality Theory [11], Enneagram [12], and the Five-Factor Personality Model [13], among others. While these theories may differ in their specific dimensions and categories, they all recognize the importance of introversion, extroversion, and ambiversion as key aspects of personality. The concept of introversion/extroversion was introduced in 1910 by Carl Gustav Jung [14], existing as part of a continuum with each personality type at separate ends of the scale. In 1947, psychologist Eysneck [11] added another category as *ambivert personality* for individuals showing traits of introverted personality in some situations and in others, they behave as an extrovert personality type.

Psychometric NLP can analyse language data to identify linguistic cues associated with the extroversion, introversion, and ambiversion dimensions of personality. Here are some examples of linguistic cues associated with each of these dimensions:

- *Extroversion:* People who score high on the extroversion dimension tend to be outgoing, social, and assertive. Linguistic cues that may indicate high extroversion include the use of more positive emotion words, more words related to social activities, and more assertive language.
- *Introversion:* People who score high on the introversion dimension tend to be quiet, reserved, and reflective. Linguistic cues that may indicate high introversion include the use of more negative emotion words, more words related to introspection and reflection, and fewer words related to social activities.
- *Ambiversion:* Ambiverts have characteristics of both extroversion and introversion and may display a balance of social and introspective behaviours. Linguistic cues that may indicate ambiversion include the use of a mix of positive and negative emotion words, a moderate use of words related to social activities, and a moderate use of introspective language.

Conversational data can provide valuable insights into a person's personality traits, as the way someone speaks or communicates can reflect their personality. Conversational data can be in the form of chat transcripts, emails, social media messages, or any other written or spoken communication between individuals. By analysing conversational data, it is possible to identify patterns in the language used, such as the frequency of certain words or phrases, the tone of the conversation, and the overall sentiment. These patterns can be used as features to train machine learning models that can predict personality traits. Especially as within the field of psycholinguistic, which studies the interrelation between linguistic factors and mental aspects, researchers have confirmed that individuals' personalities can be revealed by their written language and styles. For example, a person who frequently uses words related to aggression or dominance in their conversation may be classified as having a more dominant personality trait. Similarly, a person who uses more positive words and phrases may be classified as having a more positive and outgoing personality. While several pertinent efforts made to develop datasets and benchmarks for low-resource NLP task including sentiment analysis [15] and sarcasm detection [16], few attempts have recently been made to address the personality detection challenges in languages other than English. These include indigenous and low-resource ones such as Arabic [17], Persian [18,19], Bahasa Indonesia [20], and Indian languages like Hindi [3, 21], and Bengali [22]. Most of the Indian languages are low resource, which means that they have relatively lesser data available for training NLP systems, especially conversational systems. Moreover, the datasets created are based on MBTI and Big-5 frameworks

As individual tendency towards the outer world (extroversion) or the inner world (introversion) can have a significant influence on career choice, relationships and overall lifestyle, this research puts forward a personality detection model, HindiPersonalityNet, for assorting dialogues into three personality categories {ambivert, extrovert, introvert}. A conversational dataset called 'शख्सियत' (pronounced as Shakhsiyat) is created using dialogues from an Indian crime-thriller drama series in Hindi, Aarya. We demonstrate the use of deep neural network with static embedding that can

Just Accepted

learn and generalize patterns in the data, resulting in accurate and efficient personality detection. We run intensive simulations on the dataset and evaluate several performance metrics. The major contributions of this work can be summarised below:

- Creation of a novel text-based Hindi dataset, शख्सियत dataset (pronounced as Shakhsiyat) with three class labels {extrovert, introvert, ambivert} which can be used as a benchmark dataset for personality detection in Hindi language.
- Implementation and evaluation of HindiPersonalityNet model that uses BioWordVec embeddings with GRU for detecting personality traits in the conversational Hindi dataset.
- Comprehensive comparison of word embeddings for personality detection in Hindi dataset using Psychometric natural language processing.

The rest of the paper is organized as follows. Section 2 discusses some recent work done in the field of personality detection, specifically in Indian languages. Section 3 provides an overview of the fundamental concepts related to static and dynamic word embeddings, as well as deep learning models that are commonly used for NLP tasks. Section 4 delves into the specifics of the Hindi personality detection dataset that was curated for our study and presents the details of HindiPersonalityNet model for the task of personality detection in Hindi conversational data. The experimental results obtained are discussed in Section 5. Section 6 presents the concluding remarks and the scope for future improvement for this work.

2 RELATED WORK

Significant literature account for studies using machine learning (including the state-of-the-art transformer-based models) to detect and predict various psychological traits such as personality, behaviour as well as issues like depression and anxiety using linguistic cues from text. But a large body of work has been reported only in English language, especially from social media postings. Specific to personality detection public datasets and benchmark models have been well-reported and evaluated using MBTI and Big-5 personality traits [23,24]. Recent works also report few studies on personality detection other than English. In 2018, Adi et al. [20] presented optimization techniques for machine learning (ML)-based Bahasa Indonesian automatic personality detection. Fatehi et al. [18] put forward a deep neural network using fastText embeddings with LSTM and attention layers to automate a personality detection system in Persian language. The dataset was collected from Twitter and annotated into MBTI personality traits. The authors also used an interpretability technique to explain the results. In 2022, Anari et al. [19] proposed a lightweight deep convolutional neural network for categorizing personality into 9 types from handwritten Persian text. Salem et al. [17] proposed a ML technique for Egyptian dialect automatic personality detection.

Lately, a small number of studies have been reported on automatic personality detection in Indian languages. In one of the earliest works in personality detection in Hindi text, Singh et al. [25] introduced a psycho-lexical approach where the data from different Hindi novels and sources were collected and further studied based on three major trigunas such as sattvic, rajasic, and tamasic. In 2017, Singh and Raad [26] extended this work with the participation of 1250 different Hindi speaking people of young generation and recorded the personality traits by observing the tri-gunas and the big-5 personality traits. In 2020. Khan et al. [21] proposed a multimodal Hindi conversational dataset with audio, video, and Hinglish utterances (Hindi+English). One of our recently published works [3] reports the use of an ensemble of SVM kernels with soft voting for MBTI personality detection in a novel Hindi text-dataset विशेष चरित्र_MBTI (pronounced as vishesh charitr). Apart from Hindi language, in 2020, Rudra et. al. [22] has reported the use of various ML and deep learning models to detect Big-5 personality traits for informal Bangla transcripts. Our approach demonstrates the effectiveness of deep neural networks with static embeddings for personality detection in Hindi conversational data. GRU is used to model the temporal relationships between the words in the input text, as it allows to capture the nuanced and subtle patterns associated with specific personality traits in textual data. Simultaneously, BioWordVec embeddings are trained on a large corpus of biomedical literature, including scientific texts, articles, and abstracts, which enables them to learn the specific language used in the biomedical domain. This makes them effective at capturing the nuances of language use in specific domains, which can be beneficial for personality detection in conversational data.

3 PRELIMINARIES

Conversational AI is becoming more prevalent, with chatbots and virtual assistants being used for customer service, sales, and other tasks. NLP is a key component of conversational AI, as it enables these systems to understand natural language

Just Accepted

inputs and generate appropriate responses. In the past, NLP models relied heavily on handcrafted features and rule-based systems, which were limited in their ability to understand natural language. However, with the advent of word embeddings and deep learning, NLP models can now learn from large amounts of data and automatically extract meaningful features that capture the nuances of human language.

3.1 Word Embedding

Word embedding refers to the process of representing words as dense, low-dimensional vectors in a way that captures their semantic and syntactic properties. There are two main categories of word embeddings used in natural language processing (NLP): static and dynamic embeddings. Static embeddings, such as GloVe [27, 28], Crawl [29], Wikipedia [30], BioWordVec [31-32], GoogleNews [33], and PubMed [34], are pre-trained on a large corpus and remain fixed throughout the NLP task. They provide a fast and efficient approach to encoding words with semantic and syntactic information.

In contrast, dynamic embeddings, such as ElMo [35], fastText [36-37], and BERT [4], are generated during the training of a neural network for a specific NLP task and evolve over time. They can capture more context-dependent information, including the latest language trends, and are optimized for the specific NLP task at hand. However, they are computationally more expensive and require more training data than static embeddings. In this study, we perform a comprehensive comparison of word embeddings (static and dynamic) for personality detection in Hindi dataset using Psychometric natural language processing.

3.2 Deep Learning Models

Deep learning models have become increasingly popular in the field of natural language processing (NLP) due to their ability to handle large amounts of data and learn complex relationships between words and phrases. Some commonly used deep learning models for NLP include [38, 39]:

- CNN (Convolutional Neural Network): A neural network model that is commonly used for image recognition but can also be applied to NLP tasks such as text classification and sentiment analysis. The model uses convolutional filters to identify local patterns in the input text.
- GRU (Gated Recurrent Unit): A type of recurrent neural network (RNN) that is designed to overcome the vanishing gradient problem of traditional RNNs. GRUs use gating mechanisms to selectively update and reset the hidden state of the network.
- LSTM (Long Short-Term Memory): Another type of RNN that is designed to handle the problem of vanishing gradients and address the issue of forgetting long-term dependencies. LSTMs use a memory cell that can selectively retain or forget information over time.
- BiLSTM (Bidirectional Long Short-Term Memory): A variant of LSTM that processes input sequences in both forward and backward directions, allowing the model to capture both past and future information.
- Attention: A mechanism that can be added to various neural network architectures, including RNNs and transformers, to allow the model to selectively focus on specific parts of the input text. Attention has been shown to improve the performance of many NLP tasks, including machine translation and sentiment analysis.
- Transformer-based models: The transformer architecture has revolutionized the field of NLP, and researchers are constantly working to develop new and more powerful models. Some of the most recent examples include: GPT (Generative Pre-trained Transformer), RoBERTa (Robustly Optimized BERT Pretraining Approach), T5 (Text-to-Text Transfer Transformer), ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately), XLM (Cross-lingual Language Model), RoBERTa (Robustly Optimized BERT Pretraining Approach), XLNet

In this research, we utilized GRU [35, 40] and performed a comparative analysis involving LSTM, BiLSTM and CNN. Fig. 1 showcases the diverse combinations of embeddings and deep learning models employed in this study.

Just Accepted

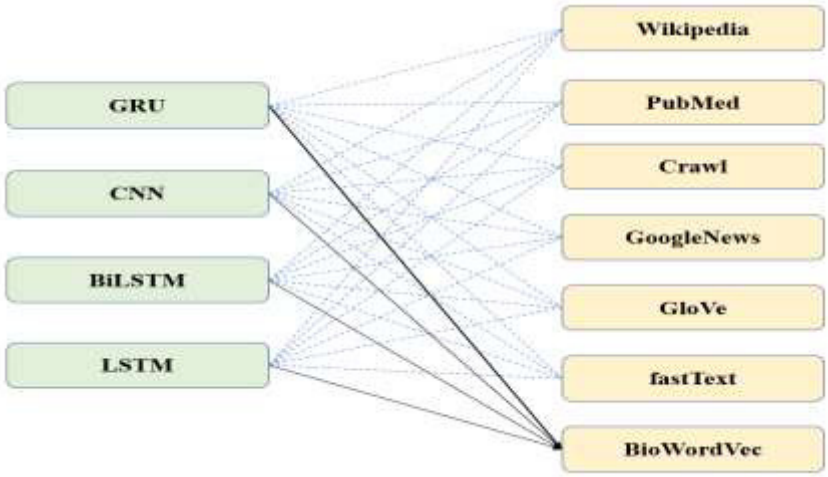


Fig.1. Deep learning and embedding employed

4 METHODOLOGY

This section presents the methodology employed to curate the 'Shakhsiyat' dataset and develop the 'HindiPersonalityNet' model, contributing to the advancement of natural language processing (NLP) in the Hindi language and personality detection tasks.

4.1 शख्सियत (pronounced as Shakhsiyat) Dataset

To create the "Shakhsiyat" dataset, which means "personality" in Hindi, we transcribed the dialogues of the Indian crime-thriller drama series called Aarya in Hindi. We manually classified the dialogues into three personality categories: introvert, extrovert, and ambivert, using various differentiators and context to guide our classifications. In order to ensure the accuracy of the annotations, we conducted a validation process in collaboration with a team of two highly knowledgeable and qualified post-graduate students specializing in Psychology from a reputable public university located in Delhi, India. Table 1 provides statistical details of the dataset.

Table 1. Shakhsiyat (**शख्सियत**) Dataset statistics

Description	Statistics
Total number of instances	6734
Total number of "Extrovert" instances	1710
Total number of "Introvert" instances	495
Total number of "Ambivert" instances	4529

The dataset comprises 6734 rows and 3 columns, with the first row serving as the header. The first column contains the speaker's name, such as Aarya, Veer, Tej, Adi, etc. The second column contains the actual utterance spoken by the speaker, while the last column provides the label for the utterance (extrovert, ambivert, or introvert). Table 2 illustrates some sample dialogues along with their corresponding categories. Our proposed Shakhsiyat (शख्सियत) dataset can serve as a benchmark for personality detection in the Hindi language, as well as other text classification or NLP-based tasks related to Hindi language.

Just Accepted

Table 2. Snapshot of Shakhsiyat (**शख्सियत**) dataset

Speaker (Character)	Utterances (Hindi)	Label
AARYA	**हाय स्वीटी।**	Extrovert
	Hi sweetie.	
ADI	**-माँ, वीर मुझे परेशान कर रहा है!**	Introvert
	Mother, Veer is troubling me!	
AARYA	**मैं जानता हूँ। घबराओ मत।**	Ambivert
	I know. Do not panic.	
AARYA	**घबराना बंद करो।**	Introvert
	stop panicking	
VEER	**आदि, बैठो।**	Ambivert
	Adi, sit down.	
AARYA	**आप अभी उठे हो**	Extrovert
	you just woke up	
VEER	**पापा, उससे कहो कि मुझे दे दो**	Introvert
	papa, tell him to give it to me	
VEER	**वहाँ बैठो!**	Ambivert
	sit there!	
ARU	**माँ, मुझे पैसे चाहिए।**	Extrovert
	Mother, I need money.	

The data extracted from the transcripts of the Hindi series contains various special characters, URLs, etc. This part of the transcript adds non-pertinent noise to the dialogues, and they need to be removed. To tackle this, we perform various text-cleaning techniques used extensively in Natural Language Processing. Since our dataset is in the Hindi language, we also use iNLTK library to better process the obtained Hindi dialogues. The iNLTK library is publicly available with basic in-built functions for Natural Language Processing in Indian Languages.

4.2 The HindiPersonalityNet Model

The model uses BioWordVec embedding to train GRU (Gated Recurrent Unit). BioWordVec embeddings are trained on a large corpus of biomedical literature, including scientific texts, articles, and abstracts, which enables them to learn the specific language used in the biomedical domain. This makes them effective at capturing the nuances of language use in specific domains, which can be beneficial for personality detection in conversational data. GRU is a type of recurrent neural network (RNN) architecture that is capable of processing sequential data. It is designed to overcome some of the limitations of traditional RNNs, such as the vanishing gradient problem, which can make it difficult to train deep models. GRUs use a gating mechanism to selectively update the hidden state at each time step, allowing them to capture long-term dependencies in the input sequence. To build the NLP model, we first initialize the GRU with random weights and then train it on the BioWordVec embeddings. The model takes in a sequence of word embeddings as input, which are processed by the GRU layer. The output of the GRU layer is then fed into a fully connected layer for the classification tasks. Fig.2 depicts the system architecture of the proposed model.

Just Accepted

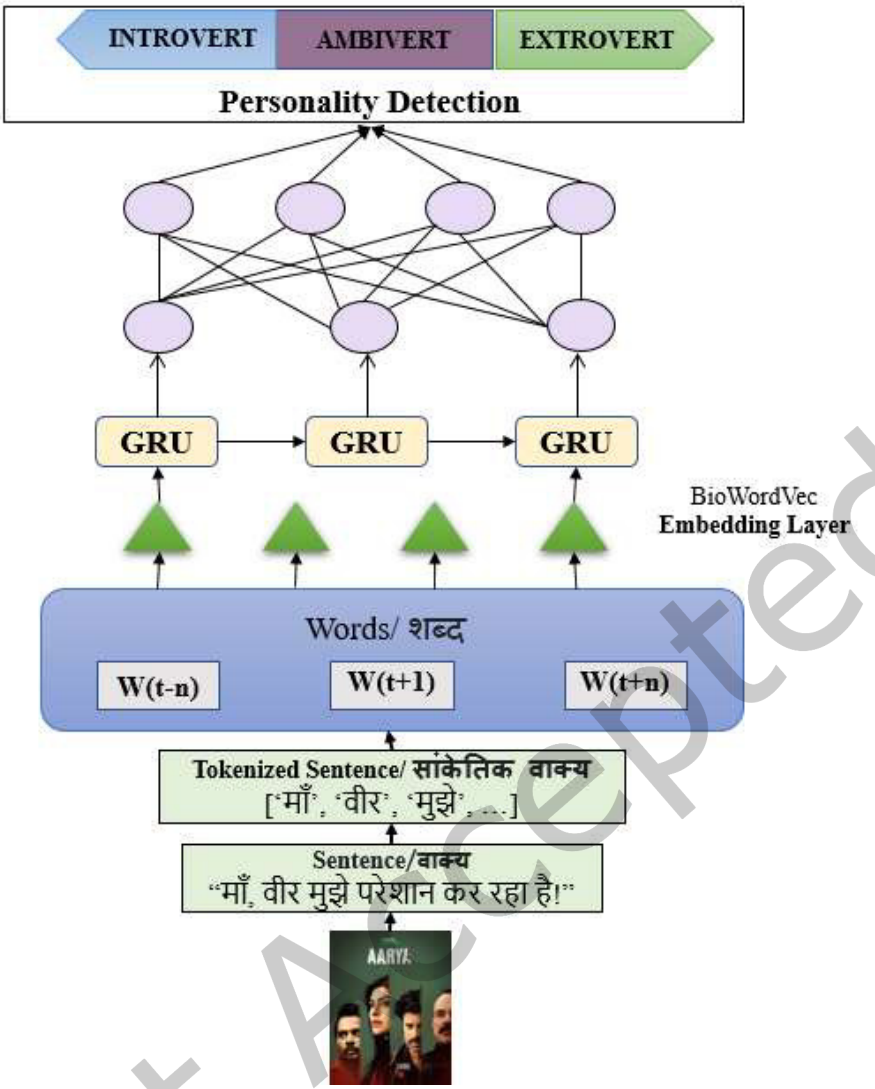


Fig.2. The HindiPersonalityNet Model

5 RESULTS AND DISCUSSION

The evaluation metrics utilized to estimate the performance of the proposed HindiPersonalityNet model include accuracy, precision, recall, and F1-score [3]. These metrics provide a comprehensive measure of the model's efficacy. While accuracy is considered, F1-score is also incorporated to address the sensitivity of accuracy towards imbalanced datasets and to assess the classifier's performance in terms of specificity and sensitivity. To establish a benchmark, the results of the proposed model are compared with those of the state-of-the-art model. Table 3 showcases the performance results of the proposed model, providing a comprehensive overview of its effectiveness.

Table 3. Comparative Performance Analysis of Hindi Personality Detection Models

Model	Dataset	Accuracy	F1-score	Precision	Recall
HindiPersonalityNet	शख्सियत dataset	0.739	0.701	0.704	0.739
KBSVE-P [3]	विशेष चरित्र_MBTI dataset	0.668	0.679	0.674	0.701

This model achieved the highest accuracy of 0.739 and recall of 0.739, indicating that it performed well in correctly identifying positive instances. It also had a relatively high F1-score of 0.701 and precision of 0.704. These results suggest

Just Accepted

that the HindiPersonalityNet model with GRU-BioWordVec embeddings is effective in capturing the patterns and features necessary for accurate predictions in the given task.

Table 4 presents a comparative performance analysis of two different models for Hindi and Bangla Personality Detection. The table provides information on the models, the datasets they were trained on, and the corresponding evaluation metrics including accuracy, F1-score, precision, and recall.

Table 4. Comparative Analysis of Hindi and Bangla Personality Detection Models

Model	Dataset	Accuracy	F1-score	Precision	Recall
HindiPersonalityNet	शख्सियत dataset	0.739	0.701	0.704	0.739
C-LSTM [22]	Bangla Personality Trait dataset	0.370	0.370	0.370	0.360

By comparing the performance metrics of the two models, it can be observed that the HindiPersonalityNet model achieved higher accuracy, F1-score, precision, and recall values compared to the C-LSTM model trained on the Bangla Personality Trait dataset. This analysis provides insights into the relative performance of the two models for personality detection in Hindi and Bangla languages.

Fig.3 presents the performance comparison of all 3 Indian language personality detection models.

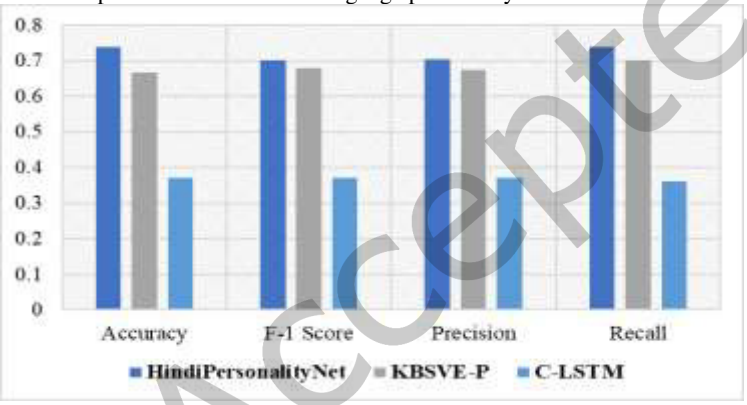


Fig.3. Performance comparison of Indian Language personality detection

The results presented in the table 5 compare the performance of different models using a combination of embedding and deep learning techniques, as illustrated in Fig 1.

Table 5. Comparative Analysis of various deep learning models with BioWordVec

Model	Accuracy	F1-score	Precision	Recall
HindiPersonalityNet (GRU-BioWordVec)	0.739	0.701	0.704	0.739
CNN BioWordVec	0.677	0.680	0.684	0.677
BiLSTM BioWordVec	0.679	0.675	0.674	0.679
LSTM BioWordVec	0.665	0.662	0.659	0.665

The CNN-based model achieved an accuracy of 0.677, which is lower than that of HindiPersonalityNet. However, it demonstrated a balanced F1-score of 0.680, precision of 0.684, and recall of 0.677. These results indicate that the CNN model performed reasonably well, but it fell short of the performance achieved by the HindiPersonalityNet model. Both BiLSTM BioWordVec and LSTM BioWordVec models exhibited similar performance. The BiLSTM BioWordVec achieved an accuracy of 0.679, an F1-score of 0.675, precision of 0.674, and recall of 0.679. On the other hand, the LSTM BioWordVec had an accuracy of 0.665, an F1-score of 0.662, precision of 0.659, and recall of 0.665. These models performed slightly lower than the HindiPersonalityNet and CNN models but still provided reasonably good results.

Table 6 presents a detailed comparative analysis of the HindiPersonalityNet model when combined with various deep learning models (CNN, LSTM, BiLSTM, GRU) and word embeddings (Crawl, fastText, GloVe, GoogleNews, PubMed,

Just Accepted

Wikipedia). The training and testing accuracies are reported for each combination, providing insights into how well the model performs on both the training and testing datasets.

Table 6. Accuracy comparison of embedding and deep learning combinations

Embedding + Deep Learning Model	Training Accuracy	Testing Accuracy
Crawl CNN	0.689	0.719
Crawl LSTM	0.735	0.724
Crawl GRU	0.742	0.721
Crawl BiLSTM	0.725	0.727
fastText CNN	0.757	0.724
fastText LSTM	0.821	0.722
fastText GRU	0.737	0.731
fastText BiLSTM	0.751	0.731
Glove CNN	0.707	0.727
Glove LSTM	0.749	0.725
Glove GRU	0.737	0.738
Glove BiLSTM	0.745	0.732
GoogleNews CNN	0.757	0.692
GoogleNews LSTM	0.853	0.698
GoogleNews GRU	0.766	0.725
GoogleNews BiLSTM	0.737	0.717
PubMed CNN	0.755	0.719
PubMed LSTM	0.784	0.727
PubMed GRU	0.752	0.731
PubMed BiLSTM	0.748	0.734
Wikipedia CNN	0.762	0.724
Wikipedia LSTM	0.752	0.712
Wikipedia GRU	0.764	0.721
Wikipedia BiLSTM	0.780	0.716

The table offers insights into the performance of different combinations of deep learning models and word embeddings for the HindiPersonalityNet model. It helps identify the most effective combinations in terms of accurately classifying instances and provides valuable information for selecting appropriate models and embeddings for similar tasks. By analysing the testing accuracies across different models, it is observed that fastText LSTM achieves the highest testing accuracy of 0.722, indicating its effectiveness in accurately classifying instances. The table also highlights the influence of different word embeddings on the model's performance. For instance, when combined with the GloVe embedding, the GRU model achieves a relatively high testing accuracy of 0.738. This suggests that the GloVe embedding provides valuable semantic information for the model to make accurate predictions.

The graph in fig.4 depicts the accuracy comparison of the seven embedding (Crawl, fastText, GloVe, GoogleNews, PubMed, Wikipedia and BioWordVec) when used with GRU. It can be observed that the HindiPersonalityNet model that uses BioWordVec with GRU outperforms the other models.

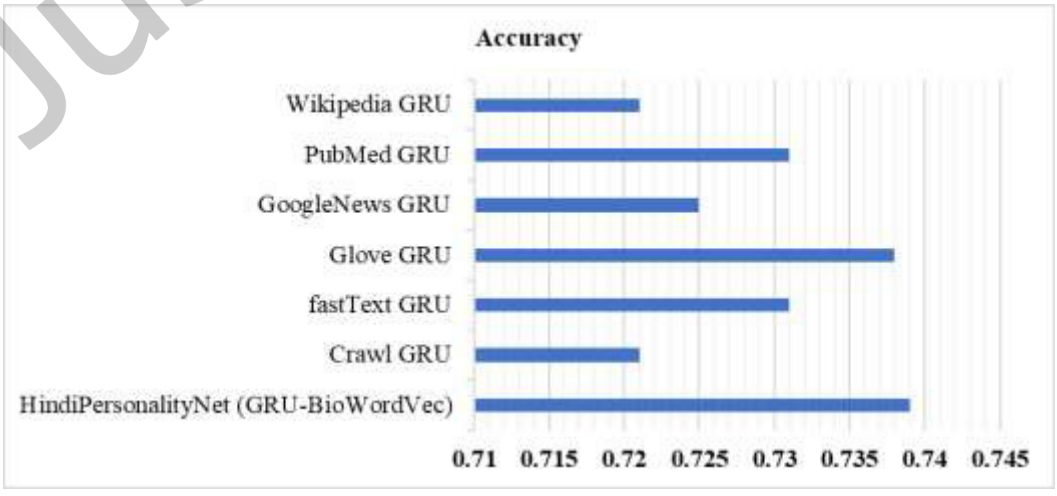


Fig.4. Accuracy comparison for embeddings with GRU

Just Accepted

6 CONCLUSION AND FUTURE WORK

In recent years, the field of natural language processing (NLP) has witnessed significant advancements in various domains and languages. In the context of Hindi language analysis, the development of comprehensive datasets and effective models becomes essential for enabling sophisticated NLP applications. This paper presents two significant contributions: the 'Shakhsiyat' dataset and the 'HindiPersonalityNet' model. The 'Shakhsiyat' dataset is curated by transcribing dialogues from the popular Indian crime-thriller drama series 'Aarya' in Hindi, meticulously classified into introvert, extrovert, and ambivert personality categories. With 6734 instances, this dataset provides a benchmark for personality detection in the Hindi language. Additionally, the 'HindiPersonalityNet' model utilizes BioWordVec embeddings and a GRU architecture, trained on a large corpus of biomedical literature, to capture domain-specific language nuances. This model demonstrates its efficacy in personality detection tasks and showcases its potential for conversational data analysis. Together, the 'Shakhsiyat' dataset and the 'HindiPersonalityNet' model contribute to the advancement of NLP in the Hindi language and offer valuable resources for researchers and practitioners in this field.

While this paper presents significant contributions to NLP in Hindi, there are several avenues for future work and improvement. The 'Shakhsiyat' dataset provides a solid foundation, but its further expansion with additional diverse dialogues and annotations from different sources can enhance its representativeness and applicability. Also, the 'HindiPersonalityNet' model shows promising results, there is scope for improvement through architectural modifications, exploring different deep learning models, or incorporating more advanced techniques such as attention mechanisms or transformer-based architectures. Further, integrating other modalities such as audio and visual cues along with text can lead to more comprehensive personality detection models, enabling a deeper understanding of human behaviour in conversations. The current classification into introvert, extrovert, and ambivert categories provides a broad categorization and therefore as another direction of future work could focus on developing models that can detect and differentiate more nuanced personality traits and dimensions. Lastly, the practical deployment of the HindiPersonalityNet model in real-world applications, such as chatbots, virtual assistants, or social media analysis, warrants further investigation to assess its effectiveness and usability.

Just Accepted

REFERENCES

1. Yoneda, T., Lozinski, T., Turiano, N., Booth, T., Graham, E. K., Mroczek, D., & Terrera, G. M. (2023). The Big Five personality traits and allostatic load in middle to older adulthood: A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, 105145.
2. Ong V, Rahmanto ADS, Williem, Suhartono D, Nugroho AE, Andangsari EW, et al. Personality Prediction Based on Twitter Information in Bahasa Indonesia. *In: Proceedings of the 2017 Federated Conference on Computer Science and Information Systems*. Prague, Czech Republic; 2017. p. 367–372
3. Kumar, A., Beniwal, R., & Jain, D. (2023). Personality Detection using Kernel-based Ensemble Model for leveraging Social Psychology in Online Networks. *ACM Transactions on Asian and Low-Resource Language Information Processing*.
4. Jain, D., Kumar, A., & Beniwal, R. (2022, July). Personality BERT: A Transformer-Based Model for Personality Detection from Textual Data. In *Proceedings of International Conference on Computing and Communication Networks: ICCCN 2021* (pp. 515-522). Singapore: Springer Nature Singapore.
5. Ren, Z., Shen, Q., Diao, X., & Xu, H. (2021). A sentiment-aware deep learning approach for personality detection from text. *Information Processing & Management*, *58*(3), 102532.
6. Cerkez, N., Vrdoljak, B., & Skansi, S. (2021). A Method for MBTI Classification Based on Impact of Class Components. *IEEE Access*, *9*, 146550-146567.
7. Shafi, H., Sikander, A., Jamal, I. M., Ahmad, J., & Aboamer, M. A. (2021). A Machine Learning Approach for Personality Type Identification using MBTI Framework. *Journal of Independent Studies and Research Computing*, *19*(2), 6-10.
8. Ashton, M. C., & Lee, K. (2007). Empirical, Theoretical, and Practical Advantages of the HEXACO Model of Personality Structure. Personality and Social Psychology Review, 11(2), 150–166. https://doi.org/10.1177/1088868306294907
9. Aghababaei, N., & Arji, A. (2014). Well-being and the HEXACO model of personality. *Personality and Individual Differences*, *56*, 139-142.
10. Ross, C., Orr, E. S., Sisic, M., Arseneault, J. M., Simmering, M. G., & Orr, R. R. (2009). Personality and motivations associated with Facebook use. *Computers in human behavior*, *25*(2), 578-586.
11. Davidson IJ. The ambivert: A failed attempt at a normal personality. *Journal of the History of the Behavioral Sciences*. 2017 Sep;53(4):313-31.
12. Riso, D. R., & Hudson, R. (2000). Understanding the enneagram: The practical guide to personality types. *Houghton Mifflin Harcourt.*
13. Montag, I., & Levin, J. (1994). The five-factor personality model in applied settings. *European Journal of Personality*, 8(1), 1-11.
14. Guilford JP, Braly KW. Extroversion and introversion. *Psychological Bulletin*. 1930 Feb;27(2):96.
15. Kumar, A., & Albuquerque, V. H. C. (2021). Sentiment Analysis Using XLM-R Transformer and Zero-shot Transfer Learning on Resource-poor Indian Language. *Transactions on Asian and Low-Resource Language Information Processing*, *20*(5), 1-13.
16. Jain, D., Kumar, A., & Garg, G. (2020). Sarcasm detection in mash-up language using soft-attention based bi-directional LSTM and feature-rich CNN. *Applied Soft Computing*, *91*, 106198.
17. Salem MS, Ismail SS, Aref M. Personality traits for Egyptian twitter users dataset. In Proceedings of the *2019 8th international conference on software and information engineering 2019* Apr 9 (pp. 206-211).
18. Fatehi S, Anvarian Z, Madani Y, Mehditabar M, Eetemadi S. MBTI Personality Prediction Approach on Persian Twitter. [28_Paper.pdf (winlp.org)]
19. Anari MS, Rezaee K, Ahmadi A. TraitLWNet: a novel predictor of personality trait by analyzing Persian handwriting based on lightweight deep convolutional neural network. *Multimedia Tools and Applications*. 2022 Mar;81(8):10673-93.
20. Adi GY, Tandio MH, Ong V, Suhartono D. Optimization for automatic personality recognition on Twitter in Bahasa Indonesia. *Procedia Computer*

Science. 2018 Jan 1; 135:473-80.

21. Khan SN, Leekha M, Shukla J, Shah RR. Vyaktitv: A multimodal peer-to-peer hindi conversations-based dataset for personality assessment. In2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM) 2020 Sep 24 (pp. 103-111). IEEE.
22. Rudra U, Chy AN, Seddiqui MH. Personality traits detection in Bangla: A benchmark dataset with comparative performance analysis of state-of-the-art methods. *In2020 23rd International Conference on Computer and Information Technology (ICCIT) 2*020 Dec 19 (pp. 1-6). IEEE.
23. Mehta Y, Majumder N, Gelbukh A, Cambria E. Recent trends in deep learning-based personality detection. *Artificial Intelligence Review*. 2020 Apr; 53:2313-39.
24. Vásquez RL, Ochoa-Luna J. Transformer-based Approaches for Personality Detection using the MBTI Model. *In2021 XLVII Latin American Computing Conference (CLEI) 2021 Oct 25* (pp. 1-7). IEEE.
25. Singh JK, Misra G, De Raad B. Personality structure in the trait lexicon of Hindi, a major language spoken in India. *European Journal of Personality*. 2013 Nov;27(6):605-20.
26. Singh JK, De Raad B. The personality trait structure in Hindi replicated. *International Journal of Personality Psychology*. 2017 Jun 29;3:26-35.
27. Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. *In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).
28. Kumar, A., Srinivasan, K., Cheng, W. H., & Zomaya, A. Y. (2020). Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data. *Information Processing & Management*, 57(1), 102141.
29. Sarma, P. K., Liang, Y., & Sethares, W. A. (2018). Domain adapted word embeddings for improved sentiment classification. *arXiv preprint arXiv:1805.04576*.
30. Sheehan, E., Meng, C., Tan, M., Uzkent, B., Jean, N., Burke, M., ... & Ermon, S. (2019, July). Predicting economic development using geolocated wikipedia articles. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2698-2706).
31. Zhang, Y., Chen, Q., Yang, Z., Lin, H., & Lu, Z. (2019). BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific data,* 6(1), 52.
32. Wang, S., Tseng, B., & Hernandez-Boussard, T. (2021). Development and evaluation of novel ophthalmology domain-specific neural word embeddings to predict visual prognosis. *International journal of medical informatics*, *150*, 104464.
33. Wang, Y., Liu, S., Afzal, N., Rastegar-Mojarad, M., Wang, L., Shen, F., ... & Liu, H. (2018). A comparison of word embeddings for the biomedical natural language processing. *Journal of biomedical informatics*, *87*, 12-20.
34. Sharma, S., & Daniel Jr, R. (2019). Bioflair: Pretrained pooled contextualized embeddings for biomedical sequence labeling tasks. *arXiv preprint* arXiv:1908.05760.
35. Kumar, A., & Sachdeva, N. (2022). A Bi-GRU with attention and CapsNet hybrid model for cyberbullying detection on social media. *World Wide Web*, 25(4), 1537-1550.
36. Jain, D. K., Kumar, A., & Sangwan, S. R. (2022). TANA: The amalgam neural architecture for sarcasm detection in indian indigenous language combining LSTM and SVM with word-emoji embeddings. *Pattern Recognition Letters*, 160, 11-18.
37. Hu, S., Kumar, A., Al-Turjman, F., Gupta, S., & Seth, S. (2020). Reviewer credibility and sentiment analysis based user profile modelling for online product recommendation. *IEEE Access,* 8, 26172-26189.
38. Ni, J., Young, T., Pandelea, V., Xue, F., & Cambria, E. (2023). Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial intelligence review*, 56(4), 3055-3155.
39. Dodge, J., Gurevych, I., Schwartz, R., Strubell, E., & van Aken, B. (2023). Efficient and Equitable Natural Language Processing in the Age of Deep Learning (Dagstuhl Seminar 22232). *In Dagstuhl Reports* (Vol. 12, No. 6). Schloss Dagstuhl-Leibniz-Zentrum für Informatik.
40. Dey, R., & Salem, F. M. (2017, August). Gate-variants of gated recurrent unit (GRU) neural networks. *In 2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS*) (pp. 1597-1600). IEEE.

Just Accepted

Social Network Analysis and Mining (2023) 13:141
https://doi.org/10.1007/s13278-023-01145-1

ORIGINAL ARTICLE

Identifying discernible indications of psychological well-being using ML: explainable AI in reddit social media interactions

Pahalage Dona Thushari[1] · Nitisha Aggarwal[2] · Vajratiya Vajrobol[2] · Geetika Jain Saxena[3] · Sanjeev Singh[2] · Amit Pundir[3]

Received: 9 May 2023 / Revised: 25 September 2023 / Accepted: 26 September 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH Austria, part of Springer Nature 2023

Abstract

Psychological well-being is a multidimensional construct and identifying it using a systematic, comprehensive approach offers insights fundamental to critical outcomes. Social networks are valuable resources for research, providing a pragmatic way of generating empirical evidence on psychological well-being based on the textual indicators across different populations. This study analyzed the information on various Reddit social media groups dedicated to mental health. The classes, namely depression, anxiety, bipolar, and SuicideWatch, using the SWMH dataset have been analyzed. The text-based interactions of persons with mental illness have common motifs like negative language and expressions like 'hopelessness,' 'emptiness,' or 'helplessness.' Topic modeling identified recurring themes and subjects that helped classify discernible factors influencing mental health. Classifiers for multiclass classification to classify targeted mental health issues based on users' network behavior and posts were trained and tested to get predictions on context (e.g., MentalBERT) and non-context-based (e.g., LR and NB) models. The MentalBERT model outperformed the other eight baseline models with an average accuracy of 76.70%, which is 4% more than reported in previous studies. Explainable AI was used to examine the trustworthiness of each model, and the explanations were evaluated using the LIME model. Explainability is crucial as mental health data characterizes syndromes, outcomes, disorders, and signs/symptoms exhibiting probabilistic interrelationships with each other. Explanations of these intricate interconnections can assist the extensive research around the model of well-being and interventions intended to improve the human condition and distill positive human functioning.

Keywords Topic modeling · BERTopic · Mental health · Reddit · LIME · Multiclass classification

✉ Amit Pundir
amitpundir@mac.du.ac.in

Pahalage Dona Thushari
thusharipahalage@gmail.com

Nitisha Aggarwal
nitisha@south.du.ac.in

Vajratiya Vajrobol
tiya101@south.du.ac.in

Geetika Jain Saxena
gsaxena@mac.du.ac.in

Sanjeev Singh
sanjeev@south.du.ac.in

[1] Department of Software Engineering, Delhi Technological University, New Delhi, Delhi 110042, India

[2] Institute of Informatics and Communication, University of Delhi, New Delhi, Delhi 110021, India

[3] Department of Electronics, Maharaja Agrasen College, University of Delhi, New Delhi, Delhi 110096, India

1 Introduction

The condition of one's mental health significantly impacts the person suffering and the community (Benrouba and Boudour 2023; Ji et al. 2022). According to the United Nations report, in 2022, one in eight people globally suffers from a mental health disorder (World mental health report: Transforming mental health for all - executive summary 2022a). The report has underscored the disparities in mental healthcare, underscoring the global prevalence of high mental health needs alongside frequently inadequate and insufficient responses. Furthermore, it has proposed measures to enhance mental healthcare systems, rendering them more accessible and effective for everyone. The delayed diagnosis of mental illnesses, resulting in delayed preventive interventions, has contributed to approximately 703,000 annual suicide-related deaths across various age groups, genders, and geographical regions. Presently, suicide stands as a leading

Published online: 25 October 2023

cause of mortality, surpassing other factors like malaria, AIDS, breast cancer, war, and homicide on a global scale. To identify high-risk demographics, nations should gather and scrutinize disaggregated data encompassing variables such as gender, age, and suicide methods. These data hold paramount importance in grasping the scope of the issue and facilitate the tailoring of interventions to suit the unique requirements of at-risk populations while adapting to evolving trends (Suicide data: Mental Health and Substance Use 2021). The diagnostic and statistical manual of mental disorders (DSM-V), published by the American Psychiatric Association, defines a mental disorder as a set of symptoms that cause significant disruption in a person's thinking, emotional regulation, or mental functioning that is consistent with psychological, biological, or developmental processes. Mental illnesses are typically accompanied by severe suffering or impairment in vital social, occupational, or other tasks (Stein et al. 2021). Unfortunately, more than 70% of individuals with mental disorders worldwide lack primary care and treatment due to limited resource environments (Kilbourne et al. 2018).

Mental health professionals study feelings, thoughts, behavior patterns, and other tests to diagnose illness. The availability of such resources is a critical constraint for conducting these studies, as health records of patients' emotional and intimate feelings are not shared due to privacy and confidentiality concerns. Data scarcity, therefore, restricts the research in this domain and severely limits designing accurate methods and techniques for diagnosing mental illness (Hanna et al. 2018). Moreover, most mental disorders have similar characteristics, making distinguishing between mental health diseases for diagnostic purposes difficult. In addition, the availability of mental health experts per million population, particularly in Low- and middle-income countries (LMICs), is deficient (Wainberg et al. 2017). Handling of few mental illnesses, such as depression, anxiety, and bipolar disorder, is challenging as they do not have cures. However, behavior therapy and medication can significantly control the illness and the quality of patients' lives may significantly improve. Therefore, reducing stress in peoples' lives is crucial by knowing the cause of illness and minimizing its impact on their daily lives.

Social network platforms allow individuals apprehensive about face-to-face interactions to share their thoughts and feelings with a larger community. These social media platforms are easy to access and do not discriminate based on age, gender, socio-economic status, race, or ethnicity. Communicating with a larger community of people suffering from similar concerns in online forums helps them regulate their emotions. Their interactions, in the form of texts, become an invaluable resource facilitating analysis of textual signs of psychological health problems (Dao et al. 2015; Kamarudin et al. 2021). Among the popular social media networks, the Reddit platform is preferred by young individuals for sharing their emotions as it provides anonymity and a handy platform for discussions on mental health issues (Boettcher 2021). Researchers using such data can analyze the emotions expressed to understand the mental issues the users suffer and design tools to find the types, stages and cost-effective solutions (Ren et al. 2021).

Recent years witnessed an abrupt growth in the analysis of social media data to investigate a variety of health issues ranging from the effects of allergies and Covid-19 (Huang et al. 2022; Zhou et al. 2021; Kathy et al. 2015) and sentimental analysis on Covid-19 vaccine (Alotaibi et al. 2023; Verma et al. 2023) to emotions and mental health conditions (Saha et al. 2016; Kim et al. 2020; Lin et al. 2023). Sentiment analysis is applied to identify the underlying sentiment or emotion expressed in a piece of text. The sentiment of a person's post, certain language and linguistic patterns can provide insights into their mental state. Natural language processing (NLP) analyzes such data and allows for diagnosing disorders and developing treatment strategies. Numerous methods have been developed to map the semantic relations between textual data (Qi and Shabrina 2023; Rizvi et al. 2023). Deep learning (DL) models recurrent neural network (RNN), long short-term memory (LSTM), transformer, convolutional neural networks (CNN), bidirectional encoder representations from transformers (BERT) and other hierarchical attention networks (HAN) have been applied to analyze text at different levels, such as document level and sentence level, for identifying mental health-related concerns. BERT and other transformer-based models employ attention-based mechanisms; hence, they offer several benefits over traditional models like CNN and LSTM in the context of NLP. This attention mechanism helps identify significant words, phrases, or patterns that contribute to the overall meaning or sentiment as they focus on different parts of the input text. They also consider the neighboring words and their relationships to allow for a better understanding of the meaning and semantics of a word within a sentence to capture contextual information effectively. Traditional models like CNN and LSTM process data sequentially and hence are limited in capturing long-range dependencies in text, while transformer models can capture dependencies between words regardless of their positional distance through parallel processing. Additionally, contextual methods, like the MentalBERT model, consider the context and relationships within the text to understand its meaning. They capture nuances and dependencies, making them suitable for mental health discussions. Non-contextual methods like Bag of Words and TF-IDF analyze words in isolation without considering the context. They extract features using statistical or rule-based techniques, providing insights without contextual nuances. Non-contextual methods may not capture the subtleties and dependencies present in the text as effectively as

contextual methods. However, they can still provide valuable insights, especially when the specific context is not crucial for the analysis. This paper uses recent developments in NLP to find the common underlying topics in communities with mental health issues using their social media interactions. Our work has three focuses of interest:

1. The extraction and evaluation of significant themes and subjects for sensitive mental health issues using topic modeling;
2. Comparative analysis of context-based and non-context-based machine learning (ML) classifiers to automate the classification process of textual data on mental health issues, and
3. Using model interpretability, analyze the secular associations present in the distinctive attributes.

Identifying the variables contributing to mental health issues based on social media interactions can reveal common factors. These factors may subsequently lead to determining new perspectives of interventions and strategies to achieve better mental health solutions. This research uncovered the main themes and patterns within mental health-related content derived from a social media platform, employing the topic modeling technique. Through these identified themes, we gain access to the underlying semantic structure embedded within the text. For instance, recurrent themes in the study encompassed topics like exams, school, college, friends, and relationships, highlighting their interconnectedness with mental health. Moreover, the outcomes revealed that context-driven models such as BERT and MentalBERT exhibited superior post-classification accuracy compared to traditional models like logistic regression (LR) and random forest (RF). To gain insights into the classification model results, we employed LIME, which indicated that BERT and MentalBERT classified posts based on contextual factors. In addition to this, our findings indicated that individuals with mental health concerns may potentially exhibit a likelihood of multiple mental disorders, a facet elucidated through the predictive capability of LIME explanations. The rest of the paper has been arranged as follows. Section 2 reports the related work, while the methodology is shown in Sect. 3. Section is about the findings of the study. Section 4 is the discussion related to the results obtained and their significance. Finally, in Sect. 5, the conclusions and future work are discussed.

2 Related work

Recent advancements in NLP and deep learning have positively influenced the analysis of social networking media interactions of communities with mental health issues. Social media data on platforms such as Twitter and Reddit (Liu et al. 2022) are of prime interest to researchers as they are real-time, readily available and help to reach a wider audience at a low cost. The research by Ji et al. (2022) describes a relation network with an attention mechanism to identify mental disorders and suicidal ideation with associated risk indicators. The SWMH (SuicideWatch and Mental Health) real-world dataset contains subreddit data divided into multiple classes, which we used in this study. They enhanced text representation and measured sentiment score and latent topics by lexicons. Their best-performing model for SWMH data achieved 64.74% accuracy for relation networks. For the same dataset, another study introduced pre-trained language models, namely MentalBERT and MentalRoBERTa (Ji et al. 2021) and achieved 72.16% best F1-score for the MentalRoBERTa model. These models were trained on domain-specific language for mental healthcare and are publicly available.

The work by Guntuku et al. (2017) investigated potential ways to use screening surveys on social media to predict mental health disorders. They detected symptoms associated with mental illness from Twitter, Facebook, and web forums and suggested that AI-driven methods can detect mental illnesses. An earlier study by Gemmell et al. (2019) investigated how to automatically recognize informal patterns in the language retrieved from online forums for borderline personality disorder patients as well as bipolar disorder patients. The top 10 phrases and terms were found to best describe each cluster using k-means clustering on the Reddit data. Another study by Kotenko et al. (2021) on the evaluation of the Mental Health of Social Network Users (Pushshift Reddit Dataset) calculated the emotion lexicon, used Latent Dirichlet Allocation (LDA) for topic modeling and classification using the fastText classifier and achieved 96% F1-score. Traditional approaches, such as LDA (Blei et al. 2003), failed to capture the semantic relationships and were not domain-specific, providing subjects irrelevant to the context.

Researchers widely prefer DL techniques in analyzing social media data to detect mental disorders. A study by Gkotsis et al. (2017) classified mental health conditions using feedforward and convolutional neural networks (CNN) based on the Reddit dataset. This Reddit dataset is further grouped through a semi-supervised technique to create subreddits of 11 mental health-related themes. The best-performing CNN model showed 71.37% accuracy in their study. Another study by Islam et al. (2018) on Facebook data detected depression using psycholinguistic measurements and ML algorithms, including the decision tree (DT), which outperformed all other techniques. Zanwar et al. (2022b) conducted a multiclass classification using hybrid and ensemble transformer models on the self-reported mental health diagnoses (SMHD) dataset and the Dreaddit dataset

using BERT, RoBERTa, and bidirectional long short-term memory (BiLSTM), achieving a macro F1-score of 31.40% across 5 folds.

Interpretability of models becomes crucial as classification alone is insufficient for sensitive applications such as mental health analysis. Explainable AI has emerged as an effective technique to address this problem (Gunning et al. 2019). One such technique, the local interpretable model-agnostic explanations (LIME) (Ribeiro et al. 2016), provides a factual implementation to explain respective predictions. This paper applied the LIME algorithm to present a trustable explanation. Hu et al. conducted a multiclass classification on Covid-19 related mental health data using the post hoc system and ante-hoc method to analyze and explain the factors that impact mental health during the pandemic (Hu and Sokolova 2021). Another recent study (Saxena et al. 2022) on CAMS (Garg et al. 2022) dataset used LIME and integrated gradient (IG) methods to find explanations for reasons related to inconsistency in the accuracy of multiclass classification. Our study found meaningful topics within mental health discussions to gain insights and interpret the findings of individuals suffering from mental health illness using a novel technique, BERTopic (Grootendorst 2022c). We analyzed the possible occurrences of topics for four clinically identified mental health issues, including bipolar disorder, anxiety, suicidal thoughts, and depression. The study has been extended using classification techniques to classify the text corpus into each category. Post classification, the trustworthiness of each model's prediction ability was investigated using local explanations for a given instance using explainable AI.

3 Methodology

3.1 Dataset

Given its novelty and accessibility, we have examined the Reddit SWMH (Ji et al. 2022) dataset. The dataset comprised texts from mental health-related reddit subreddits of anxiety, depression, bipolar, and suicidewatch, a total of 54,412 texts in five classes. This dataset is anonymous; hence, the absence of any identifiable details regarding the individuals in the dataset ensures that the privacy and confidentiality of the participants remain uncompromised. The SWMH dataset's purpose is to identify associated mental health conditions-related variables. This study removed the 'self.offmychest' class and duplicate texts. This class was removed as it did not indicate any direct relation to mental health illnesses, such as depression, anxiety, bipolar disorder, and suicidal thoughts. Our study examined 46,103 instances from the four classes, namely self.Anxiety, self.bipolar, self.depression, self.SuicideWatch. This dataset was analyzed to find themes and patterns of suicidality and mental illnesses.

The total number of texts evaluated for each dataset is shown in Fig. 1. We assessed and contrasted the data available on social media for various mental disorders with recurring themes and patterns associated with mental illnesses. The text samples in Table 1 represent each category.

Word clouds were created to visualize the importance of words in the context. The word cloud generated and shown in Fig. 2, using the text corpus, reveals the frequently used words "feel," "life," and "one" to convey emotions. Words such as 'anxiety,' 'depression,' and 'people' tend to stand out since the users have been sharing texts with phrases like "*I feel anxious around people*" and "*Is there anyone to talk about depression?*" for example. Social media users frequently utilize social platforms allowing for more open discussions about delicate subjects like mental health (Yazdavar et al. 2018).

3.2 Process architecture

The process architecture of this experiment consists of preprocessing texts, topic modeling, classification and local explanations, as shown in Fig. 3. The pre-processing pipeline was further segmented into noise removal and normalization. As for noise removal, the datasets were processed to remove duplicate text, links, text in square brackets, terms with numbers, and lowercase text conversion. In addition, stop words and largely standard English terms lacking crucial information and meaningless and misspelled words were removed. The NLTK library (natural language toolkit) was initially used for this purpose, and the tokenization step was then completed. Tokenization is a crucial process of dividing a text into token-sized pieces allowing for interpreting the text's meaning via word analysis. Several library functions, such as sci-kit-learn countvectorizer for ML models and Tensorflow Tokenizer for CNN and LSTM models, were

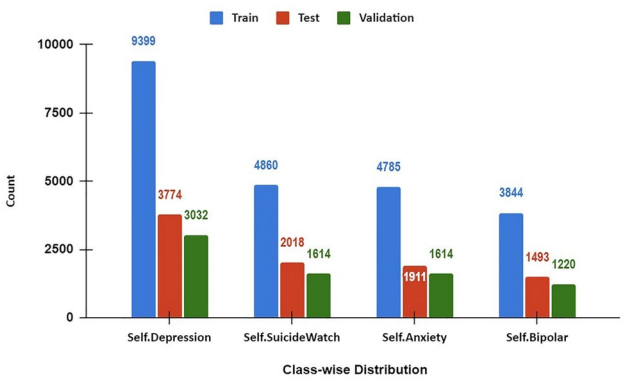


Fig. 1 Class-wise distribution of processed training, testing, and validation sets

Table 1 Samples of text from each class

Label	Text sample
Anxiety	Why do my arms, and muscles jerk when I feel anxious? When my anxiety is particularly bad, I find that my muscle control is horrible. When I try to grip something, my muscles will simply twitch. It feels and appears incredibly strange. Even turning my head in one direction causes it to jerk. I really hate this. It gives me the impression that there is a problem with me
Bipolar	Having recovered from a manic episode and considering scheduling a meeting? Since my fiancé broke up with me two months ago, I've apparently been experiencing manic episodes, and now I'm beginning to collapse
Depression	Right I must speak with someone. Every time I try to talk to my friends about my problems, they either tell me I'm acting silly or that things don't really matter, that nobody cares, or that I'm just being a little bitch. They could be correct, but those things have been bugging me for about two to three months now, and I really need to get it off my chest because it's getting exhausting. Although I frequently consider suicide, I lack the courage to carry it out
SuicideWatch	After a few days of abstinence, I cut today. Even though it wasn't much time, I felt pretty proud of myself for not doing it. And right now, slashing myself and taking my own life is all I can think about. I should have done it a week ago, but I decided not to, which was a terrible error. I overdosed four times in the last five months, and I wish one of them had killed me. I know if I leap off the nearby overpass of the highway, I'll get better results. This world is not suited for me. I have no place here. I just want to fall asleep and stay asleep forever

Fig. 2 Word cloud for training data to visualize important words in the context of mental health

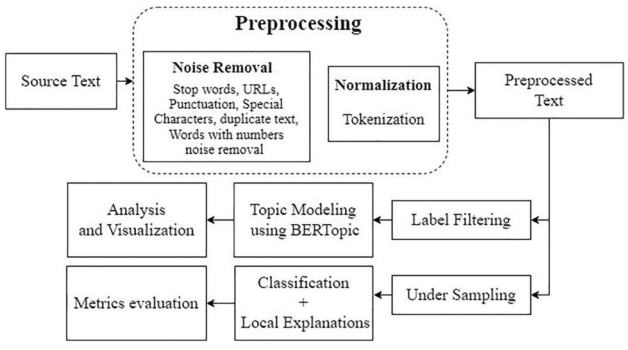


Fig. 3 Architecture of the approach used in this study to identify mental health concerns

used to complete this step. The Transformer models were used with their respective tokenizers provided by the HuggingFace platform. Normalization, a pre-processing step used to enhance the text's quality and prepare it for machine processing, was done. While conducting the classification, undersampling and class weights were applied to deal with the class imbalance.

3.2.1 Topic modeling

By examining the terms in the source texts, topic modeling seeks to identify the themes that permeate a corpus. Topic modeling is an unsupervised ML method of identifying

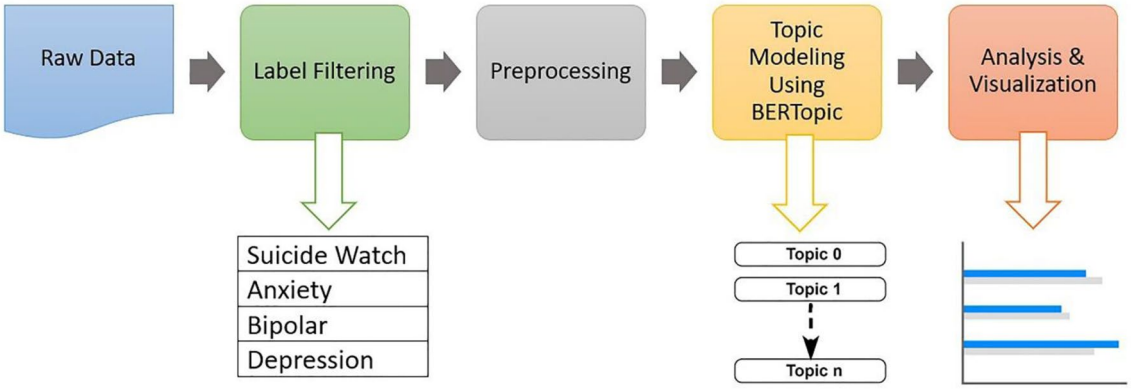


Fig. 4 Architecture of the approach used in this study for topic modeling

or extracting topics by spotting patterns, much like clustering algorithms that separate data into various sections. Transformers, C-TF-IDF and BERTopic, use dense clusters to generate simple to understand topics while preserving critical terms from the topic descriptions. Earlier BERTopic-based studies (Sangaraju et al. 2022; Abuzayed and Al-Khalifa 2021) found that it is adaptable and offers distinct topics compared to LDA. BERTopic has expanded the classical cluster embedding approach by utilizing cutting-edge language models and a class-based TF-IDF(C-TF-IDF) process to create topic representations. The model uses three phases to generate topic representations. Each document is first transformed using a trained language model into its embedding representation. The dimensionality of the generated embeddings is then decreased before clustering them to improve the clustering procedure. Due to the stochastic nature of the UMAP method used for dimensionality reduction while clustering, the model gives different results for training instances. Finally, topic representations are retrieved from the document clusters using a customized class-based form of TF-IDF. The model is significantly more flexible and easier to use because the grouping of documents and generating topic representations is carried out independently (Grootendorst 2022c). C-TF-IDF is utilized in the third stage to extract essential terms at each cluster on a class-based term frequency or inverse document frequency. In BERTopic, TF-IDF was modified to operate on a cluster/category/topic level rather than at the document level to accurately represent the subjects from the bag-of-words matrix. This modified TF-IDF representation, known as C-TF-IDF, considers the differences between documents in each cluster by treating each cluster as a single document rather than a collection of documents. The frequency of the word x in class c, where c is the cluster that we previously established, is extracted. The class-based TF representation is the outcome of this. L1-normalization is used in this representation to consider the variations in topics. A customized version of the algorithm rather than the standard TF-IDF approach is utilized in BERTopic, allowing for a far better representation. The modified algorithm Grootendorst (2022c) proposed for this computation is shown below.

$$W_{\text{tc}} = \text{tf}_{t,c} \cdot log\left(1 + \frac{A}{\text{tf}_t}\right) \tag{1}$$

Here, the term frequency represents the frequency of term t in class c in Eq. (1). The group of documents combined into a single document for each cluster is class C in this instance. Then, evaluating how much knowledge a term contributes to a class, the inverse document frequency is substituted with the inverse class frequency. It is the frequency of term t across all classes divided by the logarithm of the average number of words per class A. It is beneficial when dealing with extensive collections of documents where it is crucial to identify the most important terms for each class.

In topic modeling, a topic is a cluster of words. These words are selected based on their statistical significance in the model. The importance of words is determined by their frequency of occurrence in the cluster and their co-occurrence with other words in the same context. Therefore, it is very helpful for interpreting the groupings produced by any unsupervised clustering method. Bar charts of the C-TF-IDF scores of each topic provide two insights: Scores for each word in each topic's C-TF-IDF and a comparative study of each topic's distribution. The process used for Topic Modeling is depicted in Fig. 4.

3.2.2 Explainable AI

An emerging subset of AI called explainable AI (XAI) focuses on the readability of ML models. An explainable AI model is a set of steps and strategies that enables one to comprehend and accept ML algorithms' results. It entails specifying the model's correctness and transparency and the outcomes of decision-making assisted by the model. LIME (Ribeiro et al. 2016) is a model-independent interpretability method for individual local predictions. Model agnostic refers to the ability to generate explanations for any DL or ML model. The degree to which a person can decipher the reasoning behind a black box model's choice is known as interpretability. Instead of creating an explanation for the whole, LIME's general operating premise is to localize the problem and explain it.

3.3 Evaluation metrics

The evaluation metrics for classification models are accuracy, precision, recall, and F-score. Accuracy is the ratio of the total number of positives to the total number of classes. Precision is the ratio of true positives to the total number of predicted positives. The recall is the ratio of true positives to the actual number of positives. We can get a harmonic mean of these measures using the F1-Score that considers precision and recall. When there is a significant class imbalance, this is very helpful. The formulas are as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{2}$$

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

$$F1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

where TN, TP, FN, and FP represent true negative, true positive, false negative, and false positive, respectively in Eq. (2–5).

In general, classification models are assessed using the above metrics, but since data that have been collected and annotated may have bias compared to the real world. Additionally, LIME is employed to explain by enhancing interpretability. In our case, the SWMH data set is based on the mental health-related subreddits where nuances of language and the usage of the words need a deep understanding of the context in which they were used. Therefore, the traditional metrics might not accurately reflect the primary goal of developing the text classification model. In addition to such metrics, analyzing individual predictions using LIME can be helpful. However, there is no formal agreement on what interpretability signifies in ML and its measurement methods (Molnar n.d.). A local prediction is defined as $L(f, g, x)$ applying the regularization parameters $\Omega(g)$. The L indicates the minimized square loss function and g represents the model applied to class G (class G is defined as the set of models that have the capacity for interpretability). The faithfulness of the explanation g to the original Explanation(x) is measured to get an explanation of a local point x. The formula is as follows:

$$\text{Explanation}(x) = \arg\min_{g \in G} L(f, g, x) + \Omega(g) \tag{6}$$

where $\arg\min_g$ is defined as the value of argument *'g'* for which function $GL(f, g, x)$ attains its minimum.

3.4 Model definitions

Multiclass classification, often known as multinomial classification, categorizes cases into three or more classes in statistical classification and ML. Each data sample can be put into a specific class. A data sample, however, cannot concurrently be a member of more than one class. In other words, the classes in multiclass problems are mutually exclusive.

3.4.1 Logistic regression (LR)

By default, logistic regression can only be used to solve binary classification issues. As a result, it has to be modified to enable multiclass categorization issues. Studies (Pranckevičius and Marcinkevičius 2017) have shown that Logistic regression can be used to achieve higher performance in multiclass text classification. Logistic

Regression is helpful in this situation as it uses a sigmoid function to output a probability between zero and one. In this study, we have adapted a logistic regression model to recognize and forecast a multinomial probability distribution known as multinomial logistic regression.

3.4.2 Random forest (RF)

As an extension of bagging, the random forest (Breiman 2001) algorithm randomly chooses portions of the features utilized in each data instance. Decision trees, which are the foundation of the random forest model, are susceptible to class imbalance. Every tree is supported by a "bag," representing a uniform random data sampling. As a result, a class imbalance will, on average, bias each tree in the same direction and magnitude.

3.4.3 Stochastic gradient descent (SGD)

SGD is an optimization approach for updating a model's weights while being trained. The algorithm updates the weights based on the difference between each input text's actual and projected class. By merging various binary classifiers in an "OVA" (one versus all) framework, it can be used for multiclass classification. A binary classifier is learned for each class that can distinguish it from all other classes. For large datasets, SGD provides a quick and effective optimization approach.

3.4.4 Naive Bayes (NB)

Based on the Bayes theorem, the likelihood of a class given a set of features is proportional to the likelihood of the features given the class. Naive Bayes models are frequently employed for text classification problems since they are quick and simple to implement.

3.4.5 XGBoost (XGB)

GradientBoost was first introduced by Chen in 2016 and XGB (Chen and Guestrin 2016) is its improved version. A gradient boosting framework is used by the decision tree-based ensemble ML algorithm XGB to handle missing values in the training phase and control overfitting and split finding.

3.4.6 Long short-term memory (LSTM)

LSTM is a form of recurrent neural network (RNN). For text data sequences where the word order matters, LSTM models are very well suited. Long-term dependencies can be captured by LSTMs, which also have the ability to handle sequential dependencies in the data.

3.4.7 Convolutional neural network (CNN)

Convolutional Neural Networks can be utilized to solve text classification issues in addition to picture and video analysis. When classifying text, the input is handled like a picture, with each word acting as a feature. The fully connected layers are utilized for prediction, whereas the convolutional layers extract features from the input.

3.4.8 BERT

BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al. 2018) is a pre-trained DL model. BERT models are customized for certain classification tasks after being pre-trained on a sizable corpus of text data. BERT models are ideally suited for text categorization issues since they have a reputation for capturing the context of the words in a text.

3.4.9 MentalBERT

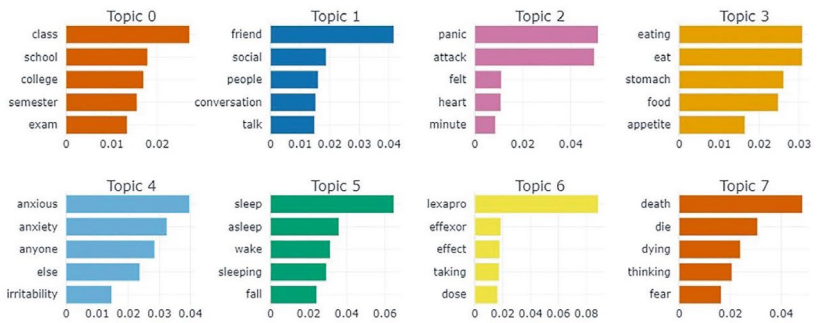
MentalBERT (Ji et al. 2021) uses the BERT model's pre-trained knowledge for contextualized language representations related to mental health. The classification head is trained using a sizable corpus of annotated text to determine the class of a given text document. The MentalBERT model's weights are adjusted during fine-tuning for optimum text categorization performance.

4 Results and discussion

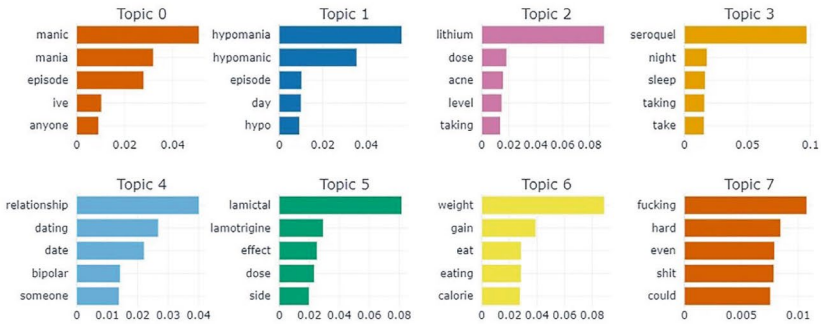
4.1 Topic modeling

We investigated four commonly identified mental health problems, examining the recurring themes and vocabulary that emerged. Plots for the eight most popular topics under each corresponding mental health condition are shown in columns of Fig. 5. Each topic cluster's five most common words are plotted on each bar chart. Each row of the bar graph with the C-TF-IDF score shows the most common terms from each cluster. It can be concluded from the results that the subjects like *class, school, holiday or exam* can trigger the symptoms of mental health problems. These results can be used to aid the therapy and medication processes and to empathize with mental health patients.

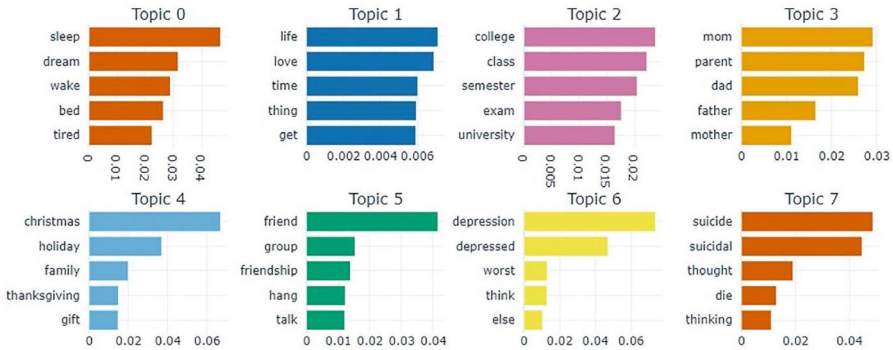
Recurring themes are observed in each class. For instance, topics related to college, work, and relationships frequently appear in every category examined. Demographically speaking, the reason for this can be that Reddit users are mainly young, tech-savvy people who use social media platforms more often to discuss their issues daily. Thus, this study can be extended to focus exclusively on the problems



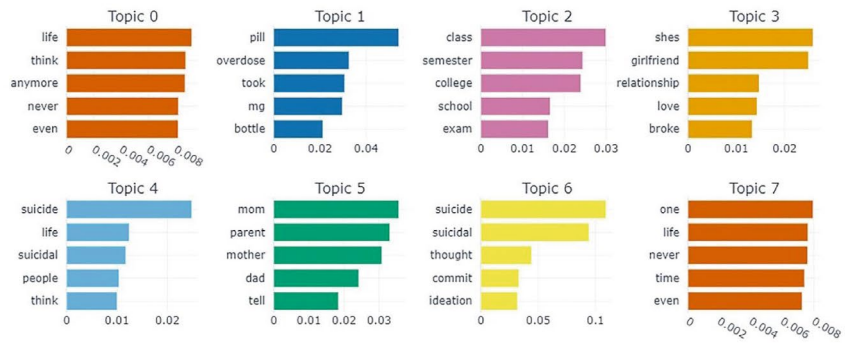
(a) Anxiety



(b) Bipolar



(c) Depression



(d) Suicide

Fig. 5 Graph of recurring words with their C-TF-IDF scores for each class

of youth in modern society that trigger mental health issues and trauma. The themes about bipolar disorder stand out from the rest, as people with bipolar disorder seek professional help compared to the rest. Thus, the topics related to medication and therapy were prominent in that category.

However, it is noteworthy that the BERTopic model has several shortcomings (Grootendorst 2022c). It presumes that each document contains one topic, while in reality, documents may have numerous topics. The word corpus was filtered using the label mental health issue, and five similar motifs in topics were found through qualitative analysis. In each health problem, the topic clusters were related. We could gather subjects that were self-contained and coherent as a result. Given the highly topical nature of Reddit's subreddits, it is expected that some topics, like bipolar disorder studies, are more specific to a particular subreddit than others. The present research facilitates comparing and examining overlapping vocabulary and underlying semantic attributes over time. The relative frequency variation of various clusters can be observed in a given mental health condition, such as depression or suicide. The procedure outlined also enables us to assess conversational shifts related to various topics within a related category and compare it with the rest. It would be helpful for mental health practitioners to understand how various conversations evolve or how they respond to particular events differently.

4.2 Classification

Classification techniques were used to perform Multiclass classification of the text corpus into each category. Since the data set is considered mutually exclusive, each text is only in one category. However, the cases may fall under two categories simultaneously. For instance, a depressed patient might also be suffering from suicidal tendencies at the same time. In addition, in some instances, it was noticed that texts must be of a substantial length to gain enough context to classify correctly.

Using several ML, DL and transformer models for classification allowed us to examine context-based against non-context-based classifier performance. The existence of an imbalanced class distribution within this dataset has repercussions on the model's effectiveness. Because the model exhibits a bias toward the majority class, its capacity to effectively learn from the minority class is compromised. To tackle this challenge, both undersampling and class weight techniques were applied to the dataset. Figure 6 showcases a comparison of the outcomes before and after addressing the data imbalance, providing a visual representation of the effects of these strategies. The domain-specific MentalBERT outperformed all other models (context and non-context-based methods). Table 2 displays the average performance evaluation of multiclass classifiers. The models performed better after undersampling the majority-class data, with the best-performing model MentalBERT achieving 76.70% mean accuracy (best-performing results are in bold in Table 2), outperforming the results published in previous studies (Ji et al. 2022, 2021) by more than 4%. To evaluate the model's classification performance, an accuracy error bar graph (Fig. 7) has been plotted for all nine classification models. Classical Models such as NB, RF, and transformer-based models BERT and MentalBERT have shown very low variance in accuracies compared to CNN and LSTM. This low variance shows that the model's performance is consistent and generalizes well across all the subsets of the dataset. Confusion matrix has also reported in Fig. 8 to understand the statistics of performance for best-performing model. MentalBert has a identify the true labels and false positives and negatives are very less for most of the classes. In literature (Ji et al. 2022), a relation network (RN) based on

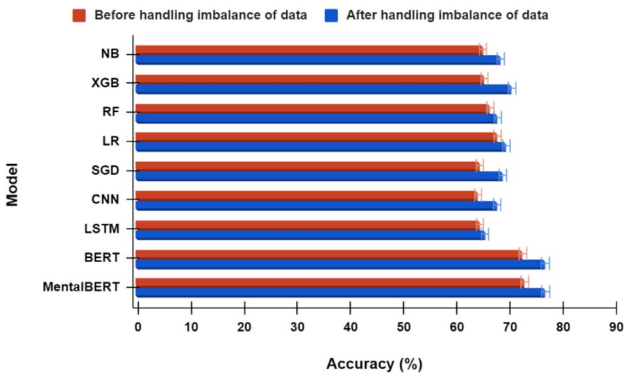


Fig. 6 Comparison of models on the basis of accuracy before and after handling of imbalanced data

Table 2 Results of classification performance metrics for all nine models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
NB	68.23	68.87	68.23	68.32
XGB	70.38	74.49	68.13	70.53
RF	67.69	67.99	67.99	67.40
LR	69.30	69.92	69.30	69.55
SGD	68.63	71.54	68.63	68.77
CNN	67.56	68.54	66.35	67.41
LSTM	65.29	67.40	61.40	64.20
BERT	76.62	76.52	76.56	76.56
MentalBERT	**76.70**	**76.66**	**76.69**	**76.63**

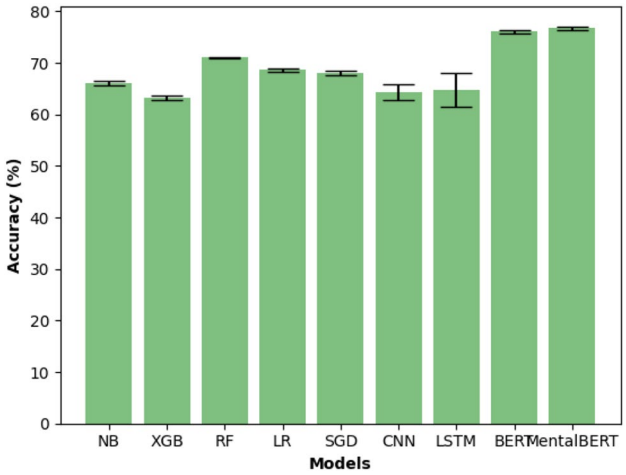


Fig. 7 Error bar graph for accuracy on all nine classification models

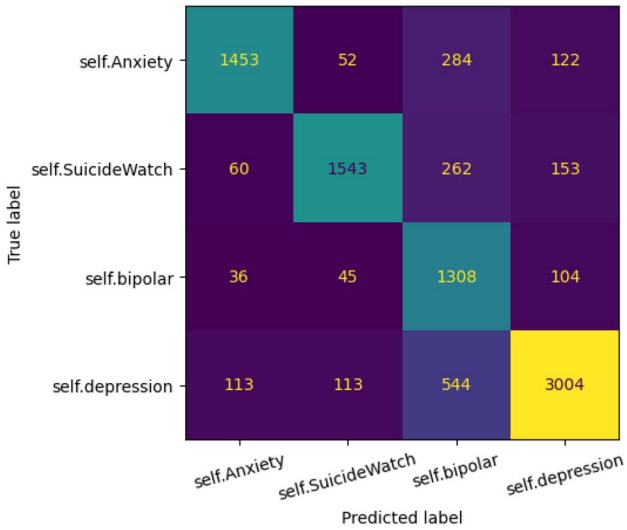


Fig. 8 Confusion matrix for best-performing model (MentalBERT)

the attention technique was used to classify the text. This RN model achieved 64.74% accuracy and 64.78% F1-score. Another study (Ji et al. 2021) applied MentalRoBERTa on the SWMH dataset and attained 72.16% F1-score.

One more experiment was designed to understand further the decision various models took. In this study, nine models are considered for classification. The SGD model outperformed four models (RF, LSTM, XGB and NB) but lagged behind the other four models (BERT, CNN, LR and MentalBERT) in terms of accuracy. Hence, 50 misclassified instances of SGD from all classes were selected and checked for their LIME explanations ("6."). Also, the other 8 models were applied to check their efficiency on these misclassified instances. As shown in Fig. 9, MentalBERT correctly classified twenty-four instances out of 50, whereas RF could correctly classify only sixteen. These posts have mixed topic themes and class-representative keywords that may confuse models, yet MentalBERT performed well compared to all other classification techniques used in this study. Despite having a smaller training corpus size compared to BERT, MentalBERT demonstrates comparable or superior performance on mental health datasets. This is due to MentalBERT's capability for capturing the specific context and sentiments prevalent in mental health-related text. While BERT demands substantial computational resources for training and inference due to its general-purpose applicability and extensive pre-training corpus, MentalBERT's specialized training results in reduced resource utilization during both the training and inference stages. This decreased resource overhead can prove advantageous in situations where computational resources are limited or when deploying the model on devices with constrained memory capacities.

An additional dataset ("depression" 2021) extracted from Facebook comments and posts on mental health-related issues, is analyzed within the same experimental setup to validate the proposed approach. This dataset comprises 6982 social media posts and is publicly available on the Kaggle platform. The dataset consists of two classes and has been annotated using a majority voting approach involving undergraduate students. All nine classification models were applied to this dataset, and the results are presented in Fig. 10. The classification outcomes were compared with the results reported in the literature (Hassan et al. 2021) and Kaggle notebooks to ensure consistency and reliability.

4.3 Explainable AI

LIME determines whether a system produces insightful justifications for its classification. The features employed in this work can be used to support ML assessments and contribute to the design of a reliable user interface. Figure 11 shows the

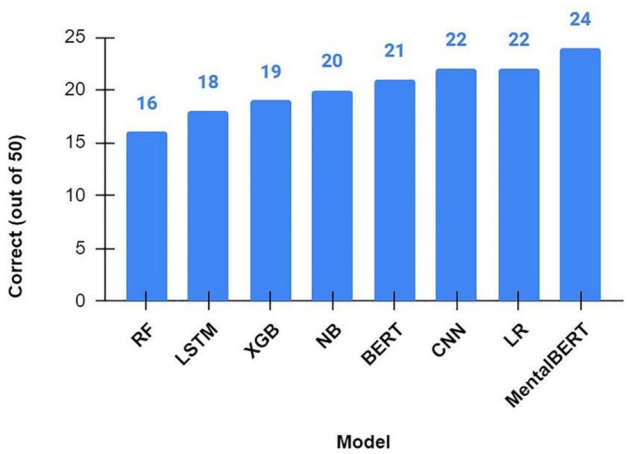


Fig. 9 Count of correct posts classified by all models except SGD

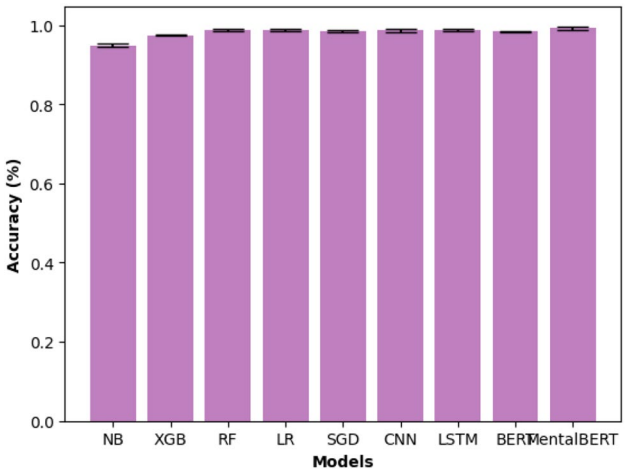


Fig. 10 Error bar graph for accuracy on all nine classification models on the Kaggle dataset

LIME algorithm outcomes for each model. The total class probabilities, the top features with their probabilities, and automatically highlighted features in the sample input text using LIME are considered to understand the results.

The class probabilities for a given model range from 0–0.99%. For instance, as shown in Fig. 11, a test text (I want to turn into an alcoholic and die in 6 years) taken for the depression class for the XGB model gives a 0.52% probability of '*self.suicidewatch*' label because it has the word *die* and the top features with their probabilities explain each class probability indicating the content is biased toward *suicidewatch* class. According to the classification from the XGB and NB models and explanations from LIME, the text consists of *suicidewatch* class (Fig. 11a, b). However, the true class (*self.depression*) of the text sample was not the same as the model's prediction because the sample text has few words (less information) and does not contain class-representative keywords (depression or depressed). The DL—LSTM model performed well for this sample text, while CNN could not because it gets confused with the words like '*deleted'* and '*die'*. RF, LR and SGD also performed well in this text but failed in others (see "6.").

Context-based models have tried to understand the context and give a higher probability to the true class (Fig. 11i). To our knowledge, the class-representative words or repetitive words (Fig. 5) have contributed to the classification of non-context-based models. While context-based models such as BERT and MentalBERT have classified text based on the overall context of the post, not on a few representative words.

The approach used in this study demonstrates a thorough and thoughtful process to enhance the performance and interpretability of NLP models. Firstly, data pre-processing techniques were implemented to clean the input data and reduce noise by removing stopwords, URLs, and punctuation. Additionally, duplicate texts were eliminated to ensure that the model was not biased by redundant information. Data imbalance, which could have led to biased predictions, was addressed through data undersampling and class weight assignment. Undersampling techniques helped balance the distribution of different classes, while class weights ensured that the model received adequate exposure to the minority class. It significantly enhanced the model's capacity to accurately predict all classes, regardless of their prevalence, compared to the previous study (Ji et al. 2021). This approach was further enriched by incorporating BERT for topic modeling. BERT, as a context-aware model, extracted meaningful themes and intricate details from text data related to mental disorders. By harnessing the capabilities of these models, we gained valuable insights into the underlying patterns within the data, thereby advancing our comprehension of mental disorders and their subtleties. Employing LIME for an explainable NLP analysis bolstered the transparency and reliability of this approach. LIME enabled us to analyze and elucidate the model's predictions by pinpointing crucial features and elucidating their contributions. It facilitated a deeper understanding of the model's decision-making process, rendering it more understandable and trustworthy in contrast to the prior study (Ji et al. 2021). Our approach amalgamates data pre-processing, the rectification of class imbalance, the utilization of BERT for topic modeling, the application of LIME for explainable NLP, culminating in a comprehensive workflow that not only enhances the model's predictive performance but also furnishes transparency in its results.

5 Conclusion and future work

This work captured complex relationships in a wide range of textual data, notably social media posts, using NLP techniques. It examined the distribution of topics in the Reddit-based online community for mental health. The social media data was analyzed for the significant themes related to mental health issues using BERTopic. The topics like relationships, exams, and school were found to impact mental health conditions directly. In addition, the themes that emerged from each category's topics were found to fit into several identifiable patterns. A few patterns related to school, friends and exams are recurring, frequently employing vocabulary from dialogues about mental health. The differences and similarities among the themes covered by the corpus of texts in each community were examined.

A comparative analysis using nine classical state-of-the-art classification techniques was done to classify the mental health disease that may help professional mental health therapists by automating the textual analysis

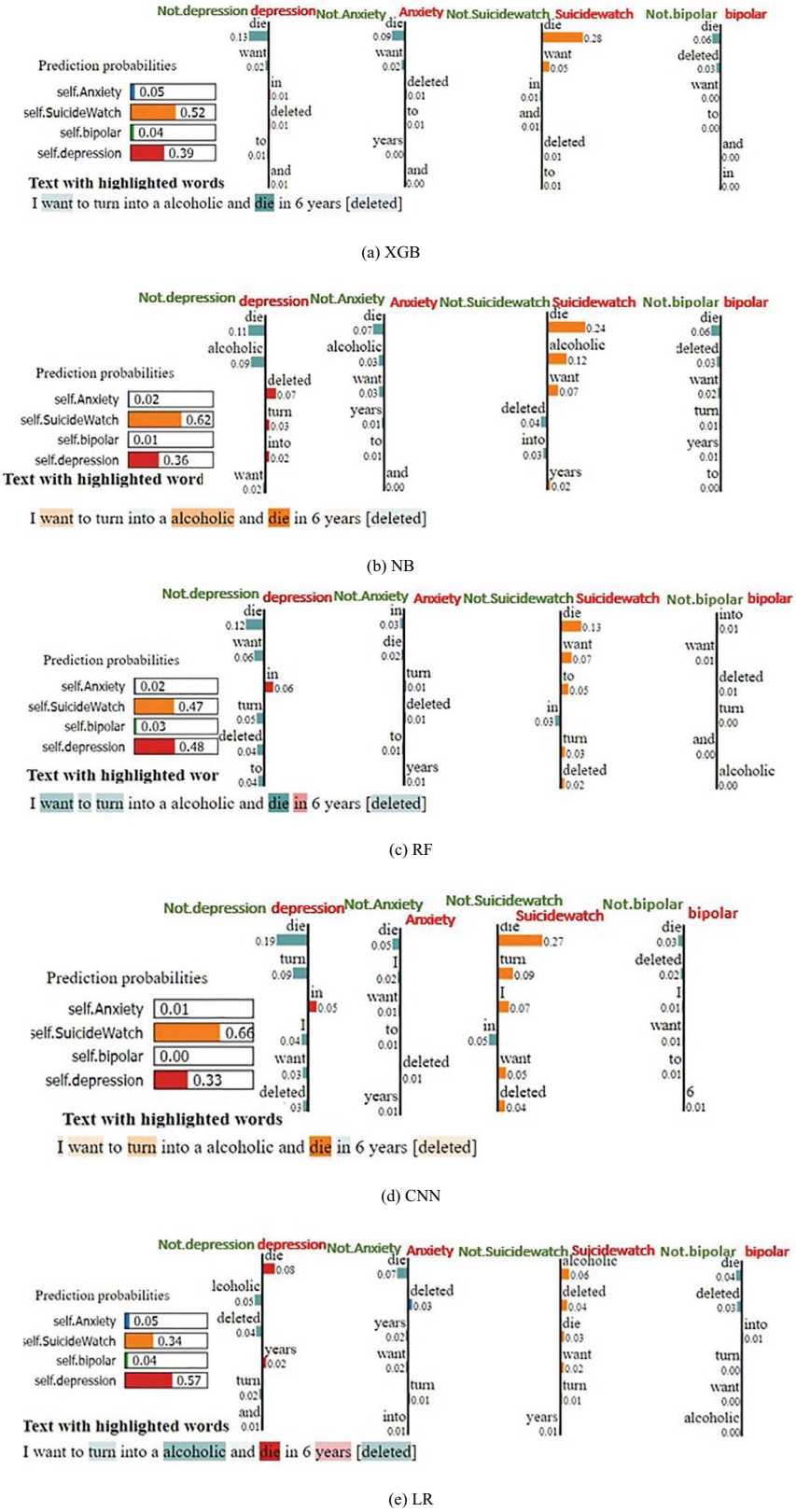
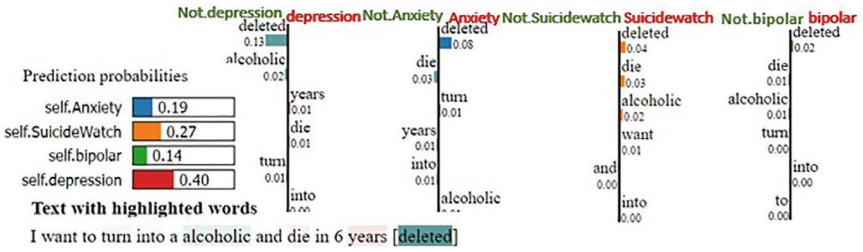
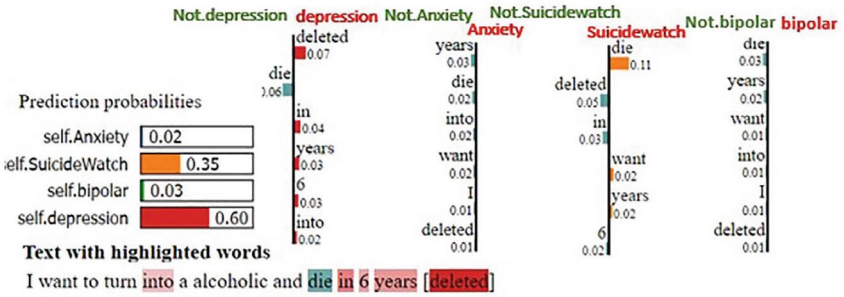


Fig. 11 Prediction probabilities of all nine models using LIME technique to explain model's decision

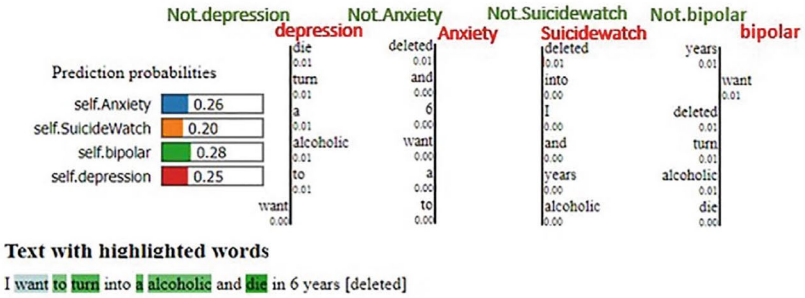
Fig. 11 (continued)



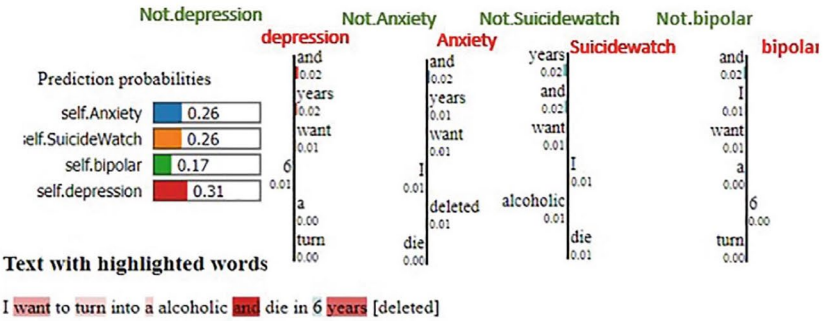
(f) SGD



(g) LSTM



(h) BERT



(i) MentalBERT

process. The transformer model, MentalBERT performed the best, considering the significance of context, and achieved 76.70% accuracy, significantly higher than the reported accuracies. The performance of the models and techniques was evaluated using the explainable AI technique–LIME. It aids in clarifying why a model is producing particular predictions and fostering trust. The results from this study can help identify users who may be at threat of mental health disease. This approach can differentiate users with potential mental health diseases, namely bipolar, depression, anxiety, and even suicidal ideation, based on symptoms extracted from their posts. The recognition of these recurring patterns and themes holds significant potential to assist mental health professionals in their practice. By identifying language patterns that signal the onset of mental health challenges, professionals can intervene proactively and extend support to individuals facing such risks. This early intervention has the capacity to curtail the progression of these mental challenges into more severe conditions. Mental health professionals can apply these insights to fine-tune their interventions and outreach efforts, increasing their effectiveness. Moreover, this study can serve as a valuable resource for policymakers, furnishing crucial data to shape mental health policies and initiatives. Comprehending the central themes and issues impacting mental health, as elucidated in this study, can inform the development of targeted mental health programs. In this research, we utilized solely English-language Reddit posts as our primary social media data source, which represents a constraint in our study. Enhancing the findings could involve incorporating information from a broader array of social media platforms beyond those using English, as well as interactions with individuals from a more extensive range of socio-economic backgrounds. This approach can be applied to larger datasets and more variable investigations in the future to improve the generalization of the study. It can be developed to explore the relationships between the uncovered topics related to mental health, like potential suicide identification or drug addiction.

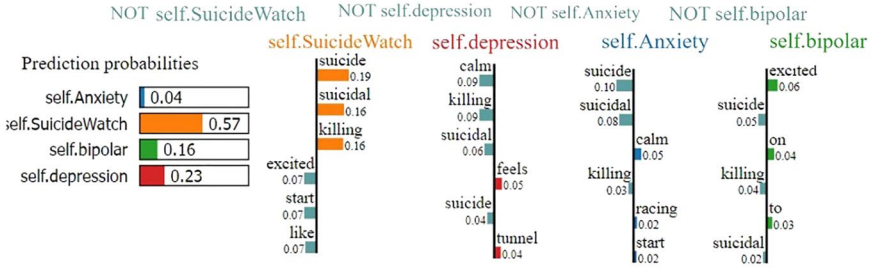
Appendix 1

A sample text (lets called 'A') '*Feelings of excitement when thinking about suicide The past few days I've been dwelling on stupid past mistakes and getting a lot of invasive suicidal thoughts, but lately when I get them I start to feel excited like I'm going some place like an amusement park or something. When I think about killing myself it feels so right, rather than feeling dark and wrong. My heart starts racing really fast and I get tunnel vision at times and think about hanging myself, but then I immediately calm down and the thoughts stop after just 10 s or so. I'm not sure if this is common or not.*' is taken from test set belongs to *self.depression* class has classified for all nine classification models. In Table 3, classification insights have been reported for all the models used in this study. XGB, NB, CNN, SGD and LSTM wrongly classified the sample text as self.SuicideWatch (because it contains keyword *suicide*), self.Anxity(it contains words *racing* and *calm*), self.bipolar, self.SuicideWatch, and self.Anxity, respectively. However, RF, LR, BERT, MentalBERT classified the this text correctly as *self.depression*.

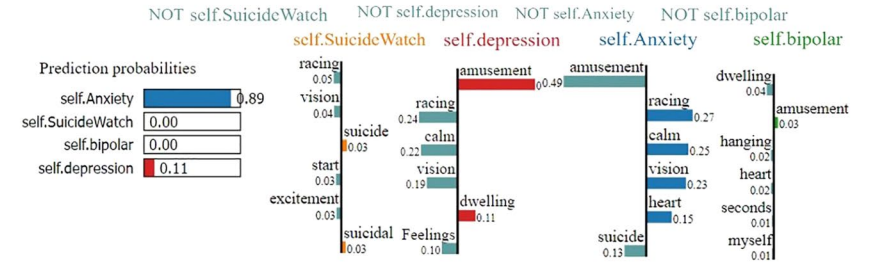
The true class for another sample text (lets called 'B') '*Back Well, after my close call that was the subject of my first post, here, I vowed to do something with my life, but it looks like I've gone around in a big fucking circle. I've liked her since high school, I've spent years becoming best friends with her and her boyfriend and she just doesn't fuckin want it. There is literally no end to this Romeo style Petrarchan lover bullshit but as much as I'm self-aware and introspective, I'm also a slave to my biochemical pitfalls. I'm seriously considering it.'* is self.SuicideWatch. Non-context-based models XGB and LR correctly identified this text where as deep learning models CNN and LSTM both misclassified sample text 'B.' Both context-based models BERT and MentalBERT also predicted actual class as shown in Table 4. Although BERT and MentalBERT gave same prediction probabilities for all classes but the selected different words to predict true class.

One more sample text (lets called 'C') from class self.depression has been taken '*depression and anxiety Attack!! Earlier I was feeling depressed and now I am having a anxiety attack. my stomach is churning, and I just want to lay down and die in all Honesty I feel like bashing my head against a wall maybe then I wouldn't feel like this. I have a sense of dread. I don't have any clue why I am feeling this way, but I just want to cry myself to sleep at least. but no. that's not even a possibility because I can't stop fidgeting and stirring and my mind won't calm down. not even breathing in and out helps.'* to understand the model's classification decision. The text contains both anxiety and

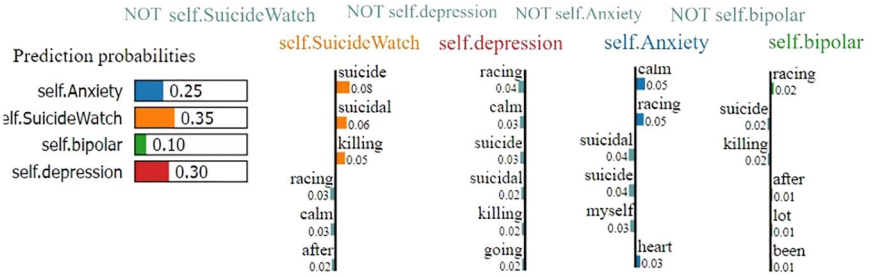
Table 3 Lime explanation for sample text A of each model



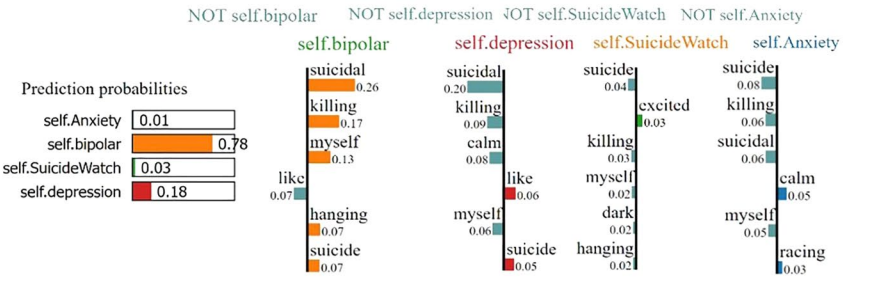
XGB



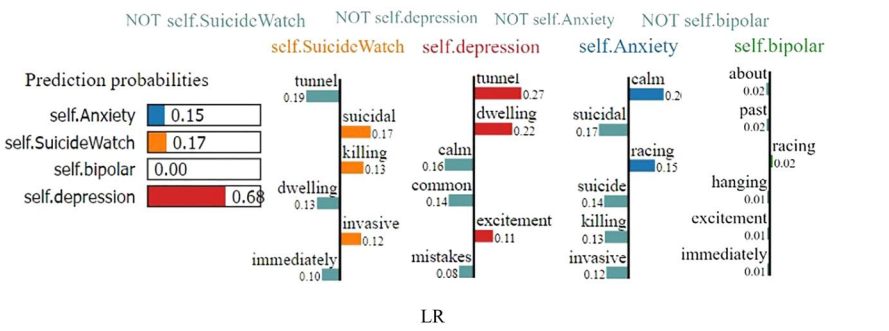
NB



RF



CNN



LR

Table 3 (continued)

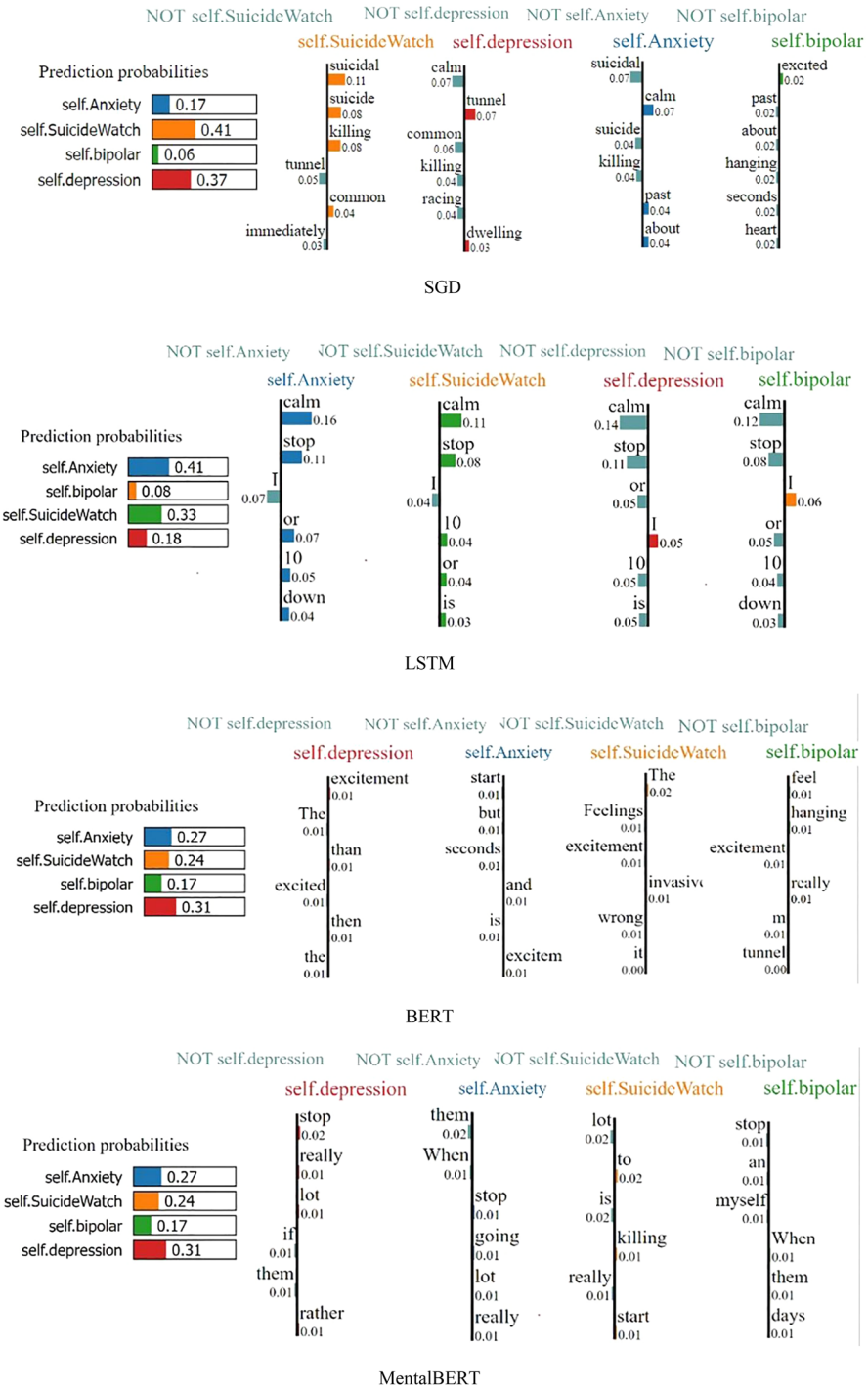


Table 4 Lime explanation for sample text B of each model

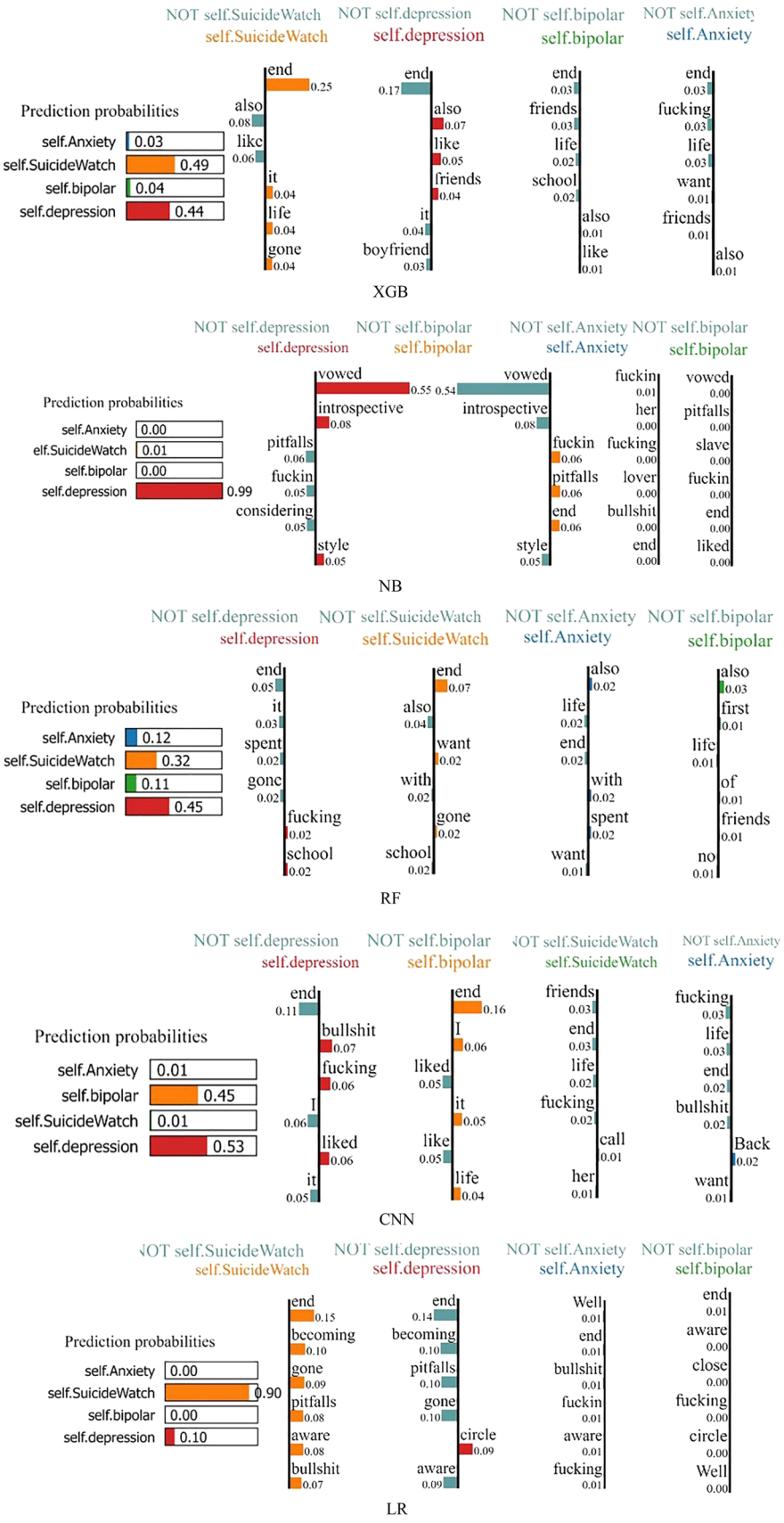


Table 4 (continued)

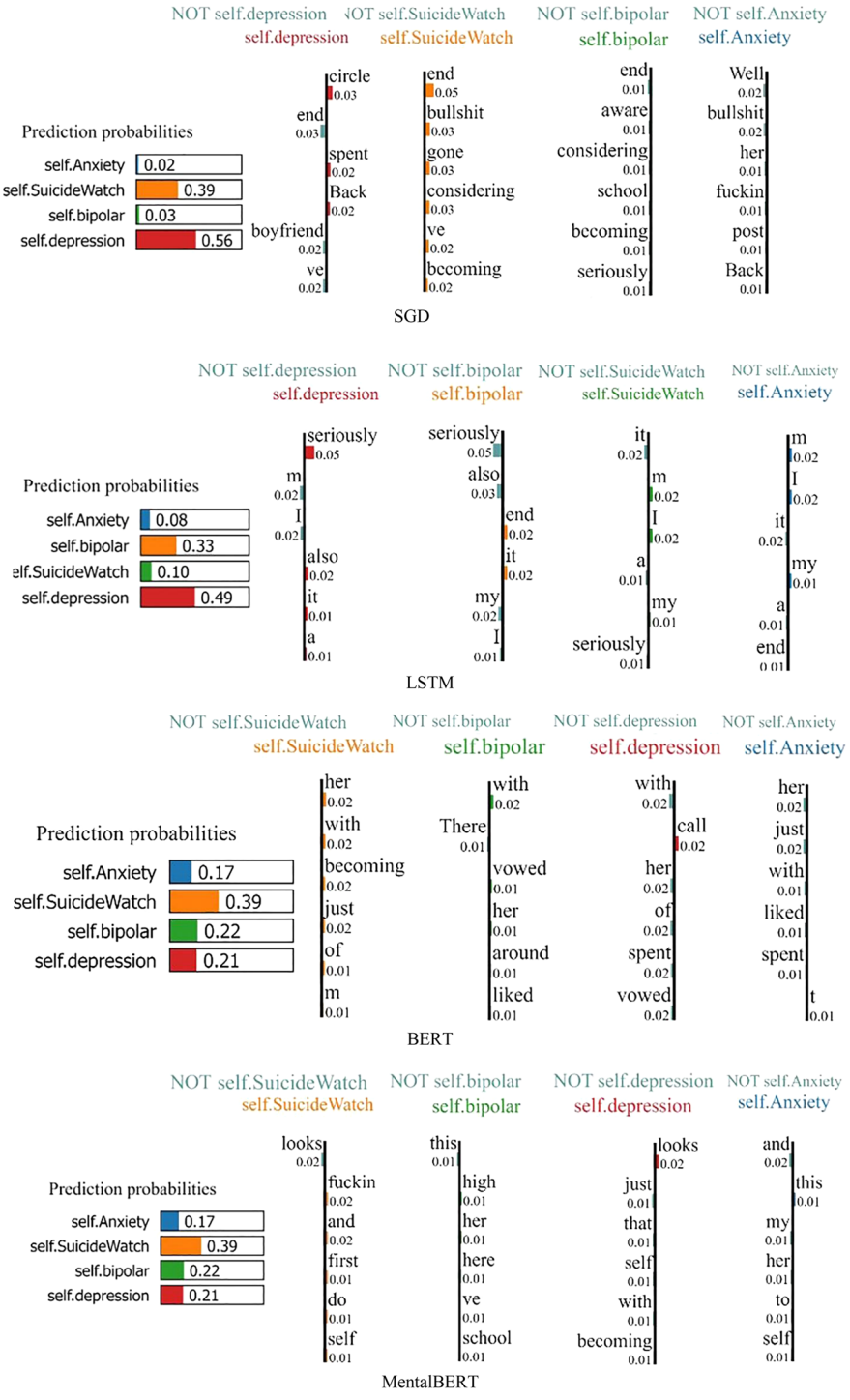


Table 5 Lime explanation for sample text C of each model

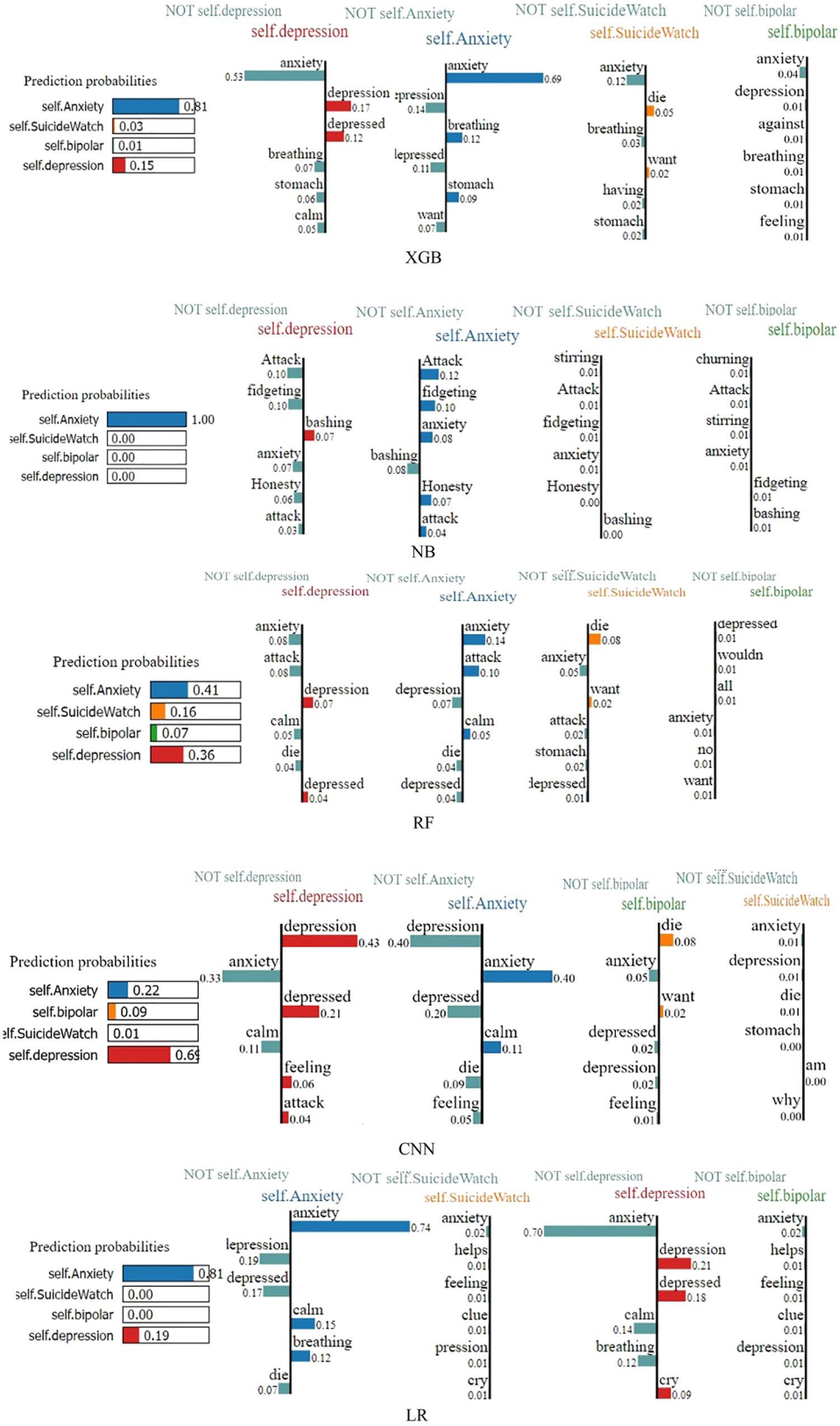
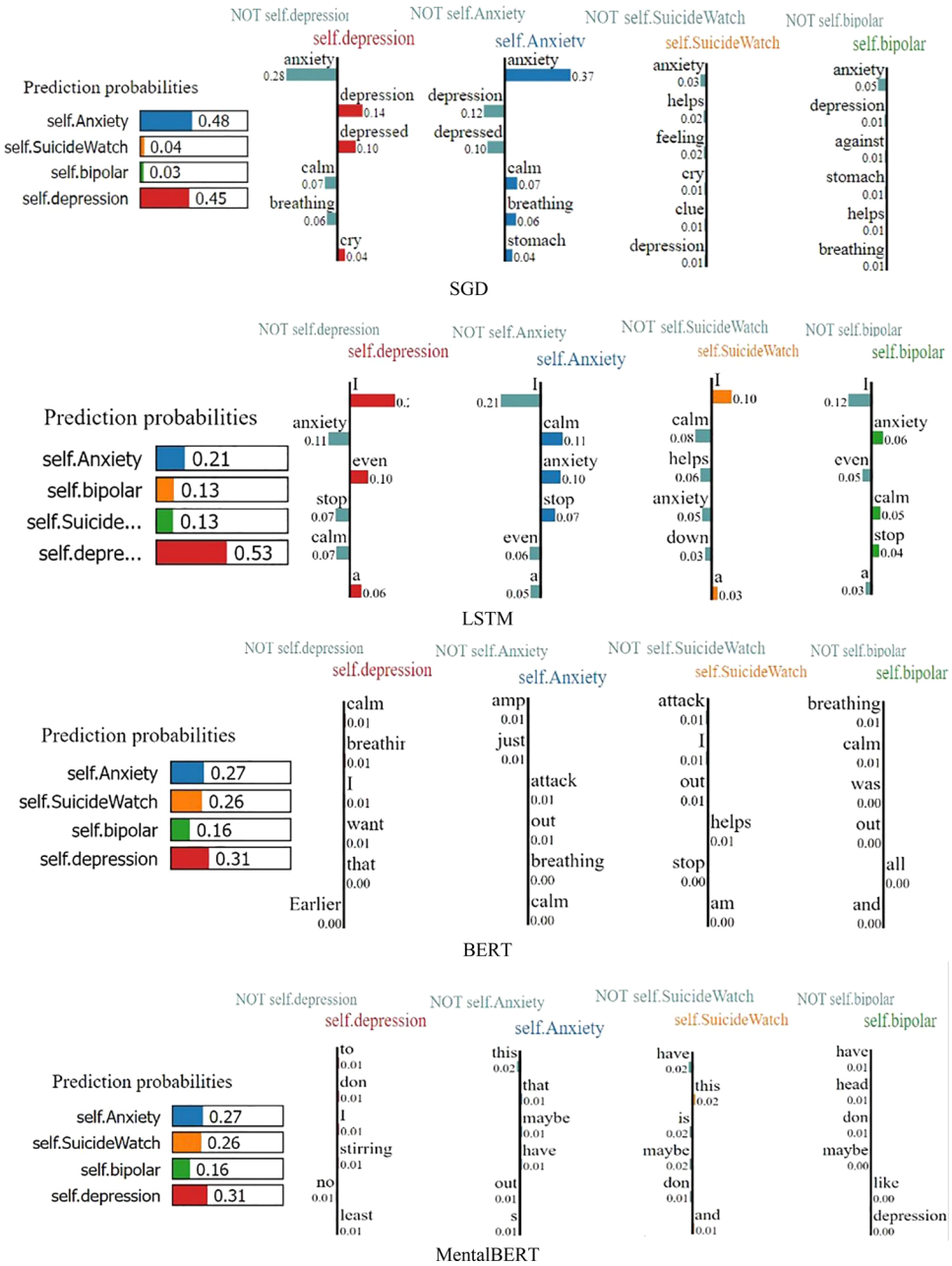


Table 5 (continued)



depression related words. Therefore, models XGB, NB, RF, SGD and LR predicted class self.Anxiety (Table 5). However, deep learning and transformer models CNN, LSTM, BERT AND MentalBERT classified actual class.

Acknowledgements The authors would like to thank Project SAMARTH, an initiative of the Ministry of Education (MoE), Government of India, at the University of Delhi South Campus (UDSC), for their support.

Authors contributions PDT was contributed to methodology and writing—original draft, NA was contributed to conceptualization, methodology, visualization, validation, and writing—original draft, VV was contributed to methodology and writing—original draft, GJS was contributed to conceptualization, visualization, validation, and writing—review and editing, **SS** was contributed to writing—review and editing, AP was contributed to validation and writing—review and editing.

Funding No funding was received for conducting this study.

Data availability The dataset analyzed in the present study can be available on request using the link: https://doi.org/10.5281/zenodo.6476179.

Code availability The two models of BERT and MentalBERT for this dataset have been released in Hugging Face hub and they can be downloaded for further studies.

MentalBERT: https://huggingface.co/tiya1012/swmh4_mtb
BERT: https://huggingface.co/tiya1012/swmh4_bert

Declarations

Conflict of interests The authors have declared that they have no conflict of interest.

Ethics approval and consent to participate Not applicable.

References

Abuzayed A, Al-Khalifa H (2021) BERT for Arabic Topic Modeling: An Experimental Study on BERTopic Technique. Procedia Computer Science 189:191–194. https://doi.org/10.1016/j.procs.2021.05.096

Alotaibi W, Alomary F, Mokni R (2023) COVID-19 vaccine rejection causes based on Twitter people's opinions analysis using deep learning. Soc Netw Anal Min 13:62. https://doi.org/10.1007/s13278-023-01059-y

Benrouba F, Boudour R (2023) Emotional sentiment analysis of social media content for mental health safety. Soc Netw Anal Min 13:17. https://doi.org/10.1007/s13278-022-01000-9

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. Journal of machine Learning research, 993–1022.

Boettcher N (2021) Studies of Depression and Anxiety Using Reddit as a Data Source: Scoping Review. JMIR Ment Health 8(11):e29487. https://doi.org/10.2196/29487

Breiman L (2001) Random Forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. ArXiv DOI 10(1145/2939672):2939785

Dao B, Nguyen T, Venkatesh S, Phung D (2015) Nonparametric discovery of online mental health-related communities. In: IEEE International Conference on Data Science and Advanced Analytics (DSAA), Paris, France, pp 1–10. https://doi.org/10.1109/DSAA.2015.7344841

Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint https://doi.org/10.48550/arXiv.1810.04805

"depression", Kaggle.com, 2021, [online] Available: https://www.kaggle.com/datasets/sahasourav17/students-anxiety-and-depression-dataset.

Garg M, Saxena C, Krishnan V, Joshi R, Saha S, Mago V, Dorr BJ (2022) CAMS: an annotated corpus for causal analysis of mental health issues in social media posts. arXiv preprint. https://doi.org/10.48550/arXiv.2207.04674

Gemmell J, Isenegger K, Dong Y, Glaser E, Morain A (2019) Comparing Automatically Extracted Topics from Online Mental Health Disorder Forums. In:International Conference on Computational Science and Computational Intelligence (CSCI), pp 1347–1352. https://doi.org/10.1109/CSCI49370.2019.00252

Gkotsis G, Oellrich A, Velupillai S, Liakata M, Hubbard TJ, Dobson RJ, Dutta R (2017) Characterisation of mental health conditions in social media using Informed Deep Learning. Sci Rep 7:45141. https://doi.org/10.1038/srep45141

Grootendorst M (2022) BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint. https://doi.org/10.48550/arXiv.2203.05794

Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang GZ (2019) XAI—Explainable artificial intelligence. Sci Robot,4(37). https://doi.org/10.1126/scirobotics.aay7120

Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC (2017) Detecting depression and mental illness on social media: an integrative review. Curr Opin Behav Sci 18:43–49. https://doi.org/10.1016/j.cobeha.2017.07.005

Hanna F, Barbui C, Dua T, Lora A, van Regteren AM, Saxena S (2018) Global mental health: how are we doing? World Psychiatry 17(3):367–368. https://doi.org/10.1002/wps.20572

Hassan MM, Khan MAR, Islam KK, Hassan MM, Rabbi MMF (2021) Depression Detection system with Statistical Analysis and Data Mining Approaches. In: International Conference on Science & Contemporary Technologies (ICSCT), Dhaka, Bangladesh, pp 1–6, https://doi.org/10.1109/ICSCT53883.2021.9642550.

Hu Y, Sokolova M (2021) Explainable multi-class classification of the camh covid-19 mental health data. arXiv preprint. https://doi.org/10.48550/arXiv.2105.13430

Huang X, Wang S, Zhang M, Hu T, Hohl A, She B, Gong X, Li J, Liu X, Gruebner O, Liu R, Li X, Liu Z, Ye X, Li Z (2022) Social media mining under the COVID-19 context: Progress, challenges, and opportunities. International Journal of Applied Earth Observation and Geoinformation: ITC Journal 113:102967. https://doi.org/10.1016/j.jag.2022.102967

Islam MR, Kabir MA, Ahmed A, Kamal ARM, Wang H, Ulhaq A (2018) Depression detection from social network data using machine learning techniques. Health Inf Sci Syst 6(1):8. https://doi.org/10.1007/s13755-018-0046-0

Ji S, Li X, Huang Z, Cambria E (2022) Suicidal ideation and mental disorder detection with attentive relation networks. Neural Comput Appl 34(13):10309–10319. https://doi.org/10.1007/s00521-021-06208-y

Ji S, Zhang T, Ansari L, Fu J, Tiwari P, Cambria E (2021) MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. arXiv https://doi.org/10.48550/arXiv.2110.15621

Kamarudin NS, Beigi G, Liu H (2021) A study on Mental Health Discussion through Reddit. In: International conference on software engineering and computer systems and 4th international conference on computational science and information management, ICSECS-ICOCSIM. https://doi.org/10.1109/ICSECS52883.2021.00122

Kathy L, Agrawal A, Choudhary A (2015) Mining Social Media Streams to Improve Public Health Allergy Surveillance. In:IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'15), pp-815–822. https://doi.org/10.1145/2808797.2808896

Kilbourne AM, Beck K, Spaeth-Rublee B, Ramanuj P, O'Brien RW, Tomoyasu N, Pincus HA (2018) Measuring and improving the quality of mental health cara global perspective. World Psychiatry 17(1):30–38. https://doi.org/10.1002/wps.20482

Kim J, Lee J, Park E, Han J (2020) A deep learning model for detecting mental illness from user content on social media. Sci Rep 10:11846. https://doi.org/10.1038/s41598-020-68764-y

Kotenko I, Sharma Y, Branitskiy A (2021) Predicting the Mental State of the Social Network Users based on the Latent Dirichlet Allocation and fastText. In:IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), pp 191–195. https://doi.org/10.1109/IDAACS53288.2021.9661061

Lin YS, Tai LK, Chen AL (2023) The detection of mental health conditions by incorporating external knowledge. J Intell Inf Syst. https://doi.org/10.1007/s10844-022-00774-w

Liu D, Feng XL, Ahmed F, Shahid M, Guo J (2022) Detecting and measuring depression on social media using a machine learning approach: systematic review. JMIR Ment Health 9(3):e27244. https://doi.org/10.2196/27244

Molnar, C (2022) Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 2. https://christophm.github.io/interpretable-ml-book

Pranckevi˘cius T, Marcinkevi˘cius V (2017) Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. Baltic J Modern Comput 5(2):221

Qi Y, Shabrina Z (2023) Sentiment analysis using Twitter data: a comparative application of lexicon- and machine-learning-based approach. Soc Netw Anal Min 13:31. https://doi.org/10.1007/s13278-023-01030-x

Ren L, Lin H, Xu B, Zhang S, Yang L, Sun S (2021) Depression detection on reddit with an emotion-based attention network: algorithm development and validation. JMIR Med Inf 9(7):e28754. https://doi.org/10.2196/28754

Ribeiro MT, Singh S, Guestrin C (2016) Why Should I Trust You?: Explaining the Predictions of Any Classifier. arXiv preprint. https://doi.org/10.48550/arXiv.1602.04938

Rizvi STR, Ahmed S, Dengel A (2023) ACE 2.0: A Comprehensive tool for automatic extraction, analysis, and digital profiling of the researchers in scientific communities. Soc Netw Anal Min 13:81. https://doi.org/10.1007/s13278-023-01085-w

Saha B, Nguyen T, Phung D, Venkatesh S (2016) A framework for classifying online mental health-related communities with an interest in depression. IEEE J Biomed Health Inf 20(4):1008–1015. https://doi.org/10.1109/JBHI.2016.2543741

Sangaraju VR, Bolla BK, Nayak DK, Kh J (2022) Topic modelling on consumer financial protection bureau data: an approach using BERT based embeddings. arXiv preprint. https://doi.org/10.48550/arXiv.2203.05794

Saxena C, Garg M, Ansari G (2022) Explainable causal analysis of mental health on social media data. Explainable causal analysis of mental health on social media data. arXiv preprint. https://doi.org/10.48550/arXiv.2210.08430

Stein DJ, Palk AC, Kendler KS (2021) What is a mental disorder? An exemplar focused approach. Psychol Med 51(6):894–901. https://doi.org/10.1017/S0033291721001185

Suicide data: Mental Health and Substance Use (2021). https://www.who.int/teams/mental-health-and-substance-use/data-research/suicide-data Accessed 5 January 2023.

Verma R, Chhabra A, Gupta A (2023) A statistical analysis of tweets on covid-19 vaccine hesitancy utilizing opinion mining: an Indian perspective. Soc Netw Anal Min 13:12. https://doi.org/10.1007/s13278-022-01015-2

Wainberg ML, Scorza P, Shultz JM et al (2017) Challenges and opportunities in global mental health: a research-to-practice perspective. Curr Psychiatry Rep 19(5):28. https://doi.org/10.1007/s11920-017-0780-z

World mental health report: Transforming mental health for all - executive summary (2022). https://www.who.int/publications/i/item/9789240049338 Accessed 28 December2022.

Yazdavar AH, Mahdavinejad MS, Bajaj G, Thirunarayan K, Pathak J, Sheth A (2018) Mental health analysis via social media data. In: IEEE international conference on healthcare informatics (ICHI), NY, USA, pp 459–460. https://doi.org/10.1109/ICHI.2018.00102

Zanwar S, Wiechmann D, Qiao Y, Kerz E (2022) Exploring Hybrid and Ensemble Models for Multiclass Prediction of Mental Health Status on Social Media. arXiv preprint. https://doi.org/10.48550/arXiv.2212.09839

Zhou J, Zogan H, Yang S, Jameel S, Xu G, Chen F (2021) Detecting community depression dynamics due to covid-19 pandemic in Australia. IEEE Trans Comput Soc Syst 8(4):982–991. https://doi.org/10.1109/TCSS.2020.3047604

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

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

Indian Journal of Pure & Applied Physics
Vol. 61, November 2023, pp. 910-914
DOI: 10.56042/ijpap.v61i11.3151

Insights into Structural, Electronic and Thermoelectric Properties of $ZnTMN_2$ (TM= Zr and Hf): A First-Principles Study

Rajesh Kumar, Ramesh Kumar, Sangeeta, Kulwinder Kumar & Mukhtiyar Singh*
Department of Applied Physics, Delhi Technological University (DTU), Delhi 110 042 India

Received 28 June 2023; accepted 14 August 2023

We employ first-principles calculations combined with semi classical Boltzmann transport theory to investigate the structural, electronic, and thermoelectric properties of $ZnTMN_2$ (TM= Zr and Hf). The negative value of formation energy confirms the stability of these compounds. We used Tran Blaha modified Becke Johnson approximation to calculate electronic properties. $ZnZrN_2$ and $ZnHfN_2$ are having indirect bands of magnitude 2.77 eV and 3.31 eV, respectively. The positive value of the Seebeck Coefficient at all studied temperatures confirms its p-type nature. The thermal conductivity slightly decreases with a rise in temperature in $ZnHfN_2$ as compared to $ZnZrN_2$. The observed value of the figure of merit is 0.80 and 0.81 at 500 K and 600 K for $ZnZrN_2$ and $ZnHfN_2$, respectively. The high figure of merit of Hf and Zr-based nitrides make them a potential material for thermoelectric energy harvesting applications.

Keywords: DFT; Becke johnson approximation; Electronic properties; Seebeck coefficient; Thermoelectric energy harvesting

1 Introduction

The unprecedented demand of energy led to over-consumption of natural resources resulting in devastating natural calamities in recent times. The green energy harvesting techniques are emerging as an alternate to conventional energy resources to sage guard the environment. Thermoelectric (TE) materials have been considered as one of the potential candidates for the green energy harvesting given its peculiar merits over the other conventional energy resources *i.e.*, wind energy and solar energy[1,2]. The efficiency these materials characterized by a dimensionless quantity[1], figure of merit which is represented as $ZT = S^2\sigma T/\kappa_e + \kappa_l$. Here, S, σ, κ_e, and κ_l denote the Seebeck coefficient, electrical conductivity, electronic thermal conductivity, and lattice thermal conductivity, respectively. To find effective TE material the extensive research has been carried out on various materials, including oxide[3], Zintl[4], and nitrides[5]. Among the semiconductors, nitrides are considered as the promising candidate owing to their high radiation resistance, high chemical stability and thermal stability which are advantageous for TE applications[6,7]. We examined the structural, electronic and transport properties using first-principles calculation in combination with Boltzmann transport theory of $ZnTMN_2$ (Tm = Zr, Hf) as a potential TE material.

*Corresponding author: (E-mail: msphysik09@gmail.com)

2 Computational Methods

The ab-initio calculations were performed using Density Functional Theory (DFT) based wien2k code[8]. The generalized gradient approximation (GGA) scheme of Perdew-Burke-Ernzerhof (PBE) was used to optimize the lattice parameters of the studied systems[9]. As GGA scheme underestimate the band gap, so for realistic prediction of various properties, we were employed Tran-Blaha modified Becke-Johnson (TB-mBJ) potential[10]which incorporates exchange and correlation energy that uses semi local quantities. The strict energy convergence of 0.0001 Ry was used. A Γ-centred k-mesh of 15×15×11was employed to sample the Brillouin zone. The transport properties were obtained by solving semi classical Boltzmann transport equation under constant relaxation time approximation which are implemented in the BoltzTraP code[11]. A denser k-mesh of 45×45×33 was used for transport calculations.

3 Results and Discussion

3.1 Structural and Electronic Properties

The $ZnTMN_2$ (Tm = Zr, Hf) exist in hexagonal structure with space group *P3m1*[12]. The crystal structure exists in TM^{4+}, Zn^{2+}, and N^{3-} environment.

In which the TM^{4+} is coupled to six N^{3-} atoms to produce TMN_6 octahedra that share corners with six ZnN_4 tetrahedra, edges with six TMN_6 octahedra, and edges with three ZnN_4 tetrahedra. There are three short bonds and three long bonds of bond length 2.13 Å and 2.33 Å, respectively in Zr-N, this bond length reduces in the case of Hf-N to 2.11 Å and 2.30 Å, respectively. The optimised lattice parameters are listed in Table 1.

The model crystal structure and optimized lattice parameters are presented in Fig. 1. The optimised structure is further used to evaluate the electronic and transport properties.

Table 1 — The lattice parameters (in Å) and formation energy of $ZnTMN_2$ (Tm = Zr, Hf).

	a	b	c	Formation energy (eV)
$ZnZrN_2$	3.26	3.26	5.25	-1.044
$ZnHfN_2$	3.25	3.25	5.27	-1.171

To further examine the properties of the purposed compound it is vital to check the thermodynamical stability. The stability analysis is done by the formation energy which is given by the formula[4]:

$$E_f = \{E_{total} - n(E_{Zn} + E_{TM} + E_N)\}/M \quad \ldots (1)$$

where, M is the total number of elements in the unit cell, E_{total} is the total energy of the cell, E_{Zn}, E_N, and E_{TM} is the energy of individual Zn, N, TM (Zr and Hf) atom, respectively.

The negative formation energy confirms the thermal stability of the proposed materials. A large negative value of formation energy shows the more stable behaviour of $ZnHfN_2$.

The electronic band structure of $ZnTMN_2$ (Tm = Zr, Hf) is obtained using TB-mBJ as shown in Fig. 2. Itis observed that $ZnZrN_2$ and $ZnHfN_2$ have indirect band gap for of 2.77 eV and 3.31 eV, respectively along M-Γ points.

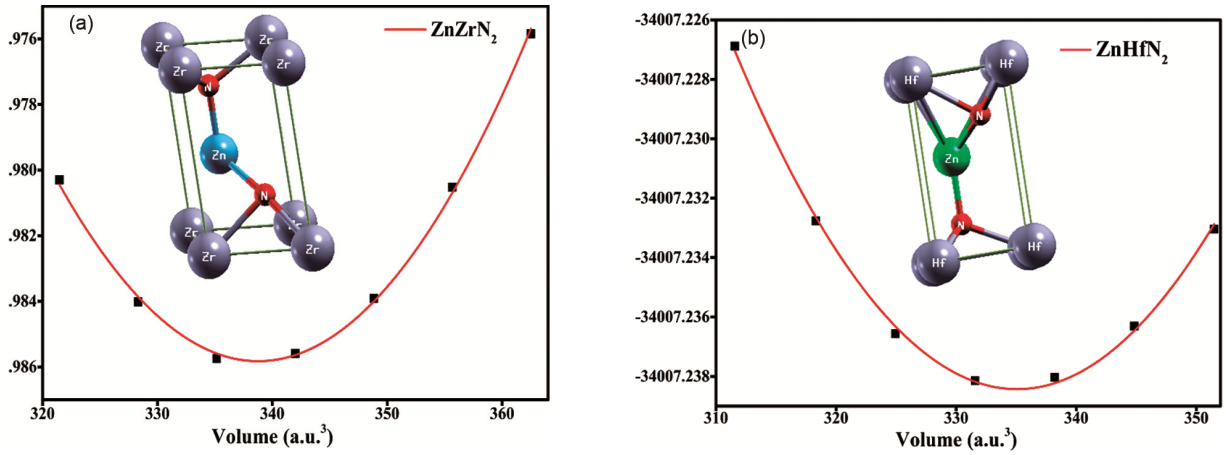


Fig. 1 — The energy vs volume curve and crystal structure (inset) of $ZnTMN_2$ (Tm = Zr, Hf).

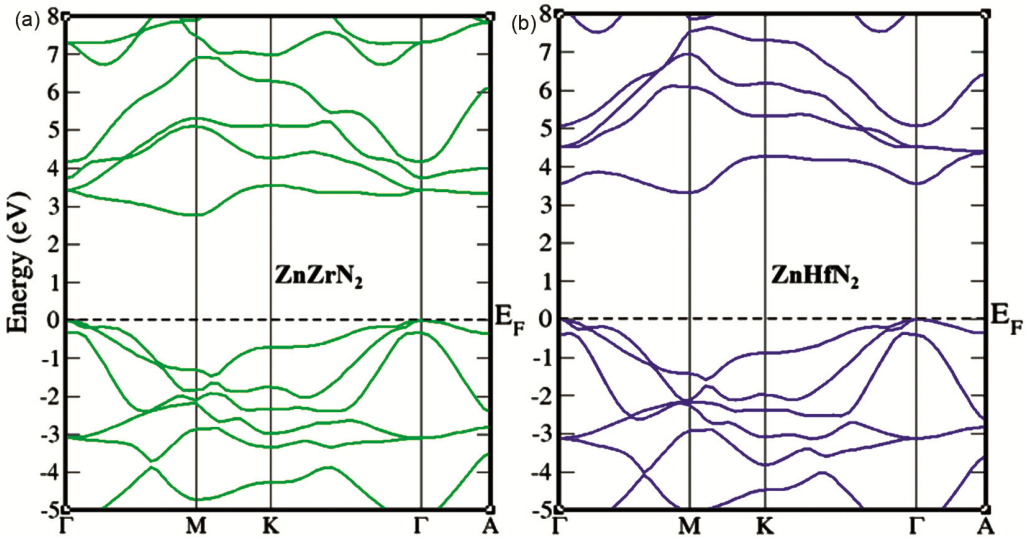


Fig. 2 — The electronic band structure of $ZnTMN_2$ (Tm = Zr, Hf). The Fermi level set to 0 eV.

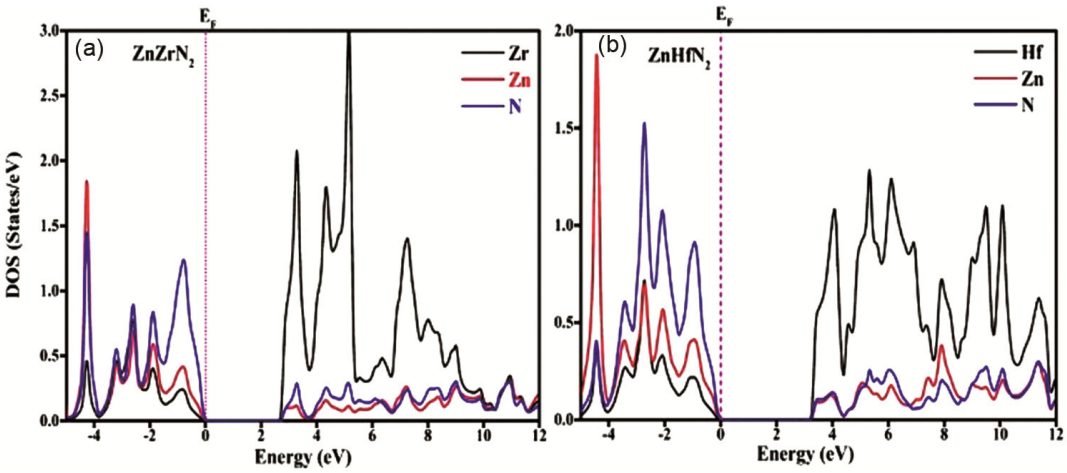


Fig. 3(a,b) — The density of states of $ZnTMN_2$ (Tm = Zr, Hf).

The density of states (DOS), Fig. 3, shows that in the valence band (VB) the N-atomic states dominant and in the deep VB the Zn-atomic states show major contribution over others. Whilst, in the conduction band (CB) Zr-atomic states more prominent over others. The magnitude of hybridization is decreases as we move from $ZnZrN_2$ to $ZnHfN_2$ this might be due to complex bonding nature of the compounds.

We have analyzed the electronic charge density contour plot along (010) plane, which shows that the mix covalent and ionic bonding in both compounds. The extent of covalent bonding is more in case of $ZnZrN_2$ (Fig. 4). The large hybridization in Zr-Zn-N atomic states in DOS is also reflected in the charge density plot.

3.2 Transport Properties

The accurate band gap estimation is important to investigate the TE behaviour of these materials for realistic comparison with future experiments. We have obtained the precise band gap using TB-mBJ potential. The TE parameters are calculated through the solution of the BTE under rigid band approximation. The S, σ, κ are the main parameters which effect TE performance. We have investigated the effect of temperature on these parameters. The variation of S with the temperature describes the ability of material to generate electric potential. The value of S rises with increase in temperature and attain the maximum value 276.59 μVK^{-1} at 900 K and 274.57 μVK^{-1} at 950 K for $ZnZrN_2$ and $ZnHfN_2$, respectively (Fig. 5 (a)).This slight variation in the values of S might be high DOS of Zr near the fermi

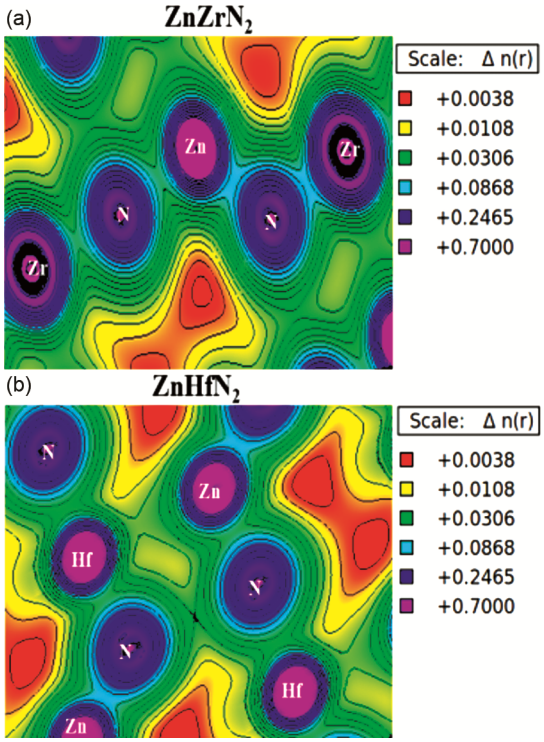


Fig. 4 — The electronic charge density of $ZnTMN_2$ (Tm = Zr, Hf).

level. The positive value of S represents the p-type behaviour of both materials.

The behaviour of σ/τ and κ/τ with temperature are shown the Fig. 5(b,c). At room temperature the value of σ/τ is $0.11x10^{19}$ $\Omega^{-1}m^{-1}s^{-1}$, whereas at 1200 K it increases to $1.32x10^{19}$ $\Omega^{-1}m^{-1}s^{-1}$ and $1.26x10^{19}$ $\Omega^{-1}m^{-1}s^{-1}$ for $ZnZrN_2$ and $ZnHfN_2$, respectively. The calculated value of κ/τ at room temperature is $0.029x10^{15}$ $WmK^{-1}s^{-1}$ whereas at 1200 K the values are

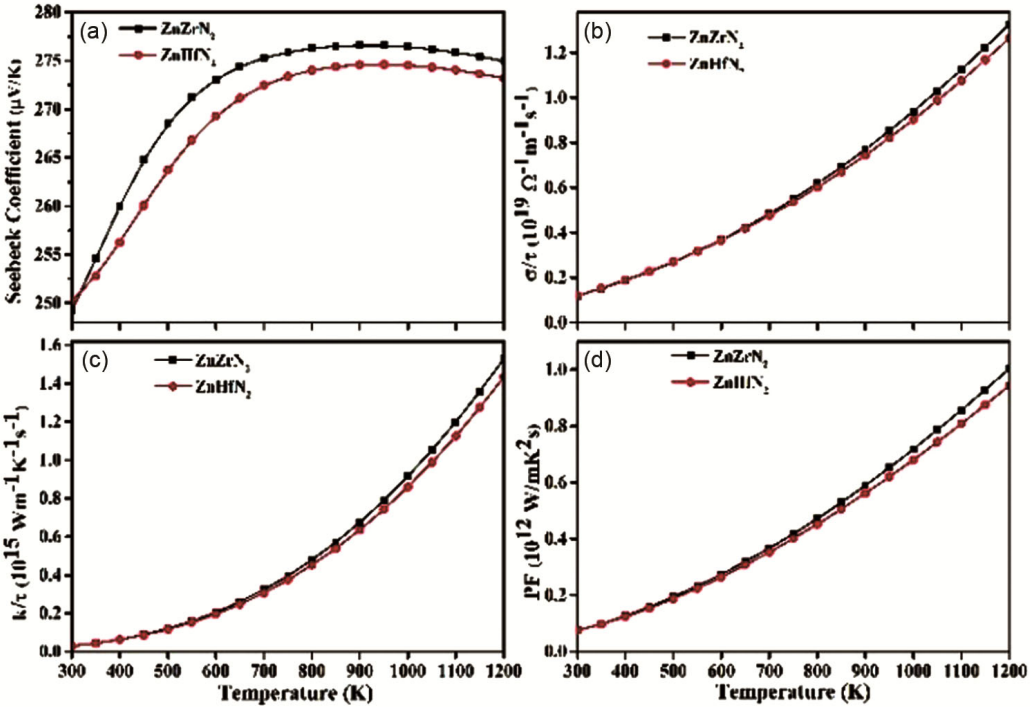


Fig. 5 — The variation of (a) Seebeck coefficient (b) electrical conductivity (c) electronic thermal conductivity (d) power factor of$ZnTMN_2$ (Tm = Zr, Hf) at different temperatures.

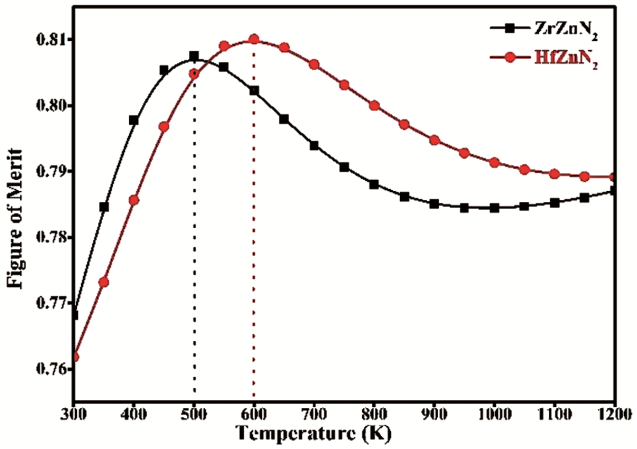


Fig. 6 — Figure of merit of $ZnTMN_2$ (Tm = Zr, Hf).

$1.529x10^{15}$ $WmK^{-1}s^{-1}$ and $1.43x10^{15}$ $WmK^{-1}s^{-1}$ for $ZnZrN_2$ and $ZnHfN_2$, respectively.

The variation in the PF with temperature is shown in Fig. 5 (d). The observed value at room temperature is $0.074x10^{12}$ $Wm^{-1}K^2s^{-1}$, whereas at 1200 K the obtained values are $1x10^{12}$ $Wm^{-1}K^2s^{-1}$ and $0.94x10^{12}$ $Wm^{-1}K^2s^{-1}$.

The variation of ZT with temperature is shown in Fig. 6. At room temperature the ZT for both materials is ~0.76. As temperature increases ZT starts increases up to 500 K with magnitude 0.80 and 600 K with magnitude 0.81 for $ZnZrN_2$ and $ZnHfN_2$, respectively. The slight increase in ZT for $ZnHfN_2$ is due to its low electronic thermal conductivity.

4 Conclusion

We have examined the structural, electronic and TE properties of $ZnTMN_2$ (Tm = Zr, Hf). The negative value of formation energy has confirmed the thermodynamical stability of purposed compounds. The indirect band of magnitude 2.77 eV and 3.31 eV have calculated for $ZnZrN_2$ and $ZnHfN_2$, respectively. The magnitude of charge sharing has been confirmed by DOS and electronic charge density contour plot. The observed value of the figure of merit are 0.80 and 0.81 at 500 K and 600 K for $ZnZrN_2$ and $ZnHfN_2$, respectively. The high figure of merit of Hf and Zr-based nitrides make them a promising candidate for thermoelectric green energy harvesting applications.

Acknowledgement

We acknowledge NSM for providing computing resources of 'PARAM SMRITI' at NABI, Mohali, which is implemented by C-DAC and supported by the Ministry of Electronics and Information Technology (MeitY) and Department of Science and Technology (DST), Government of India.

References

1 Snyder G, Toberer E, *Nature Mater*, 7 (2008) 105.
2 Sanad M F, Shalan A E, Abdellatif S O et al., *Top Curr Chem* (Z), 378 (2020) 48.
3 Kumar R, Kumar R, Vij A & Singh M, *Phys Sc*r, 97 (2022) 075813.
4 Sangeeta, Kumar R & Singh M, *J Mater Sci*, 57 (2022) 10691.
5 Eklund P, Kerdsongpanya S & Alling B, *J Mater Chem C*, 4 (2016) 3905.
6 Hultman L, *Vaccum*, 57 (2000) 1.
7 Berguzinov A, Kozlovskiy A, Kenzhina I & Shlimas D I, *Nanomaterials*, 12(2022) 1789.
8 Blaha P, Schwarz K, Madsen G K H, Kvasnicka D and Luitz J 2018 WIEN2k an augment plane wave+ local orbitals program for calculating crystal properties User's Guide, ISBN 3-9501031-1-2.
9 Perdew J P, Burke K & Ernzerhof M, *Phy Rev Lett*, 77 (1996) 3865.
10 Tran F & Blaha P, *Phys Rev Lett*, 102 (2009) 226401.
11 Madsen G K H & Singh D J *Comput Phys Commun*, 175(2006) 67.
12 Tholander C, Andersson C B A, Armiento R, Tasnádi F & Alling B, *J Appl Phys*, 120 (2016) 225102.

Evolutionary Intelligence
https://doi.org/10.1007/s12065-023-00886-4

SPECIAL ISSUE

Integrated shannon entropy and COPRAS optimal model-based recommendation framework

Neha Punetha[1] · **Goonjan Jain**[1]

Received: 25 March 2023 / Revised: 6 September 2023 / Accepted: 18 September 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract
Multi-criteria decision-making (MCDM) techniques are increasing in product recommendation decisions, which typically entail several factors. This study aims to demonstrate the application of a novel strategy based on MCDM techniques as the core element of a consumer Decision Support System by suggesting the most appropriate items from a given set of alternatives. Ranking products based on online product ratings and consumer preferences is an important area of study, but there are currently few studies on this topic. This paper proposes a method for ranking products using multi-attribute online ratings. We propose a novel mobile recommendation-ranking system-based (MCDM) method to recommend the best alternative. Our proposed model differs from previous works in the following ways: (a) Rating information of each feature is used to identify user preferences and complementary criteria; (b) Criteria weights are determined by Shannon entropy (c) The complex proportional assessment method is employed to rank the alternatives and solve the best mobile recommendation problem. (d) The sensitivity study results showed that the rankings produced by the various MCDM approaches were highly consistent with the rankings of the evaluated compromise candidates. Demonstration of the proposed approach through a mobile phone selection case study. In our experiments, we have found that our approach provides a reliable ranking while reducing time and space complexity, indicating that our optimization model is accurate and efficient. With its superior product comparison skills and ability to offer a recommendation to the user as a final ranking of alternatives, a decision-making system like this may prove to be the optimal long-term answer for e-commerce sites and websites.

Keywords MCDM · COPRAS · Alternatives · Correlation · Sensitivity analysis

1 Introduction

As e-commerce has grown in popularity and access has become more widely available, many people are choosing to do their shopping online rather than going to a brick-and-mortar store. As a result, sales of online goods have risen. In today's global marketplace, identifying customer needs is a significant concern for companies competing globally. Customers' needs must be identified accurately and quickly for businesses to remain relevant and competitive in the market. Recently, customers and planners in this field have been paying more attention to customer feedback data. The first reason is the difficulty in relying on product manufacturer information and the second reason is the increasing number of unfiltered customer feedback about their purchased products [1]. Because of this, customer ratings are now one of the most important considerations for consumers when making purchase decisions. The ease with which this information can be accessed has resulted in less satisfied customers. Business owners and researchers are increasingly using data from reviews to understand better what their customers want and how they perceive the strengths and weaknesses of their services. Only by doing so do they remain relevant in an increasingly competitive industry. As customers increasingly shop online, they have to spend much time accessing the reviews through the increasing number of daily feedbacks. As a result, customers have difficulty using product reviews to make purchasing decisions. Therefore, the rating is introduced to assess the product's qualities quickly and quickly access a large number of products in the shortest amount of

✉ Neha Punetha
nehapunetha80@gmail.com

Goonjan Jain
goonjan_jain@hotmail.com

[1] Present Address: Department of Applied Mathematics, Delhi Technological University, Shahbad Daulatpur, Main Bawana Road, Delhi 110042, India

Published online: 05 October 2023

time. A rating scale can quantify buyers' satisfaction with their purchases and shopping. Individual and customer ratings make purchasing a product more straightforward and enjoyable [2].

The majority of real-world decision-making difficulties call for the simultaneous assessment of numerous competing criteria and objectives. To provide a solution that requires consensus, MCDM integrates the performance of alternatives across several contradictory criteria and qualitative and quantitative standards. MCDM can improve engineering decision-making at every stage, from conception through production [3]. However, it shines most brightly in high-tech sectors where competitive advantage and product uniqueness are usually achieved by small adjustments to material performance [4].

1.1 MCDM for recommendation system

A Multi-Criteria Decision Making (MCDM)-based recommendation system uses decision-making algorithms to provide consumers with personalized choices. Unlike conventional recommendation systems that depend on user preferences and historical data, MCDM-based systems consider many criteria and aspects that affect user decision-making [5]. These include user preferences, item characteristics, cost, quality, and other factors. Many techniques like AHP, TOPSIS, EDAS, etc. can be used to evaluate and rank alternatives based on various criteria. Recommending the best product is combined in this study for ranking alternatives based on ratings [6].Initially, recommendations were made based on consumers' stated preferences concerning the items' core characteristics. Online sources were then used to extract product performance scores based on their feature ratings [7]. An MCDM consists of n alternatives and m criteria that form a decision matrix of order $m \times n$. Out of m criteria, some are beneficial and some are non-beneficial depending upon the priority we give the weights to these criteria [8]. There are various techniques to give weights to these criteria one of the techniques is Shannon entropy for weight estimation. In this study, we used two techniques for product recommendation such as COPRAS. We used it to rank and select the best smartphone to evaluate the framework's effectiveness. An extensive and comprehensive approach to solving purchasing dilemmas is presented in the study [9].

1.1.1 Weight estimation by shannon entropy

The Shannon entropy method can be used to assign weights to criteria in MCDM based on the principle of diversity or uncertainty among the criteria. This method is particularly useful when decision-makers want to avoid bias and ensure that the selected criteria reflect a balanced representation of their preferences. We determine the weights of each feature using a decision matrix. We follow Algorithm 1 to estimate the weights of each criterion.

1.1.2 COPRAS

The Complex Proportional Assessment technique (COPRAS) is a multi-criteria decision-making approach designed to facilitate the ranking and selection of alternatives across various criteria. COPRAS addresses the intricacies of real-world decision scenarios by considering both favorable and unfavorable impacts associated with each alternative across a spectrum of criteria. The method involves a sequence of steps encompassing criteria normalization, determination of preference coefficients, and computation of overall rankings. By evaluating net flow values for individual alternatives, which encapsulate the weighted disparities between positive and negative evaluations, COPRAS provides a systematic mechanism to manage complex and interrelated criteria. Its versatility extends to diverse domains, including project assessment, supplier choice, and resource distribution, enhancing the quality and comprehensiveness of decision-making processes. The stepwise procedure is shown in Algorithm 2.

1.2 Contributions

(1) To develop a system that facilitates better product evaluation and purchase decisions while decreasing the time and effort spent by human analysis. A practical application is provided to illustrate how the proposed method is straightforward and applicable to endeavor businesses.
(2) The integrated approach Shannon entropy and COPRAS methodologies of this system's MCDM deliver a trustworthy suggestion to the consumer in the form of a compromise rating. The Shannon entropy weighting approach was integrated with the selected MCDM methods, and the performance of these combinations was demonstrated using the illustration of selecting the best mobile phone.
(3) We also validated the MCDM framework's robustness using sensitivity analysis and correlation with the existing approaches in the field of MCDM, by taking three weights in 3 different cases.

1.3 Organization of study

The article's structure is given as: First, we look at the overall state of product recommendations and then look at how customers rated various products on the web. Section 1 is the introduction of MCDM, the recommendation system, and online services. Earlier studies are discussed in Sect. 2.

Section 3 provides the preliminaries used in this study. Section 4 includes the proposed methodology, which includes five mobile phone brands ranked, and analyses the proposed model's robustness using sensitivity analysis. Section 5 includes the conclusion, and Sect. 6 discusses the upcoming plans.

2 Related work

Various techniques based on MCDM have been used till now to make the recommendation system. Kang and Park conducted a study using sentiment analysis and the VIKOR approach to quantify and rank consumer satisfaction with mobile application services by Kang et al. [10]. Peng et al. examined Chinese reviews of mobile phones and proposed ranking alternatives using fuzzy PROMETHEE by Peng et al. [11]. Preethi et al. [12] used the AHP method to estimate the weight of mobile phones and used the TOPSIS method to make a recommendation. Rotter et al. [13] used ELECTRE-III methodology for content-based image retrieval. All objective criteria are captured using MATLAB's toolbox for fuzzy inference system (FIS) by the author [14]. Abirami et al. [15] used the word count method for weight estimation and ranking; they performed the MCDM-based SAW and TOPSIS methods on the hospital dataset. Liang et al. [16] used the frequency of words for the weighting method and the VIKOR technique for the ranking method and a case study was done on the hotel dataset. Material selection in marine environments can be improved using a TOPSIS-PSI approach proposed by Yadav et al. [17]. Annette et al. [18] propose an MCDM recommender system for ranking cloud renders farm service. Prasad et al. [19] applied fuzzy AHP-TOPSIS for selecting coating magnesium alloy. The author [20] applied MCDM techniques to Big data. The author [21] used TOPSIS and VIKOR to analyze and estimate digital cameras' and cell phones' weights using the webometric standard. The author [22] used the AHP method for weight estimation and TOPSIS, VIKOR, and PROMTHEE-II for mobile phone recommendation. Yang et al. [23] used TF-IDF for weight estimation and IVIF-TOPSIS for ranking purposes of cars. The author [24] proposed methods for optimizing steel towers made from prefabricated cylindrical stacks presented in this paper, which are commonly used in the oil and gas industry. Zhang et al. [25] estimated the weights using AHP, used a ranking method using TODIM, and performed the case study on phones. Vyas et al. [22] used the AHP method and recommended using AHP, FMADM, and VIKOR to conduct the case study on the bank dataset. The author [26] presents the application and results of one aspect of the general BOCR method, which was solved using the WASPAS method to verify the results. The researcher [27] uses genetic algorithms and a decision-making process that considers multiple perspectives to find an optimal heat transfer coefficient. Qin et al. [28] used the entropy method, and ranking was done based on ELECTRE and VIKOR. The author proposes intuitionistic fuzzy sets (IFS) for product recommendation [29]. Bueno et al. established the MCDM methodology, which determines the best appropriate technique for each business case [30]. The author [31] used MADM for Television advertisement recommendations.

2.1 Research gap

Previous research employed numerous fuzzy-MCDM techniques to determine the best product. Fuzzy sets and fuzzy decision-making strategies can effectively deal with the imprecision and fuzziness that frequently accompany decision-making. In recent years, there has been considerable research on the theoretical and applied sides of MCDM and fuzzy MCDM. Fuzzy Logic control methods have one main flaw: they are reliant on human knowledge and experience. This article addresses the limitations observed in prior studies that utilized fuzzy logic like (IF-Multi-MOORA) [29], (IF-TOPSIS) [32], (IF-TODIM) [33], (IF-VIKOR) [34]. In many research, authors utilized a scoring system that relied on proprietary fuzzy membership values, lacking standardization and shaped for convenience through human intervention. In contrast, our methodology derives fuzzy values directly from star ratings associated with mobile features, providing a realistic and empirical foundation for our ranking process, as this data originates from customer feedback. In our approach, we initiate the process by assigning weights to each criterion, representing various mobile features. Subsequently, we apply the COPRAS (Complex Proportional Assessment) technique to rank the alternatives, which in our context pertain to different brands of mobile phones. This systematic and data-driven method ensures the objectivity and credibility of our evaluation process. Therefore, utilizing the proposed recommended model will increase precision while decreasing time, space, and computational complexity.

3 Preliminaries

This section discusses the theory behind MCDM, Shannon Entropy, and COPRAS methods in Sects. 3.1, 3.2 and 3.3.

3.1 MCDM

Multi-Criteria Decision Making (MCDM) is a method that takes into account the strengths and weaknesses of several potential solutions and then selects one that meets

the needs of all stakeholders. An MCDM problem has two different kinds of criteria: cost criteria and benefit criteria. It is preferable to minimize the cost criterion while increasing the benefit criterion as much as possible. A typical MCDM problem with m alternatives (Al_1, Al_2,... Al_m) and n criteria ($(\vartheta)_1$, $(\vartheta)_2$,..., $(\vartheta)_n$) can be presented in the following form.

$M = [x_{ij}]_{m\times n}$, , $W = [w_j]_n$,

Here M stands for the decision matrix, x_{ij} stands for ith alternative's performance concerning the jth criterion, and W stands for the weight vector, w_j denotes the weight of the jth criterion. The initial decision matrix M is typically incomparable because different criteria are expressed using several units of measurement. Numerous MCDM methods and strategies have been developed, proposed, and put into practice since the 1960s. The goal of multi-criteria decision-making (MCDM) is not simply to offer the best alternative, but to help decision-makers choose an alternative that satisfies their needs and is consistent with their preferences. If you want to make good decisions early on, you must acquaint yourself with MCDM methods and the perspectives of the people who will be making them.

3.2 Weight estimation through the Shannon Entropy Method

Shannon's entropy is a well-known method for determining an MCDM problem, particularly when obtaining a good weight based on preferences is difficult and decision-making experiments are not viable. Shannon's entropy technique can be expressed in a few steps. First is the normalization (ω_{ij}) of the decision matrix using Eq. (1).

$$\omega_{ij} = \frac{x_{ij}}{\sum_{j=1}^{m} x_{ij}} \qquad \begin{matrix} j = 1, 2 \ldots \ldots \ldots m. \\ i = 1, 2 \ldots \ldots \ldots n \end{matrix} \tag{1}$$

All measurement irregularities are eliminated by normalizing the raw data. This procedure converts numerous scales and units from various criteria into standard measurable units to compare different criteria. Then Compute entropy (E_i) using Eq. (2).

$$E_i = -E_o \sum_{j=1}^{m} \omega_{ij}. \ln \omega_{ij} \qquad i = 1, \ldots \ldots n, \tag{2}$$

$E_0 = (\ln m)^{-1}$ ω_{ij}. $\ln$ ω_{ij} is defined as 0 if $\omega_{ij} = 0$. Equation 3 denotes the degree of diversification (D_i).

$$D_i = 1 - E_i \tag{3}$$

The degree of importance of attribute i is denoted (I_{ij}) and evaluated by Eq. (4).

$$I_{ij} = \frac{D_i}{\sum_{S=1}^{n} D_S} \approx \frac{1 - E_i}{\sum_{S=1}^{n} (1 - E_s)} \tag{4}$$

The range of entropy values E_i is [0 1]. The larger the E_i is, the greater the differentiation degree of index i is, and more information can be derived. Here normalizing the entropy value to [0, 1] makes it more interpretable. A value of 0 indicates that all criteria are perfectly certain or homogeneous, meaning there is no uncertainty in the decision process. A value of 1, on the other hand, indicates maximum uncertainty among the criteria. In summary, normalizing the entropy value to the [0, 1] range in the Shannon Entropy method simplifies interpretation and comparison, making it a valuable tool for decision-making in various applications.

3.3 COPRAS method

The author [35] presented the COPRAS technique. Several researchers concur with the dependability and reliability of the COPRAS method, and it is now employed to handle a variety of engineering and management multi-attribute issues. In addition, the COPRAS approach can calculate both maximization and minimization criteria. It is now possible to derive both qualitative and quantitative metrics using this approach. The COPRAS approach has many advantages over other multi-criteria decision-making methods. If you compare two or more products, you can see whether one is superior or inferior. There are seven steps in this procedure.

As with any MCDM problem, a decision matrix must be constructed. The COPRAS method begins by constructing the problem's decision matrix. The decision matrix's columns and rows contain the criteria and the alternatives. The decision matrix demonstrates the relative merits of multiple alternatives in light of several criteria. D stands for the decision matrix, which stores the ith alternative's performance value relative to the jth criterion (where m and n stand for the total number of alternatives and criteria, respectively). x_{ij} represents the ith alternative's performance following the jth criteria. Equation 5 is a matrix representation of the decision matrix. C_1 C_2 $\cdots$ C_n

$$D = \begin{matrix} A_1 \\ A_2 \\ \vdots \\ A_m \end{matrix} \begin{bmatrix} x_{11} & x_{12} & \ldots.. & x_{1n} \\ x_{21} & x_{22} & \vdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \tag{5}$$

The mathematical normalization method converts the performance of the various alternatives under consideration into comparable, dimensionless values. In the COPRAS approach, normalization is accomplished using Eq. (6).

$$X_{ij}^{*} = \frac{x_{ij}}{\sum_{i=1}^{m} x_{ij}} \tag{6}$$

Where, x_{ij} is the performance value of the ith alternative concerning the jth criteria, $\dot{X}_{ij}$ is its normalized value, m is the number of alternatives, and n is the number of criteria. Then we create a weighted normalized decision-making matrix (d_{ij}) given by using Eq. (7), where ($\dot{X}_{ij}$) is the performance and (I_{ij}) is the weight, which we had estimated using Shannon entropy.

$$d_{ij} = X_{ij}^{*} . I_{ij} \tag{7}$$

Equation 7 creates a weighted normalized decision-making matrix. All criteria are given weights according to Shannon entropy. After that, each alternative is classified as either index-maximizing or index-minimizing based on Eqs. (8) and (9).

$$S_{i+} = \sum_{j=1}^{k} d_{ij} \qquad j = 1,\ 2, \ldots,\ k \text{ maximizing index.} \tag{8}$$

$$S_{i-} = \sum_{j=k+1}^{n} d_{ij} \qquad j = k+1,\ k+2 \ldots n \text{ minimizing index.} \tag{9}$$

Where k denotes the number of profit-maximizing criteria. The alternative's performance metrics are based on cost-type criteria when all values match up with beneficial or profit-type criteria. Equation 10 determines the relative importance of each alternative (Q_i).

$$Q_i = S_{+i} + \frac{\min_i S_{-i} \sum_{i=1}^{m} S_{-i}}{S_{-i} \sum_{i=1}^{m} \frac{\min_i S_{-i}}{S_{-i}}} \cong S_{+i} + \frac{\sum_{i=1}^{m} S_{-i}}{\sum_{i=1}^{m} \left(\frac{1}{S_{-i}}\right)} \tag{10}$$

The (Q_i) value corresponding to the ith alternative indicates how satisfied the provided requirements are with that alternative. The higher the value (Q_i), the better the alternative's relative performance, thus the higher its position. Because of this, the most reasonable and efficient decision matrix unit should have the maximum (Q_i), i.e., the best value. Alternatives are ranked according to their relative importance based on a weighting formula. Equation 11 determines the alternative with the largest relative weight (rank).

$$A^{\otimes} = \{A_i | \max_i Q_i\} \tag{11}$$

Equation 12 is used for calculating Utility (U_i) values and is given by Eq. (12). They are ranked from the most significant to the least significant to determine the most viable alternatives.

$$U_i = \frac{Q_i}{Q_{\max}} \times 100 \tag{12}$$

4 Methodology

This work presents a framework for product ranking. We use ratings of various features of products and overall customer ratings through the combination of Shannon entropy and the COPRAS model. The framework is divided into three phases:

(i) Extract product features and customer ratings;
(ii) Design a decision matrix with criteria; and.
(iii) Implement MCDM techniques.

During the initial phase of feature extraction, we assess consumer satisfaction with the features of alternative products. Information on product features was collected at the feature level. We transform the obtained features into a decision matrix in the next step. Lastly, feature weights are retrieved using the Shannon entropy method, and COPRAS is used to rank the products. To ensure the framework's dependability and efficacy, a sensitivity analysis is performed at the conclusion. Figure 1 depicts the processing steps of the proposed model. The procedures involved in completing each stage are outlined below.

Step 1: Data Collection.

Data about mobile brand features like battery life, storage capacity, cost-effectiveness, and customer rating are gathered from online sources. This methodology was used to rank and select the best smartphone based on the performance of the developed framework from Amazon's feature rating. Table 1 shows the five smartphones released in 2019.

Once the alternatives are secured, the features of the available alternatives are extracted. We gather brand-specific features for each mobile brand from online sources. The mobile devices were assessed based on eight web-scraped features. Table 2 displays the extracted features and the criteria nomenclature.

Step 2: Determine weights using Shannon Entropy.

In the next step, we determine the weights of each feature using a decision matrix. We follow Algorithm 1 to estimate the weights. The initial decision matrix is given in Table 3. Then we normalize the decision matrix using Eq. (1) and the normalized decision matrix is shown in Table 4.

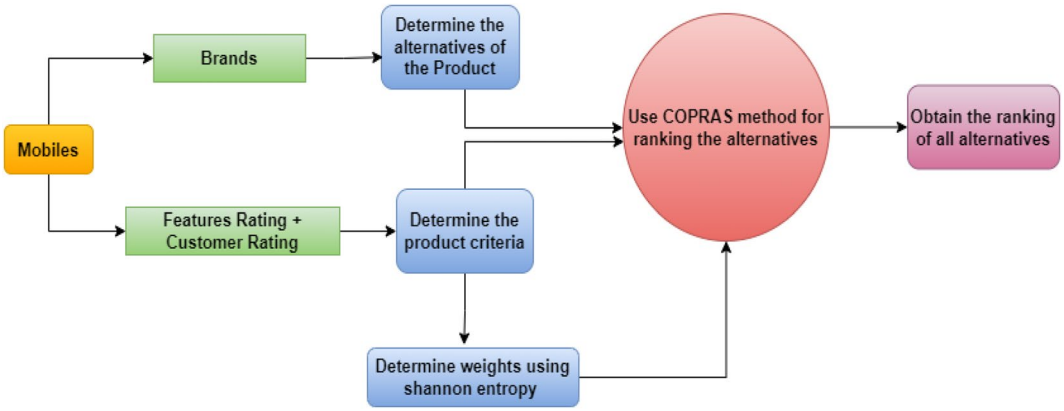


Fig. 1 Proposed model's pipeline

Table 1 Alternatives used to rank the products using the proposed model

Sr. no.	Alternatives	Mobile brand
1	Al_1	*"Samsung Galaxy A50, A505G, 64GB"*
2	Al_2	*"Google - Pixel 3a with 64GB."*
3	Al_3	*"Xiaomi Redmi Note 81 Pro 128GB"*
4	Al_4	*"Moto G7 with Alexa Hands- 64 GB"*
5	Al_5	*"LG G8 ThinQ with Alexa Hands-128 GB"*

Table 2 Criteria used to rank the products using the proposed model

Sr. no	Criteria numbering	Mobile features	B/NB
1	$(\vartheta)_1$	Battery life	Beneficial
2	$(\vartheta)_2$	Value for money	Non Beneficial
3	$(\vartheta)_3$	Face recognition	Beneficial
4	$(\vartheta)_4$	Fingerprint reader	Beneficial
5	$(\vartheta)_5$	Easy to use	Beneficial
6	$(\vartheta)_6$	Screen quality	Beneficial
7	$(\vartheta)_7$	Camera quality	Beneficial
8	$(\vartheta)_8$	Customer reviews	Beneficial

In the next step, we evaluate the entropy using Eq. (2). In the third step, we evaluate the degree of diversification using Eq. (3). In the fourth step, we evaluate the degree of importance using Eq. (4), given in Table 4. Final weights for each criterion using Eqs. (1–4) are given in Table 4.

Step 3: COPRAS Method.

To evaluate mobile recommendation performance, we consider five mobiles from five distinct brands. We utilize eight distinct criteria and five alternatives. Tables 1 and 2 contain the following criteria and alternatives. The decision matrix is constructed as in Table 3, and the weights obtained by the Shannon entropy method (Table 4) will be used as input for the *COPRAS* method following Algorithm 2. In the sequence of applying the COPRAS method. The first step is the normalization of the decision matrix using Eq. (5). In the second step, a weighted normalized decision-making matrix is constructed and evaluated using Eq. (6). The third step is maximizing and minimizing each alternative using Eqs. (7) and (8). In the second last step relative weight of each alternative (Q_i) is calculated using Eq. (9). As a final step, the best mobile phone brand is determined. After evaluating COPRAS, it has been determined that the Al_1 alternative

Algorithm 1: Weight estimation using Shannon Entropy

1: Construct a decision matrix x_{ij}.

2: Normalize the decision matrix. (ω_{ij}), $\omega_{ij} = \dfrac{x_{ij}}{\sum_{j=1}^{m} x_{ij}}$

3: Calculate entropy (E_i). $E_i = -E_o \sum_{j=1}^{m} \omega_{ij} . \ln \omega_{ij}$

4: Calculate the degree of diversification (D_i). $D_i = 1 - E_i$.

5: Calculate the degree of importance (I_{ij}). $I_{ij} = \dfrac{D_i}{\sum_{S=1}^{n} D_S} \approx \dfrac{1-E_i}{\sum_{S=1}^{n}(1-E_s)}$

Table 3 Decision matrix along with criteria and alternatives

Alternatives	Criteria↓							
	$(\vartheta)_1$	$(\vartheta)_2$	$(\vartheta)_3$	$(\vartheta)_4$	$(\vartheta)_5$	$(\vartheta)_6$	$(\vartheta)_7$	$(\vartheta)_8$
Al_1	4.2	4	4.5	3.9	4.1	4	4	4.3
Al_2	4.3	4.8	3.7	4.5	4.4	4.4	4.7	4.4
Al_3	4.7	4.7	4	4.6	4	4.1	4.3	4.6
Al_4	4	4.5	4	4.4	5	4.6	1	4.3
Al_5	4.4	4.6	4.2	4.5	4	4.4	4.8	4.4

Table 4 Degree of diversification ($1 - E_j$) and Degree of importance (I_j)

Criteria→	$(\vartheta)_1$	$(\vartheta)_2$	$((\vartheta)_3$	$(\vartheta)_4$	$(\vartheta)_5$	$(\vartheta)_6$	$(\vartheta)_7$	$(\vartheta)_8$
$1 - E_j$	− 0.15993	− 0.15956	− 0.15946	− 0.15978	− 0.15823	− 0.16003	− 0.09493	− 0.16074
I_j	0.13188	0.13158	0.131497	0.131757	0.130483	0.131962	0.07828	0.132553

performs best because it ranks one. According to the other alternatives, Al_2 performs worse than them because it ranks fifth, so it can be considered that Al_2 performance is the least satisfactory. The ultimate alternative ranking is given in Table 5, calculated using Eq. (10). Lastly, the Utility value index (U_i) using Eq. (11) in Table 5. Equations 4 and 5 calculate the optimal (maximizing/minimizing) index values visible in Table 5.

Equation 9 informs that the best Q_i value is Al_1, shown in Table 5 as the best alternative. Equation 6 is used to determine the relative importance of each alternative.

$Q(Al_1) = 0.20917$, $Q(Al_2) = 0.19779$, $Q(Al_3) = 0.20527$, $Q(Al_4) = 0.20801$, $Q(Al_5) = 0.20232$.

The alternative's rank is displayed in Table 5. The alternatives with ranks one and two are highly recommended for purchase by customers, the alternatives with rank three have

Algorithm 2: Ranking alternatives using the COPRAS method

1: Construct a decision matrix x_{ij}.

2: Normalize the decision matrix (X^*_{ij}), $X^*_{ij} = \frac{x_{ij}}{\sum_{i=1}^{m} x_{ij}}$

3: Calculate Normalized Decision-Making (d_{ij}), $d_{ij} = X^*_{ij} . w_j$

4: Calculate Maximum (S_{i+}) and Minimum Indexes (S_{i-}) for Each Alternative.

$$S_{i+} = \sum_{j=1}^{k} d_{ij} \quad \text{and} \quad S_{i-} = \sum_{j=k+1}^{n} d_{ij}$$

5: Calculate Weighted averages (Q_i), $Q_i = S_{+i} + \frac{\min_i S_{-i} \sum_{i=1}^{m} S_{-i}}{S_{-i} \sum_{i=1}^{m} \frac{\min_i S_{-i}}{S_{-i}}} \cong S_{+i} + \frac{\sum_{i=1}^{m} S_{-i}}{\sum_{i=1}^{m} \left(\frac{1}{S_{-i}}\right)}$

6: Calculate the order of alternatives $(A^{\otimes})$, $A^{\otimes} = \{A_i \mid \max_i Q_i\}$

7: Calculate the Utility value (U_i), $U_i = \frac{Q_i}{Q_{\max}} \times 100\%$

Table 5 Index values: maximizing and minimizing and each alternative's relative weight (Q_i) and ultimate ranking of alternative and utility value index (Ui)

Alternatives↓	S_{i-}	S_{i+}	Q_i	Rank	Recommendation rate	Ui
Al_1	0.11987	0.07794	0.20917	1	Highly recommended	100
Al_2	0.13138	0.07805	0.19779	5	Least recommended	94.5589
Al_3	0.12745	0.08185	0.20527	3	Average recommendation	98.1346
Al_4	0.11953	0.07641	0.20801	2	Highly recommended	99.4449
Al_5	0.12937	0.08072	0.20232	4	Weakly recommended	96.725

Table 6 Weight estimation in different cases

Sr. no.	Weighting methods	Description
Case 1	Entropy method	For this, we evaluate using the entropy method
Case 2	B-50%, NB-50%	50% weightage is given to beneficial criteria and 50% to non-beneficial criteria
Case 3	B-60%, 40%-NB	60% weightage is given to beneficial criteria and 40% to non-beneficial criteria

an average recommendation, the alternatives with rank four are weakly recommended, and the alternatives with rank five are the least recommended to buy or sell because they are the least popular among consumers. The utility (U_i) value index is calculated in Table 5. Microsoft Excel (2019 version) was used for MCDM operations on the entire dataset. A laptop running Windows 10 and powered by AMD RYZEN 3 dual-core CPU (2.6 GHz, 64 bits) and 8GB of RAM executed the programs (64-bit).

4.1 Comparison of the proposed method with other MCDM methods

The comparative analysis of the proposed model with other existing models is shown in this section. For this objective, the proposed work is contrasted with a few MCDM approaches, including (IF-Multi-MOORA) [29], (IF-TOPSIS) [32], (IF-TODIM) [33], (IF-VIKOR) [34]. The advantages of the proposed work over alternative MCDM techniques are discussed in more depth below. A sensitivity analysis of the criteria weights is conducted to comprehend the consequences of changing the weight values in the ranking order. This examination demonstrates the suggested work's robustness. The proposed work's consistency is further evaluated using Spearman correlation.

4.1.1 Sensitivity analysis

A sensitivity analysis (Sen_{ij}) is performed to examine the ranking's stability and consistency concerning the criteria weights. The study aims to improve a model's output in terms of quality and quantity, showing the sensitivity of decision-making when confronted with uncertain input values. Reduce subjectivity by creating a series of filter selections and guiding the selection of complex alternatives with the help of the multi-criteria approach. The same problem can have multiple solutions because different MCDM methods use different mathematical procedures. The researchers' MCDM method determines which alternatives to consider. Despite this, there is no consensus regarding the quality of the decision-making technique known as sensitivity analysis. Table 6 shows three weighted scenarios for testing the proposed model's robustness. The values of weights of the three cases corresponding to 8 criteria are given in Table 7.

Table 7 Criteria with different weights

Criteria	Entropy method (Case 1)	B-50%, NB-50% (Case 2)	B-60%, 40%-NB (Case 3)
$(\vartheta)_1$	0.13188628	0.125	0.15
$(\vartheta)_2$	0.13158042	0.125	0.15
$(\vartheta)_3$	0.13149667	0.125	0.1
$(\vartheta)_4$	0.13175688	0.125	0.1
$(\vartheta)_5$	0.13048336	0.125	0.15
$(\vartheta)_6$	0.13196213	0.125	0.1
$(\vartheta)_7$	0.078281	0.125	0.1
$(\vartheta)_8$	0.13255326	0.125	0.15

Table 8 Performance (Q_i) of Alternatives under different weights

Alternatives	Case 1	Case 2	Case 3
Al_1	0.20917	0.20763	0.20726
Al_2	0.19779	0.19527	0.19832
Al_3	0.20527	0.20328	0.20587
Al_4	0.20801	0.21581	0.21188
Al_5	0.20232	0.19929	0.20221

Table 9 Ranking of alternatives in three different cases

Different cases	Alternatives ranking
Case 1	$Al_1 > Al_4 > Al_3 > Al_5 > Al_2$
Case 2	$Al_2 > Al_5 > Al_3 > Al_1 > Al_4$
Case 3	$Al_2 > Al_5 > Al_3 > Al_1 > Al_4$

The proposed COPRAS model is implemented using these weights in three different cases, and the relative weight index is given in Table 8. Then the final ranking of alternatives was done in all three cases in Table 9.

In this *case 1*, when we used the entropy method for weight estimation, the ranking of alternatives is shown in Table 9. In *case 2*, we gave 50% weightage to beneficial criteria and 50% weightage to non-beneficial criteria; weights are visible in Table 7, and the corresponding ranking is given in Table 9. So, in *case 2*, the rank of Al_3, Al_4, and Al_5 remained the same as in *case 1*, but the rank of Al_1 and Al_2

get exchanged with the variation in the weights same follow in *case 3* as it is in *case 2*. This shows that with changing weights, the rank of the alternative is almost the same in 3 alternatives (Al_3, Al_4, Al_5) out of 5 alternatives (Al_1, Al_2, Al_3, Al_4, Al_5). Only alternatives with consecutive ranks get exchanged, which is quite evident that the ranks of the two best alternatives can get exchanged as we did not see much difference in the ranks of alternatives in different cases of weight estimation. Therefore, this shows the robustness of our method.

Criteria characteristics and weighting distribution among criteria both have an impact on MCDM ranking outcomes [36]. In this section, we will examine the effect of varying the weight of criteria on the ranking results obtained by each method and the robustness of the employed MCDM methods. To accomplish this, the MCDM methods depicted in Fig. 2 demonstrate the variation of criterion weights and observe the changes in the final ranking of alternatives. There are significant changes in the weights of criteria, and this panel shows how those changes affect the rankings of alternatives. Furthermore, the sensitivity coefficient (Sen_{ij}) for the ith method has been calculated using the jth criterion). Equation 13 determines the average change in alternative ranking results when the COPRAS methodology alters the criterion weight.

$$Sen_{ij} = \frac{\sum_{w=1}^{W} D_{ij}^{w}}{H} \forall i, j, w \in \{0.1, 0.2, 0.3, ...0.9, 1\} \tag{13}$$

$$Sen_{ij} = \frac{2}{4} = 0.5$$

Where D_{ij}^{w} is the number of changes in the method i's alternative ranking result when the weight of criterion j changes within a specific range, and H stands for how often the criteria weights change. The weight of each criterion changes between 0.1 and 1. Sen_{ij} It is 0.5, which indicates that variation in the weights has the most negligible impact on the ranking of alternatives. Therefore, COPRAS is effective at strengthening the decision-making process.

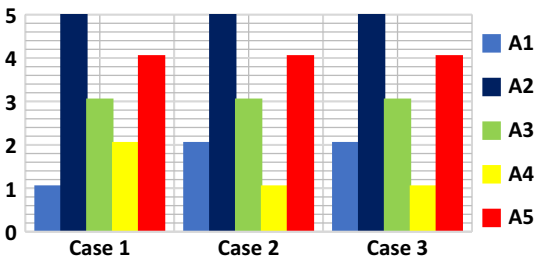


Fig. 2 Alternatives ranking under different weights

Hence, it is clear from Fig. 2 ranking pattern that irrespective of different weights in three cases, the variation comes in the ranking of alternatives. Two changes were made in the consecutive ranks of the alternatives due to the different weights (*case 1*, *case 2*, and *case 3*). *Alternative 1* and *Alternative 2* ranks are affected due to variations in the weight in the three cases. However, only the rank of two consecutive alternative changes, i.e., Al_1 and Al_2, get exchanged, whereas the rank of three alternatives, Al_3, Al_4, *and* Al_5, *is* unaffected by the variation in weights.

4.1.2 Correlation and correlation coefficient

Correlation in Multiple Criteria Decision Making (MCDM) refers to the statistical relationship between different criteria used to assess alternatives. It serves as a critical analytical tool for decision-makers in understanding how changes in the values of one criterion are associated with changes in the values of another. It is the measure of the degree of dependency or interconnection between criteria. It provides insights into whether changes in one criterion are associated with changes in another and whether these relationships are positive, negative, or negligible.

The correlation coefficient (r_{coeff})is a numerical measure used to quantify the degree and direction of correlation between two criteria. It is a value that ranges from −1 to 1. If the correlation coefficient is 1 indicates a perfect positive correlation, meaning that as one criterion increases, the other also increases linearly. Similarly, if the correlation coefficient is − 1 indicates a perfect negative correlation, suggesting that as one criterion increases, the other decreases linearly. A correlation coefficient of 0 signifies no linear relationship between the criteria; they are not correlated.

4.1.2.1 Comparison based on correlation Correlation coefficients have been used extensively for comparing rankings from different decision-making methods. This coefficient value can be calculated using Eq. (14) below [37].

$$r_{coeff} = \frac{\sum_{i=1}^{n} (a_{ij} - a_{ij}^{*})(b_{ij} - b_{ij}^{*})}{\sqrt{\sum_{i=1}^{n} (a_{ij} - a_{ij}^{*}) \times \sum_{i=1}^{n} (b_{ij} - b_{ij}^{*})^2}}, j = 1, 2, ...n \tag{14}$$

This study found that all correlation coefficients greater than 0.6 indicated that the method developed in this paper is compatible with other techniques. However, it is essential to note that our method and the IF-TOPSIS method have a higher correlation coefficient than (0.98).

According to Table 10. the rankings produced by the proposed model are compared to those produced by existing MCDM approaches MCDM methods (IF-Multi-MOORA) [29], (IF-TOPSIS) [32], (IF-TODIM) [33], (IF-VIKOR) [34]

Table 10 Results from different methods and Spearman's correlation coefficients

Methods	Score	Ranking	Ranking correlation				
			COPRAS	IF-Multi MOORA	IF-TOPSIS	IF-TODIM	IF-VIKOR
COPRAS	$Q(Al_1)=0.20917$, $Q(Al_2)=0.19779$, $Q(Al_3)=0.20527$, $Q(Al_4)=0.20801$, $Q(Al_5)=0.20232$	$Al_1>Al_4>Al_3>Al_5>Al_2$	1	0.6	0.98	0.5	0.6
IF-Multi-MOORA	–	$Al_1>Al_4>Al_2>Al_5>Al_3$	–	1	0.6	0.3	0.9
IF-TOPSIS	$CI(Al_1)=0.53$, $CI(Al_4)=0.41$, $CI(Al_3)=0.23$, $CI(Al_5)=0.22$, $CI(Al_2)=0.16$	$Al_1>Al_4>Al_3>Al_5>Al_2$	–	–	1	0.5	0.6
IF-TODIM	$CI(Al_1)=0.53$, $CI(Al_4)=0.41$, $CI(Al_3)=0.23$, $CI(Al_5)=0.22$, $CI(Al_2)=0.16$	$Al_1>Al_3>Al_2>Al_4>Al_5$	–	–	–	1	0.3
IF-VIKOR	$\delta(Al_1)=0.87$, $\delta(Al_4)=0.8$, $\delta(Al_2)=0.63$, $\delta(Al_5)=0.25$, $\delta(Al_3)=0$	$Al_1>Al_4>Al_2>Al_5>Al_3$	–	–	–	–	1

Table 11 Efficiency of the proposed model in different cases

Efficiency	Best case	Average case	Worst case
Time complexity	$\Omega(m+n)$	$\theta(mn)$	$O(mn)$
Space complexity	$\Omega(m+n)$	$\theta(mn)$	$O(mn)$

offered in this field to determine the validity of the proposed model's ranking. Comparative studies were conducted using data previously used in similar studies. Our method shows a positive correlation with all other previously proposed methods, and our result is at par with this method.

5 Discussion

An integrated model is used to identify the best alternatives, and the robustness and validity of the results are evaluated. In this section, we discuss the advantages and disadvantages of the proposed integrated model.

5.1 Time and space complexity

Assessing the algorithm's efficiency and usefulness requires knowing its time and space complexity. The time complexity of an algorithm is based on the number of operations it must perform relative to the size of the input dataset to complete its task. The algorithm's space complexity evaluates how much space it takes up during execution for varied input sizes. In Table 11, we can see how the algorithm's run time and space complexity vary depending on whether we have m options or n criteria.

5.2 Advantages

Many studies claim that the Shannon entropy method is the best weighting approach for predicting weights [38]. It is reliable enough to identify both the magnitude of contrast and the dispersion of responses and appropriately calculate their weights. The Shannon entropy can measure the uncertainty of a random process [39]. Similarly, the COPRAS method evaluates a multi-criteria system of variables to maximize or minimize the values. This technique makes it simple to compare and validate the measurement data. Zheng et al. [40] state that the COPRAS technique (i) Is easy to understand and implement, (ii) Uses several perspectives to decide rank, and (iii) Takes into account the type of criteria when determining the final ranking. Because of this, the writers decided to look into the COPRAS recommendation approach. A sensitivity analysis is done to see

how the ranking order would change if the criteria weights were changed. Figure 2 shows the rank values of several sets of criteria weights. Figure 2 demonstrates the adaptability of the proposed work, and the ultimate ranking order is $Al_1 > Al_4 > Al_3 > Al_5 > Al_2$. Spearman's correlation is used to examine the consistency of the suggested work by comparing the rank values to those of the other MCDM approaches Spearman correlation is evaluated to see the correlation between the existing and the proposed method. Table 10 shows that our proposed MCDM technique positively correlates with the existing MCDM approaches. The highest correlation coefficient recorded is 0.98 with IF-TOPSIS, which is very near 1. The proposed study appears to be reasonably consistent with the other MCDM techniques. With a thorough grasp of each criterion, COPRAS can generate a ranking order that is fairly distinct from other methods. The proposed method is unsupervised, so it does not require training. It can be implemented in any recommendation system to recommend the best alternative to the customer and the seller. The impact of varying the weights of various criteria on the ranking results and the methodology of under-consideration is examined by sensitivity analysis. As a result, the model's credibility is enhanced. Thus, the proposed method can be used in any real-world context.

5.3 Shortcoming

The primary limitation of this study is that MCDM methodologies for ranking alternatives are not always consistent. The results obtained in COPRAS can be sensitive to the weights assigned to the criteria. Small changes in these weights can lead to significantly different outcomes, and determining the most appropriate weights can be challenging. Another shortcoming is that COPRAS can become computationally intensive when dealing with a large number of criteria and alternatives. This can make it impractical for complex decision scenarios with extensive data requirements. Despite these shortcomings, COPRAS can be a valuable tool for certain decision contexts, especially when constraints play a significant role in the decision-making process. However, careful consideration of its limitations and appropriate application is essential to ensure reliable and meaningful results.

6 Conclusion

This study suggests a novel decision-making strategy based on the fusion of many mathematical optimization techniques for producing rational conclusions with minimal human participation and subjectivity. In addition, the weights of both criteria and experts are determined using incomplete information. This research presents the best alternative recommended system based on the MCDM model. Different features of mobile brands are used as criteria in the evaluation process, and different brands are used as alternatives. The Entropy weight estimation method is used to evaluate the weights of each criterion. The COPRAS method combines the performance of each mobile criterion to produce a final performance score for each alternative (mobile brands). It provides us with good recommendations for selecting the best brand. In conclusion, the proposed work could be improved by including more conceptual frameworks, such as using transcendental functions to solve problems in industrial and scientific applications. Research can explore methods to make COPRAS more robust to variations in criteria weights, helping to reduce sensitivity and improve the stability of results. Future work could investigate ways to enhance the scalability and efficiency of COPRAS, making it more suitable for large-scale decision problems. Developing more sophisticated constraint-handling techniques within COPRAS to address complex constraints effectively would be valuable. Exploring data-driven approaches to criteria evaluation could reduce the reliance on subjective judgments and improve the robustness of COPRAS. Future research in these areas can enhance the capabilities of the COPRAS technique, making it a more powerful tool for addressing complex decision-making problems with constraints while mitigating some of its current limitations.

Author contributions NP and GJ contributed equally.

Declarations

Conflict of interest All the authors declare no conflict of interest. Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

References

1. Zha ZJ, Yu J, Tang J et al (2014) Product aspect ranking and its applications. IEEE Trans Knowl Data Eng 26:1211–1224. https://doi.org/10.1109/TKDE.2013.136
2. Liu Y, Bi JW, Fan ZP (2017) Ranking products through online reviews: a method based on sentiment analysis technique and intuitionistic fuzzy set theory. Inf Fusion 36:149–161. https://doi.org/10.1016/J.INFFUS.2016.11.012
3. Punetha N, Jain G (2023) Game theory and MCDM-based unsupervised sentiment analysis of restaurant reviews. Appl Intell 53:1–22. https://doi.org/10.1007/s10489-023-04471-1
4. Chen YS, Chuang HM, Sangaiah AK et al (2019) A study for project risk management using an advanced MCDM-based DEMATEL-ANP approach. J Ambient Intell Humaniz Comput 10:2669–2681. https://doi.org/10.1007/S12652-018-0973-2/FIGURES/7
5. Punetha N, Jain G (2023) Aspect and orientation-based sentiment analysis of customer feedback using mathematical

optimization models. Knowl Inf Syst 65:1–30. https://doi.org/10.1007/s10115-023-01848-z
6. Punetha N, Jain G (2023) Game theory and MCDM-based unsupervised sentiment analysis of restaurant reviews. Appl Intell https://doi.org/10.1007/s10489-023-04471-1
7. Chawra VK, Gupta GP (2021) Optimized coverage-aware trajectory planning for AUVs for efficient data collection in underwater acoustic sensor networks. Evol Intell https://doi.org/10.1007/s12065-021-00667-x
8. Punetha N, Jain G (2023) Sentiment analysis of stock prices and News Headlines using the MCDM Framework. In: 2022 4th international conference on artificial intelligence and speech technology (AIST), pp 1–4. https://doi.org/10.1109/aist55798.2022.10065221
9. Vaz WS (2022) A comparative approach to multiobjective optimization of AISC thin-walled members. Evol Intell 15:655–668. https://doi.org/10.1007/s12065-020-00541-2
10. Kang D, Park Y (2014) Review-based measurement of customer satisfaction in mobile service: sentiment analysis and VIKOR approach. Expert Syst Appl 41:1041–1050. https://doi.org/10.1016/J.ESWA.2013.07.101
11. Peng Y, Kou G, Li J et al (2014) A Fuzzy PROMETHEE approach for mining customer reviews in Chinese. Arab J Sci Eng 39:5245–5252. https://doi.org/10.1007/S13369-014-1033-7
12. Preethi JS, Abirami AM, Askarunisa A et al (2015) Applying MCDM techniques for ranking products based on online customer feedback. Int J Knowl Based Comput Syst https://doi.org/10.21863/ijkbcs/2015.3.2.009
13. Rotter P (2014) Relevance feedback based on n-tuplewise comparison and the ELECTRE methodology and an application in content-based image retrieval. Multimed Tools Appl 72:667–685. https://doi.org/10.1007/S11042-013-1384-1/FIGURES/13
14. Rajak AK, Niraj M, Kumar S (2016) Designing of fuzzy expert heuristic models with cost management toward coordinating AHP, fuzzy TOPSIS and FIS approaches. Sadhana Acad Proc Eng Sci 41:1209–1218. https://doi.org/10.1007/s12046-016-0548-x
15. Abirami AM, Askarunisa A (2017) Sentiment analysis model to emphasize the impact of online reviews in healthcare industry. Online Inf Rev 41:471–486. https://doi.org/10.1108/OIR-08-2015-0289/FULL/HTML
16. Liang R, Wang J (2019) A linguistic intuitionistic cloud decision support model with sentiment analysis for product selection in E-commerce. Int J Fuzzy Syst 21:3. https://doi.org/10.1007/S40815-019-00606-0
17. Yadav S, Pathak VK, Gangwar S (2019) A novel hybrid TOPSIS-PSI approach for material selection in marine applications. Sadhana Acad Proc Eng Sci https://doi.org/10.1007/s12046-018-1020-x
18. Annette JR, Banu A (2019) Ranking cloud render farm services for a multi criteria recommender system. Sadhana Acad Proc Eng Sci https://doi.org/10.1007/s12046-018-0981-0
19. Prasad RV, Rajesh R, Thirumalaikumarasamy D (2020) Selection of coating material for magnesium alloy using Fuzzy AHP-TOPSIS. Sadhana Acad Proc Eng Sci https://doi.org/10.1007/s12046-019-1261-3
20. Muruganantham A, Gandhi GM (2020) Framework for social media analytics based on multi-criteria decision making (MCDM) model. Multimed Tools Appl 79:3913–3927. https://doi.org/10.1007/S11042-019-7470-2/TABLES/4
21. Aldhaban F, Daim T, Harmon R, Basoglu N (2020) Technology adoption in emerging regions: case of the smartphone in Saudi Arabia. Int J Innov Technol Manag https://doi.org/10.1142/S0219877020500030
22. Vyas V, Uma V, Ravi K (2020) Aspect-based approach to measure performance of financial services using voice of customer. J King Saud Univ Comput Inform Sci https://doi.org/10.1016/j.jksuci.2019.12.009
23. Yang L, Li Y, Wang J, Sherratt RS (2020) Sentiment analysis for E-Commerce product reviews in chinese based on sentiment lexicon and deep learning. IEEE Access 8:23522–23530. https://doi.org/10.1109/ACCESS.2020.2969854
24. Cicconi P, Castorani V, Germani M et al (2020) A multi-objective sequential method for manufacturing cost and structural optimization of modular steel towers. Eng Comput 0:475–497. https://doi.org/10.1007/s00366-019-00709-0
25. Zhang D, Li Y, Wu C (2020) An extended TODIM method to rank products with online reviews under intuitionistic fuzzy environment. J Oper Res Soc 71:322–334. https://doi.org/10.1080/01605682.2018.1545519
26. Liang J, Yang J (2021) Application of the AHP method on the optimization with undesirable priorities. Eng Comput 1:1–17. https://doi.org/10.1007/S00366-021-01359-X
27. Khosravi R, Teymourtash · AR, Mohammad · et al (2021) Numerical study and optimization of thermohydraulic characteristics of a graphene-platinum nanofluid in finned annulus using genetic algorithm combined with decision-making technique. Eng Comput 37:2473–2491. https://doi.org/10.1007/s00366-020-01178-6
28. Qin Y, Wang X, Xu Z (2022) Ranking tourist attractions through online reviews: a novel method with intuitionistic and hesitant fuzzy information based on sentiment analysis. Int J Fuzzy Syst 24:755–777. https://doi.org/10.1007/s40815-021-01131-9
29. Heidary Dahooie J, Raafat R, Qorbani AR, Daim T (2021) An intuitionistic fuzzy data-driven product ranking model using sentiment analysis and multi-criteria decision-making. Technol Forecast Soc Change 173:121158. https://doi.org/10.1016/j.techfore.2021.121158
30. Bueno I, Carrasco RA, Ureña R, Herrera-Viedma E (2022) A business context aware decision-making approach for selecting the most appropriate sentiment analysis technique in e-marketing situations. Inf Sci 589:300–320. https://doi.org/10.1016/J.INS.2021.12.080
31. Alipour-Vaezi M, Tavakkoli-Moghaddam R, Mohammadnazari Z (2022) Optimization of a television advertisement scheduling problem by multi-criteria decision making and dispatching rules. Multimed Tools Appl 81:11755–11772. https://doi.org/10.1007/S11042-022-12027-7/TABLES/12
32. Boran FE, Genç S, Kurt M, Akay D (2009) A multi-criteria intuitionistic fuzzy group decision making for supplier selection with TOPSIS method. Expert Syst Appl 36:11363–11368. https://doi.org/10.1016/j.eswa.2009.03.039
33. Wu C, Zhang D (2019) Ranking products with IF-based sentiment word framework and TODIM method. Kybernetes 48:990–1010. https://doi.org/10.1108/K-01-2018-0029
34. Büyüközkan G, Feyzioğlu O, Gocer F (2016) Evaluation of hospital web services using intuitionistic fuzzy AHP and intuitionistic fuzzy VIKOR. IEEE Int Conf Ind Eng Eng Manag 2016(Decem):607–611. https://doi.org/10.1109/IEEM.2016.7797947
35. Beheshti M, Amoozad Mahdiraji H, Zavadskas EK (2016) Strategy portfolio optimisation: a copras g-modm hybrid approach. Transform Bus Econ 15:500–519
36. Maliene V, Dixon-Gough R, Malys N (2018) Dispersion of relative importance values contributes to the ranking uncertainty: sensitivity analysis of multiple criteria decision-making methods. Appl Soft Comput J 67:286–298. https://doi.org/10.1016/j.asoc.2018.03.003
37. Chitsaz N, Banihabib ME (2015) Comparison of different multi criteria decision-making models in prioritizing flood management alternatives. Water Resour Manage 29:2503–2525. https://doi.org/10.1007/s11269-015-0954-6

38. Şahin M (2021) A comprehensive analysis of weighting and multicriteria methods in the context of sustainable energy. Int J Environ Sci Technol 18:1591–1616. https://doi.org/10.1007/S13762-020-02922-7/TABLES/6
39. Célio SRL, Rotela Junior P, Aquila G et al (2021) Toward a robust optimal point selection: a multiple-criteria decision-making process applied to multi-objective optimization using response surface methodology. Eng Comput 37:2735–2761. https://doi.org/10.1007/s00366-020-00973-5
40. Zheng Y, Xu Z, He Y, Liao H (2018) Severity assessment of chronic obstructive pulmonary disease based on hesitant fuzzy linguistic COPRAS method. Appl Soft Comput J 69:60–71. https://doi.org/10.1016/J.ASOC.2018.04.035

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

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

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/355616420

Low Power Pre-charge Voltage Level and Low Swing Logic Based 8T SRAM Cell for High Speed CMOS Circuits

Conference Paper · September 2021

DOI: 10.1109/MIEL52794.2021.9569198

CITATION
1

READS
39

2 authors, including:

Ram Murti Rawat
Delhi Technological University
18 PUBLICATIONS **51** CITATIONS

SEE PROFILE

All content following this page was uploaded by Ram Murti Rawat on 14 December 2022.

The user has requested enhancement of the downloaded file.

Low Power Pre-charge Voltage Level and Low Swing Logic Based 8T SRAM Cell for High Speed CMOS Circuits

R. M. Rawat and V. Kumar

Abstract – This article describes the memory cell proposed design of a low-power 8T SRAM cell. Two PMOS transistors are employed in this proposed approach, one of which is connected to the node voltage Q and the other one is connected to the Qbar. In order to remove the swing voltage at the bit and bit bar lines. When the cell is in working mode, the swing voltage is minimized, which reduces the dynamic power dissipation. Pre-charge voltage level is a technique in which a PMOS transistor acts as a switch and an NMOS transistor acts as a resister, resulting in a series of leakage current reductions when the transistors move from sleep to active and vice versa. The findings of the comparison show that the read, write, and hold mode operation for the proposed 8T SRAM cell is superior to that of the 6T SRAM cell. This is frequently because greater static noise margins are obtained, ensuring good bit cell write ability. All simulation results of read and write on the proposed 8T SRAM cell and current models of 6T SRAM cell have been applied to 130 nm CMOS technology utilizing the Symica DE tool.

I. Introduction

Low-power advanced semiconductor memory cell design is a critical component of many digital systems, including personal and workstation computers, commercial and military satellite equipment, telecommunications, consumer technology, and so on. Because the battery powers high-speed portable devices such as notebook computers, laptops, smartphones, cellular phones, and so on, its demand is steadily increasing. For a good reaction, high-speed transportable devices require primary memory. SRAM is used for this function since it is faster and does not require regular replenishment. The most serious issues with large speed SRAM cells involve dynamic power dissipation and leakage current, which diminish the backup power lifetime of portable electronics. As a result, an SRAM cell design with low static and dynamic power dissipation is required [1].

"For numerous years, there has been a focus on low-power architecture and style innovation on increasing the speed and operation frequency of digital systems. The development of complementary Metal Oxide Semiconductor (CMOS) technology enables integrated circuits to have a higher density and greater performance (ICs). When an IC's density, and thus its power consumption, grows, temperature management becomes more challenging [2]". Every bit in an SRAM is recorded on two cross-coupled inverters comprised of four transistors [3]. The better performance demand, however, limits supply voltage scaling. The scalability of supply voltage alone is insufficient to keep the power density within bounds for power-sensitive applications [4]. Circuit approaches and system level strategies, in addition to supply voltage scaling, are required to achieve low power designs [5]. Because of the forced scaling of the gadgets, sub-threshold leakage rises, as well as other bad impacts, such as better Drain Induced Barrier Lowering (DIBL), threshold voltage (Vth) spiel off, lowered on current to off current ratio, and increased source to vacant resistance [6]. The severity of Vth on the channel length increases as Vth recites. If the channel length varies slightly, the device properties become unpredictable, resulting in large threshold voltage variations. As these gadgets are scaled, oxide thickness scaling and improved, non-uniform doping should be introduced to prevent these small channel lighting tricks [7]. Due to the thin oxide thickness, a large electric field is created, resulting in a large direct tunnelling current [8]. "A lower memory transparency Adaptive Body Bias (ABB) circuitry is expected to increase SRAM constancy and yield by balancing age and process fluctuations [9]. A threshold voltage sensor circuit with an analogue controller is part of the intended ABB circuit for power saving. The Multi-Threshold Complementary Metal Oxide Semiconductor (MTCMOS) technology offers low leakage and superior performance by utilizing high speed, Low Threshold Voltage (LVT) transistors throughout the active region and low leakage, High Threshold Voltage (HVT) transistors throughout the sleep mode, reducing the static power loss of the SRAM circuit [10][11]". The CMOS cell that is being developed consumes low power and has faster read and write rates. The cell is based on the notion of single-ended memory storage. This strategy is based on the lowering of static power. A symmetrical architecture based SRAM cell has also been predicted for greater stability [12]. In a concept [13], a 5T SRAM cell is anticipated to have quick performance, high density, and power efficiency. Two transistors separate the cell storage nodes from the read operations path to maintain the cell's information reliability. This topology improves the data stability. "It describes a power-efficient multiport multiport

R. M. Rawat and V. Kumar are with the Department of Computer Science and Engineering, Delhi Technological University, Delhi, Bawana Rd, Shahbad Daulatpur Village, Rohini, Delhi, 110042, India. E-mail: rammurtirawat@dtu.ac.in, vinod_k@dtu.ac.in

978-1-6654-4528-3/21/$31.00 ©2021 IEEE

SRAM featuring cross-point write word lines, shared write bit lines, and shared write row access transistors. The design has one write, one read and two write, two read multi-port SRAMs for registering file applications in Nano scale CMOS technology" [14]. "The cell has a cross-point write word line configuration, which reduces the risk of writing half select and increases the static noise margin (SNM). The write bit lines (WBLs) and write row access transistors are exchanged with neighbouring bit cells to minimize cell transistor consumption and space. Since the design halves the amount of WBL, there is less leakage and power usage. In contrast, for the reading stack, column based virtual ground direct is used to reduce read power usage. The low power consumption of the SRAM cell is attributed to its chained linked drivers, which are powered by bit lines, and read buffers, which give stack consequences" [15].

During this research, a novel low-power 8T SRAM cell was presented. To reduce leakage currents and static power dissipation throughout mode transition, a pre-charge voltage level approach is used. Two voltage sources are employed at the output layer to reduce the fluctuating voltage. As an outcome, the dynamic energy dissipation throughout switch movement is reduced. The efficiency parameters of the suggested SRAM cell are evaluated and compared to those of the competing SRAM designs.

The following is how the paper is structured: Section II goes over existing SRAM cells, whereas Section III goes through the circuit design and operation of the proposed unique 8T SRAM cell. Section IV covers the full examination of the suggested cell's features and comparisons with existent SRAM cells, and Section V brings the paper to a close.

II. Existing SRAM Cell Designs

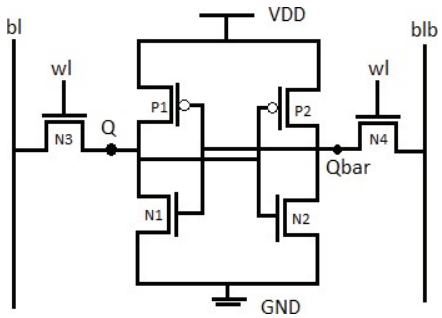


Fig. 1. Conventional 6T SRAM cell [16].

Figure 1 depicts a standard 6T SRAM cell. It has the same six transistors as an SR latch. 4 cross-coupled CMOS inverters (P1, P2 and N1, N2) and two NMOS transistors (N3 and N4) provide accessibility to the cell. The access transistors are switched on when word line (WL) is selected [16]. As an outcome, the supply voltage (Vdd) is generated via the pull-up transistor P1 of the node storing '1'. On the other hand, current passes from the bit bar line's pre-charged state to ground, draining the bit bar lines.

III. Proposed SRAM Cell

A pre-charge voltage level and minimal swing logic based low power 8T SRAM cell are implemented to achieve low static and dynamic power losses for read and write processing, as well as better stability, in order to solve the problems existing with standard SRAM cells and other established SRAM cells.

A. Operation

The presented design can be seen in Figure 2 Two PMOS transistors, M7 and M8, are joined to the cross-coupled node volts Q and Qbar, respectively, in the presented architecture. Throughout the writing and reading operations, PMOS transistors are turned on and off automatically. The pre-charge voltage level reduces the swing voltage throughout read and write operations in the suggested 8T SRAM cell [17]. The dynamic energy dissipation is reduced by lowering the swing voltage. The comparative results show that the suggested 8T cell performs better in read, write, and hold modes than the 6T cell.

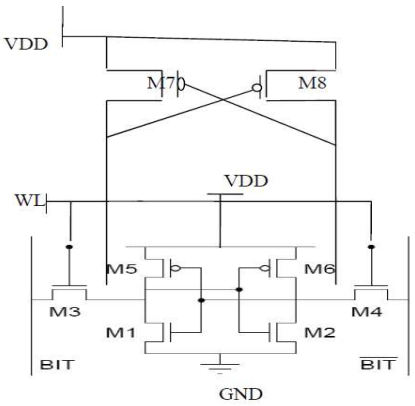


Fig. 2. Proposed 8T SRAM Cell [17].

IV. Results and Discussions

The system is defined by its use of 130 nm technology and a 1.2 volt voltage supply. The simulation is run in the SymicaDE tool. This study examines the suggested 8T SRAM cell in regards to write capability, read stability, and holding static noise margin. This value is then compared to a 6T SRAM cell that is standard.

A. Hold Stability

The most popular method for determining the cell's hold ability is to use the static noise margin (SNM).

Whenever the SRAM cell is in the hold state, hold stability is determined. The word lines are turned off in the holding state, so the cell is completely separated from the bit lines. Before switching the content of the cell, SNM sets the maximum amount of noise that can be placed on the storage nodes. The hold static noise margins of 6T and suggested 8T SRAM cells are shown in Figures 3 and 4, respectively.

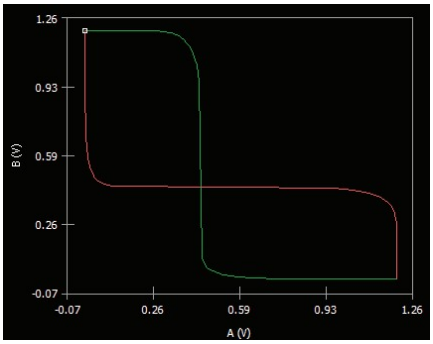


Fig. 3. HSNM 6T SRAM CELL.

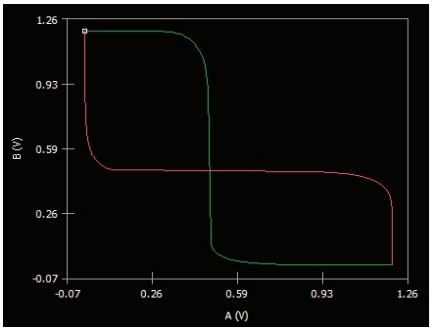


Fig. 4. HSNM PROPOSED 8T SRAM CELL.

B. Read Stability

In an SRAM, read static noise margin (RSNM) is used to determine read stability. Due to the remoteness of the storing nodes, the suggested 8T SRAM cell has a superior RSNM than a normal 6T SRAM cell. The read reliability of 6T and projected 8T SRAM Cells is depicted in Figures 5 and 6, correspondingly.

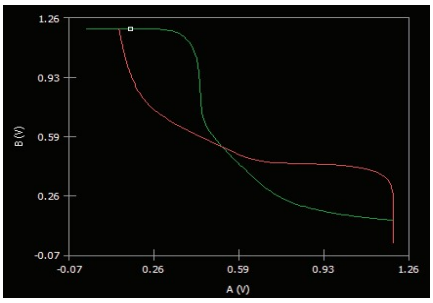


Fig. 5. RSNM 6T SRAM CELL.

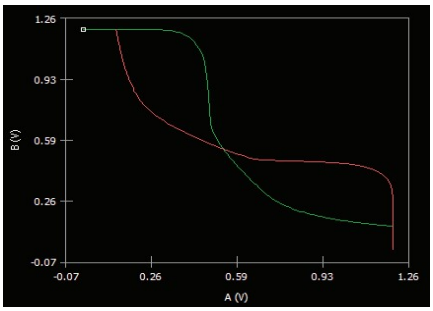


Fig. 6. RSNM PROPOSED 8T SRAM CELL.

TABLE I
COMPARISON OF HOLD SNM AND READ SNM

Cell	Read SNM (in Volts)	Hold SNM (in Volts)
6T SRAM	0.0711	0.125
Proposed 8T SRAM	0.1000	0.125

C. Write Stability

Write static noise margins are used to determine write stability (WSNM). Due to the remoteness of the storage nodes, the suggested 8T SRAM cell has a superior WSNM than a regular 6T SRAM cell. The write reliability of 6T SRAM and projected 8T SRAM cells is depicted in Figures 7 and 8, correspondingly.

The write trip level of a cell is a measurement of its ability to write. It demonstrates how challenging it is for the cell's storage nodes. When Q and Qbar switch their content, the bit line voltage is swept from 0 to Vdd, and the cell switching is registered. The Write trip point is defined as the value of bitline voltages at the intersection of internal storage nodes Q and Qbar.

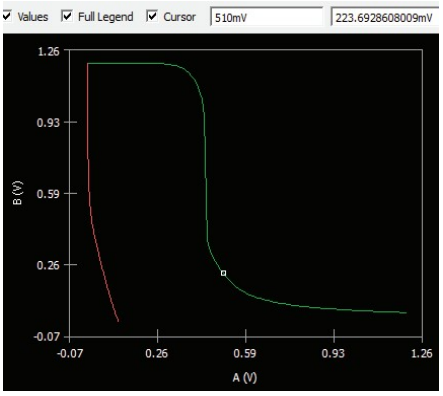


Fig. 7. 6T SRAM Write Trip Point.

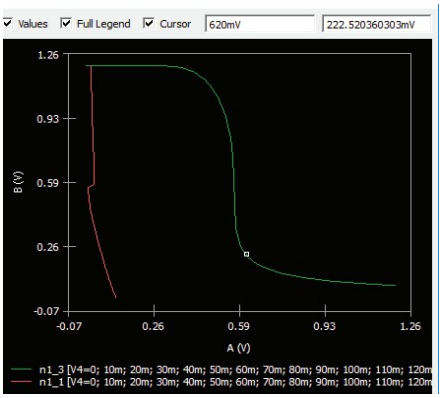


Fig. 9. Proposed 8T SRAM Write Trip Point.

TABLE II
COMPARISON FOR WRITE TRIP POINT OF 6T SRAM AND 8T SRAM CELLS

Cell	Write Trip Point
6T SRAM	675 mv
Proposed 8T SARM	900 mv

V. CONCLUSION

A unique low power 8T Cell and low swing oriented 8T SRAM Cell is employed in this study, which consumes lesser dynamic power throughout write/read operations due to lower swing voltages supplied by the Node voltages Q and Qbar. As a result, the suggested 8T SRAM cell uses reduced static power. Low leakage current and Node voltages offer better stability. According to the simulation, the suggested 8T SRAM cell consumes less dynamic power, has better stability, and has less current leakage throughout mode transitions to standard 6T SRAM cells. Albeit the size increases in compared to 6T SRAM cells, the reduced power consumption and higher stability can overcome the disadvantages. The large increases in results gained by the use of pre-charge voltage levels and low swing logic would be relevant to future limited energy memory designs such as laptop, cellphones, and programmable logic device, among others.

REFERENCES

[1] P. Upadhyay, S. Ghosh, R. Kar, D. Mandal, and S. P. Ghoshal, "Low static and dynamic power MTCMOS based 12T SRAM cell for high speed memory systems", in *Proc. 11th IEEE international joint conference on computer science and software engineering (JCSSE-14)*, 2014, pp. 212–7.

[2] G. Pasandi and S. M. Fekhraie, "A 256-kb near-threshold SRAM with 1 k cells per bit line and enhanced write and read operations", *IEEE Transactions on VLSI Systems*, 2015, vol. 23, no. 11, pp. 1063–8210.

[3] K. Gavaskar and S. Priya, (2013). "Design of efficient low power stable 4-bit memory cell", *International Journal of Computer Applications*, 2013, vol. 84, no. 1, pp. 0975–8887.

[4] K. Gavaskar, U. S. Ragupathy, "An efficient design and comparative analysis of low power memory cell structures", in *Proc. IEEE international conference on green computing, communication and electrical engineering (ICGCCEE'14)*, 2014, pp. 489–493.

[5] S. Borkar, "Design challenges of technology scaling", *IEEE Micro*, 1999, vol. 19, no. 4, p. 23.

[6] J. Brews, *High speed semiconductor devices* , pp. 139–210, 1990, New York: Wiley.

[7] K. Roy and S. C. Prasad, (2000). *Low power CMOS VLSI circuit design*, 2000, pp. 27–29, New York: Wiley.

[8] Y. Taur and T. H. Ning, *Fundamentals of modern VLSI devices*, 1998, pp. 285–286. New York: Cambridge University Press.

[9] H. Mostafa, M. Anis, and M. Elmasry, "Adaptive body bias for reducing the impacts of NBTI and process variations on 6T SRAM cells", *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2011, vol. 58, no. 12, pp. 2859–2871.

[10] J. Kao, A. Chandrakasan, and D. Antoniadis, "Transistor sizing issues and tool for multi threshold cmos technology", in *Proc. IEEE 34th design automation conference (DAC)*, 1997, pp. 409–14.

[11] J. Kao, S. Narenda, and A. Chandrakasan, "MTCMOS hierarchical sizing based on mutual exclusive discharge patterns", in *Proc. IEEE 35th design automation conference (DAC)*, 1998, pp. 495–500.

[12] R. Saeidi, M. Sharifkhani, and K. Hajsadeghi, "A sub-threshold symmetric SRAM cell with high read stability", *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2014, vol. 61, no. 1, pp. 26–30.

[13] R. Gupta and S. S. Gill, "A novel low leakage and high density 5T CMOS SRAM cell in 45 nm technology", in *Proc. Conference on recent advances in engineering and computational sciences*, 2014, pp. 1–6.

[14] D. P. Wang, H. J. Lin, C. T. Chuang, and W. Hwang, "Low-power multiport SRAM with cross-point write word-lines, shared write bit-lines, and shared write row access transistors", *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2014, vol. 61, no. 3, pp. 188–192.

[15] A. Islam and H. Mohd, "Variability aware low leakage reliable SRAM cell design technique", *Journal of Microelectronics Reliability*, 2012, vol. 52, pp. 1247–1252.

[16] T. H. Kim, J. Liu, J. Keane, C. H. Kim, "Circuit techniques for ultra-low power subthreshold SRAMs", In IEEE international symposium on circuits and systems (ISCAS), 2008, pp. 2574–77.

[17] R. M. Rawat, "A Novel Low power and Swing Restoration SRAM Logic Circuit Technique", 2018, *International Journal of Engineering and Technical Research*, vol. 7, no. 02.

Physica Scripta

INCLUSIVE PUBLISHING
TRUSTED SCIENCE

PAPER

Mechanical and surface characterization of $Er_2O_3/La_2O_3/CeO_2$ doped carbide coating developed using high velocity oxy fuel (HVOF)

To cite this article: Mohit Vishnoi *et al* 2023 *Phys. Scr.* **98** 115618

View the article online for updates and enhancements.

You may also like

- Effect of nano-Al_2O_3 addition on the microstructure and erosion wear of HVOF sprayed NiCrSiB coatings
Ayyappan Susila Praveen and Arun Arjunan
- Structure and optical properties of Cd-substituted ZnO ($Zn_{1-x}Cd_xO$) nanostructures synthesized by the high-pressure solution route
Manoranjan Ghosh and A K Raychaudhuri
- Influence of laser texturing pre-treatment on HVOF-sprayed WC-10Co-4Cr+GNP coatings on AISI 304
Vijay Kumar, Rajeev Verma, Kundan Chauhan et al.

This content was downloaded from IP address 14.139.251.98 on 03/11/2023 at 05:40

Physica Scripta

PAPER

Mechanical and surface characterization of $Er_2O_3/La_2O_3/CeO_2$ doped carbide coating developed using high velocity oxy fuel (HVOF)

Mohit Vishnoi[1,2,*], **Qasim Murtaza**[1,*] and **Paras Kumar**[1]

[1] Department of Mechanical Engineering, Delhi Technological University, Delhi 110042, India
[2] Department of Mechanical Engineering, JSS Academy of Technical Education, Noida 201301, India
[*] Authors to whom any correspondence should be addressed.

E-mail: vishnoi.mohit06@gmail.com, qasimmurtaza@dce.ac.in and paraskumar@dtu.ac.in

RECEIVED
29 July 2023

REVISED
25 August 2023

ACCEPTED FOR PUBLICATION
13 October 2023

PUBLISHED
25 October 2023

Keywords: high velocity oxy fuel, rare earth, carbide coating, contact angle, energy dispersive spectroscopy

Abstract

This study is mainly focused on effect of rare earth oxides on a composite coating applied to martensitic steel (SS410) using the high velocity oxy fuel (HVOF) process. The composite coating consists of WC-10Co-4Cr doped with rare earth (erbium oxide/lanthanum oxide/cerium oxide; 0.2 wt%. each). The range of deposited coating thickness was 282–300 μm. Mechanical testing, including tensile, flexural, and hardness tests, along with surface characterization, such as examining morphology, porosity, elemental composition, and phase identification with wettability testing has been conducted on both the substrate and coated samples. The hardness of the coatings is increased by approximately four times to the substrate. The coated sample has porosity in the range of $\geqslant$1 to $\leqslant$ 2%. The contact angle of the coated surface has enhanced ($\approx$134°) than the substrate ($\approx$61°), this indicates that the coated surface exhibits hydrophobicity.

1. Introduction

Rare earth elements, such as cerium (Ce), lanthanum (La), niobium (Nb), and zirconium (Zr), have found widespread application in various industries, particularly in the metallurgical field. These elements are utilized to enhance the properties of materials used in electronics, ceramics, lasers, machinery, and other related areas [1–3]. In addition, rare earth elements have been employed in surface engineering techniques, including chemical treatment, flame spraying, electric plating, laser alloying, and the development of antibacterial coatings for bone [4–7]. Due to its huge atomic radius and chemical activities, rare earth (RE) may react with many elements, including H, O, N, S, and Si, Moreover, they play a vital role in the production of steels [2]. They refine grains and stop grain development during heating as well as provide purifying properties that can get rid of undesirable contaminants. As a result, it was anticipated that rare earth-doped coating would have slower grain development [4]. Surface roughness and the degree of brittle fracture can both be efficiently reduced by rare earth [8]. Erbium oxide (Er_2O_3), a rare earth oxide, stands out due to its remarkable attributes, including a high dielectric constant, elevated bulk refractive index, and wide band gap. The thin films of Er_2O_3 have potential applications as up-conversion layers in optoelectronic devices [9]. Moreover, rare earth elements have proven to be effective in altering the surface characteristics of coatings applied using diverse techniques such as thermal spray, thermal spray welding (TSW), and laser-alloyed coatings [10]. Thermal spray methods such as flame, detonation guns and plasma sprayed are frequently used in power plants and other industrial fields to protect the surface against corrosion and wear [11]. The high-velocity oxygen fuel (HVOF) technique is a well-established method in thermal spraying. It is known for its high gas and particle velocities, which are typically three to four times the speed of sound (Mach 3 to 4). The combustion of fuels with oxygen produces both thermal and kinetic energy that is used to heat and accelerate feedstock powder in HVOF process. The Amorphous steel coatings, known for their exceptional resistance to wear, corrosion, and oxidation, are commonly produced using thermal spraying processes such as HVOF, Moreover, enhanced properties like hardness, density, and strength

© 2023 IOP Publishing Ltd

[12]. Due to the exceptional glass forming capacity of alloys such as yttrium doped amorphous coating has been manufactured using the HVOF process [13].

The high-velocity oxygen fuel (HVOF) process has been widely used to produce WC-Co coatings as protective coatings on engineering components due to its good bond strength with the substrate, high hardness, wear, and corrosion resistance [10, 14]. The above three key factors that cause mechanical failures of components are wear, corrosion, and their combination. In the oil, chemical, coal mining, and electric power industries, equipment such as pump casings, vanes, valves, bent tubes, T-tubes, and rollers are commonly utilized in flowing slurries and is consequently susceptible to erosion–corrosion, which is the primary cause of failure [15]. The extensive research is being conducted to modify powder coating materials in order to reduce production costs and enhance the service life of WC-Co coatings. This includes exploring the use of alternative materials such as ultrafine powders, nano-structured powders, and additives, including rare earth (RE) elements [16–19]. Carbide-based coatings are one of the coatings deposited using a thermal spray that is most frequently. The components are frequently employed at temperatures below 550 °C to provide protection against various types of wear, including sliding, abrasion, fretting, and erosive wear [20]. Incorporating rare earth oxides into plasma-produced coatings yields several effects, including the refinement and purification of the coating's microstructure and the facilitation of new phase formation. Erbium oxide (Er_2O_3) coatings were successfully generated using plasma arc source assisted deposition. Additionally, yttrium oxide (Y_2O_3) and erbium oxide (Er_2O_3) coatings were created through radio frequency (RF) sputtering. These coatings exhibited almost stoichiometric compositions and demonstrated resistance levels surpassing the requirements for fusion reactors (102–104 ohm-meter) [21]. A flame-sprayed nickel-based alloy coating was produced, and two variations were manufactured - one with the inclusion of CeO2 (cerium oxide) and another without it. To study the impact of rare earth elements on the composition and properties of the resulting coating on the counter face metal surface, several techniques were utilized. These techniques include x-ray diffraction (XRD), electron probe microanalysis (EPMA), and x-ray photoelectron spectroscopy (XPS). Through these methods, the influence of rare earth elements on the film's characteristics and composition was investigated. Friction and wear characteristics of the coating were evaluated using a block-on-ring friction and wear tester. Subsequently, the impact of CeO_2 on wear and its underlying mechanism were proposed [22]. Various researchers have suggested the advantages, disadvantages, and application of various thermal sprayed coatings. The papers have also reviewed and suggested the various powders and substrate materials used for coatings, along with their applications [23–26].

The insertion of rare earth (RE) elements into materials has been found to impact various properties, including: (1) Improvement in their mechanical properties such as hardness, strength, and elastic modulus in the material. (2) Reduction in the porosity and cracks in the material, leading to a more refined microstructure. (3) Enhances the tribological properties of coatings. This includes improved adhesion of the coatings to the substrate and enhanced resistance to wear, corrosion, and erosion. (4) Improves the oxide morphology and increases the resilience to thermal shock in thermal shock cycles. It has lowered the rate of chromium (Cr) consumption, promoted the creation of spherical agglomerates, and hindered the formation of spinel-like structures on the surface [27, 28].

A matrix along with WC, for example; WC-Co, has earlier been used by researchers to elevate wear and cavitation resistance [29–31]. Moreover, the inclusion of Chromium (Cr) in the matrix, like; WC-Co-Cr, WC-Cr-C-Ni, and Cr_3C_2-Ni-Cr can add up to further corrosion resistance and reduce cavitation damage [32]. It has been earlier observed that WC-Co-Cr coating set down under high combustion pressure (0.8 MPa) has elevated CE resistance [33]. This experimentation elucidates the mechanical and surface characterization of WC-Co-Cr Coating powder in correlation with rare earth oxide on SS410 substrate. This experimentation was designed after an extensive literature review.

The incorporation of rare earth elements, such as erbium oxide (Er_2O_3), through doping in carbide coatings is a relatively less explored area of research. Even mixing of all mentioned rare earth is also a less explored area of research.

2. Materials and methods

As shown in figure 1, the SS410 commonly known as Martensitic stainless steel (received) has been considered as a substrate for deposition of rare earth doped tungsten carbide coating using the high velocity oxy fuel (HVOF) process. The chemical composition of the substrate is depicted in table 1.

The substrate (SS410) has been blasted using the pressure blasting machine before deploying a coating to the surfaces. The surface blasting was done using brown alumina with 20 mesh particle sizes. The surface blasting parameters were considered such as air pressure (5 bars), blasting distance (4.5'), and blasting angle. The

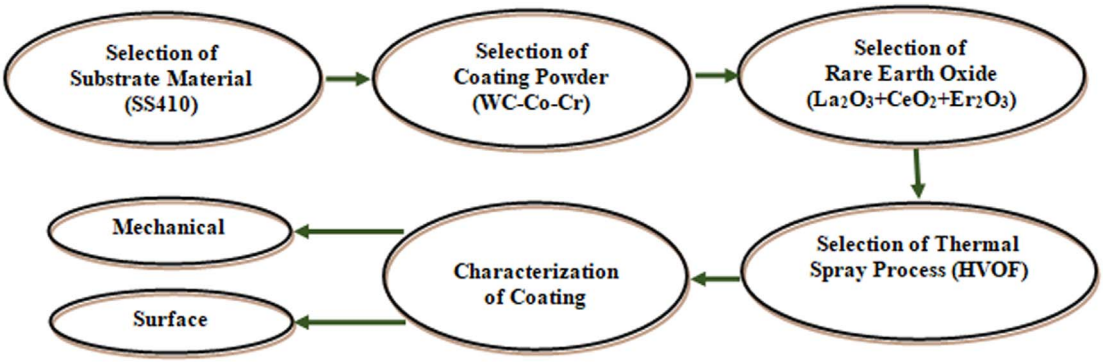


Figure 1. Selection of substrate, coating powder, process, and characterization of developed coating.

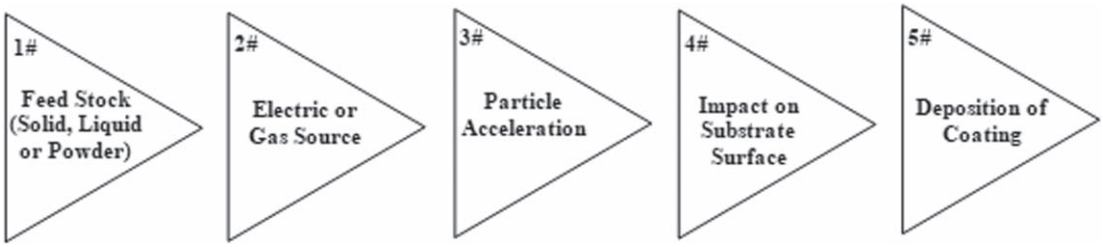


Figure 2. Basic steps involved in manufacturing coating on SS410.

Table 1. Presenting the composition of martensitic stainless steel (SS410).

Element	C	Si	Mn	P	S	Cr	Ni	Fe
Wt.%	0.14	0.52	0.64	0.03	0.004	12.42	0.18	Rest

developed rough texture on the surface could potentially offer advantages in terms of facilitating a strong mechanical bond between the coating and the substrate [34].

To create the coating, a WC-10Co-4Cr composition of agglomerated sintered carbide powder was employed. The particle size of the powder ranged from −45 μm to +15 μm. The spherical rare earth oxides with a purity of 99% were utilized as doping material. These rare earth oxides had particle sizes ranging from 5 μm to 10 μm. Specifically, a mixture of CeO_2, La_2O_3, and Er_2O_3 at a concentration of 0.2 wt% each was appropriately blended with the carbide powder to ensure proper dispersion.

The carbide powder was doped with 0.6 wt% of rare earth elements. Previous research suggests that the optimal range for rare earth doping is typically between 0.4 wt% and 1.2 wt%. Within this range, researchers have observed significant improvements in various coating properties. However, it should be noted that when the doping percentage exceeds 1.2 wt% of rare earth, there is a decline in coating properties, as indicated by studies [3, 11, 14, 15].

Figure 2 shows the basic steps that have been involved in the deposition of rare earth doped coating on SS410. There were five basic steps; starting from powder to deposition of that powder in molten form by utilizing the excess heat and velocity on the prescribed substrate i.e. martensitic steel.

The schematic view of high velocity oxy fuel (HVOF) has shown in figure 3. This thermal spraying process was used to develop a coating on Martensitic stainless steel. At MEC Pvt. Ltd in Jodhpur, the chosen combination of powders was applied onto a substrate of Martensitic stainless steel using a commercial high velocity oxy fuel spray system called the HIPOJET-2700 (model name). The coating thickness ranged between 282–300 μm, as illustrated in figure 8. The development of the coating involved adhering to a set of standard operational parameters, which are mentioned in table 2.

2.1. Surface analysis of deposited coatings

The scanning electron microscope (SEM) model JSM-LA was utilized to examine the morphology of the deposited coating. Additionally, an Energy Dispersive x-ray (EDX) detector integrated with the SEM was

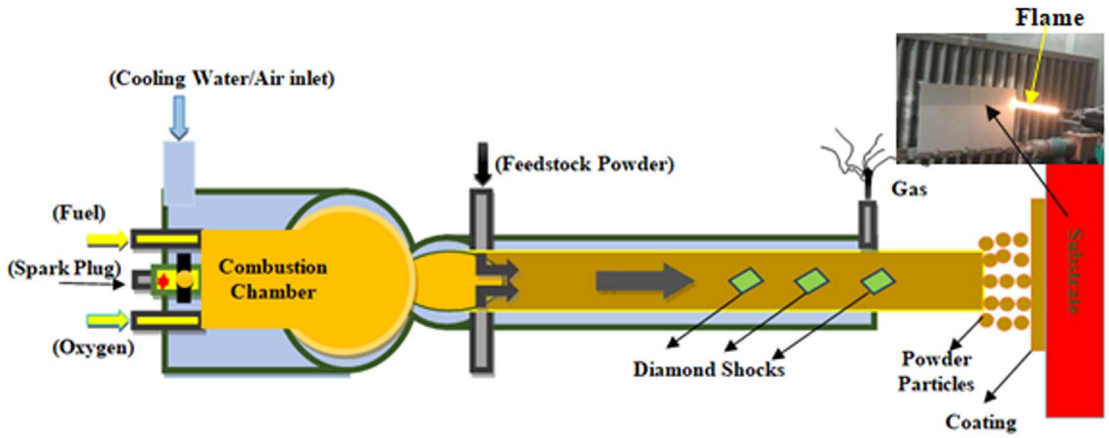


Figure 3. Schematic representation of the HVOF process [35].

Table 2. Standard process parameters of HVOF coating.

Gun		HIPOJET-2700
Oxygen	Pressure (kg/cm^2)	10
	Flow (Slpm)	250
LPG	Pressure (kg/cm^2)	7.0
	Flow (Slpm)	55
AIR	Pressure (kg/cm^2)	6
	Flow (Slpm)	550
Powder Feed Rate (kg/min)		40
Spray Distance (cm)		17.78
Particle Velocity (m/sec)		446
Flame Temperature (°C)		2322
Angle of Deposition (Degree)		90

employed for compositional analysis of the coating. To measure the static water contact angle on both coated and uncoated surfaces, a contact angle meter (DMe-211 plus) was utilized. The porosity level of the coated surface was evaluated using a metallurgical microscope (Axiovert A1, manufactured by Carl Zeiss) with a magnification range spanning from 100X to 1000X.

2.2. Mechanical analysis

The mechanical properties of both the coated and uncoated specimens were assessed using a Universal Testing Machine (Instron-1195). Tensile and flexural (three-point bend) tests were conducted on this machine, which had a maximum capacity of 100kN. The crosshead speed during the tests ranged from 0.005 to 40 mm min^{-1}, and the maximum crosshead displacement achievable was 700 mm.

To determine the hardness of both the coated and uncoated samples, the Micro Vickers hardness test was performed using a Vickers hardness tester (model Matrix DVS 1AT8). The instrument had a maximum hardness capability of 1600HV, and a maximum load of 1 Kg was applied during the hardness test.

3. Results and discussion

The result obtained from each test on coated and uncoated samples has been discussed in this section.

3.1. Surface morphology and elemental identification on deposited coating

The operational parameters of energy dispersive x-ray analysis (EDAX) have been given in table 3. The individual elements (like C, Cr, Co, La, Ce, Er, and W) have been identified by different colors as shown in figures 4(a) to (g) and collectively all the identified elements have illustrated in figure 4(h). The used elements in powder such as Tungsten, Carbon, Cobalt, Chromium, Lanthanum, Cerium, and Erbium in the coating have been verified using EDS as well. The rare earth elements are uniformly distributed in a matrix of WC-Co-Cr as shown in figures 4(d) to 4(f).

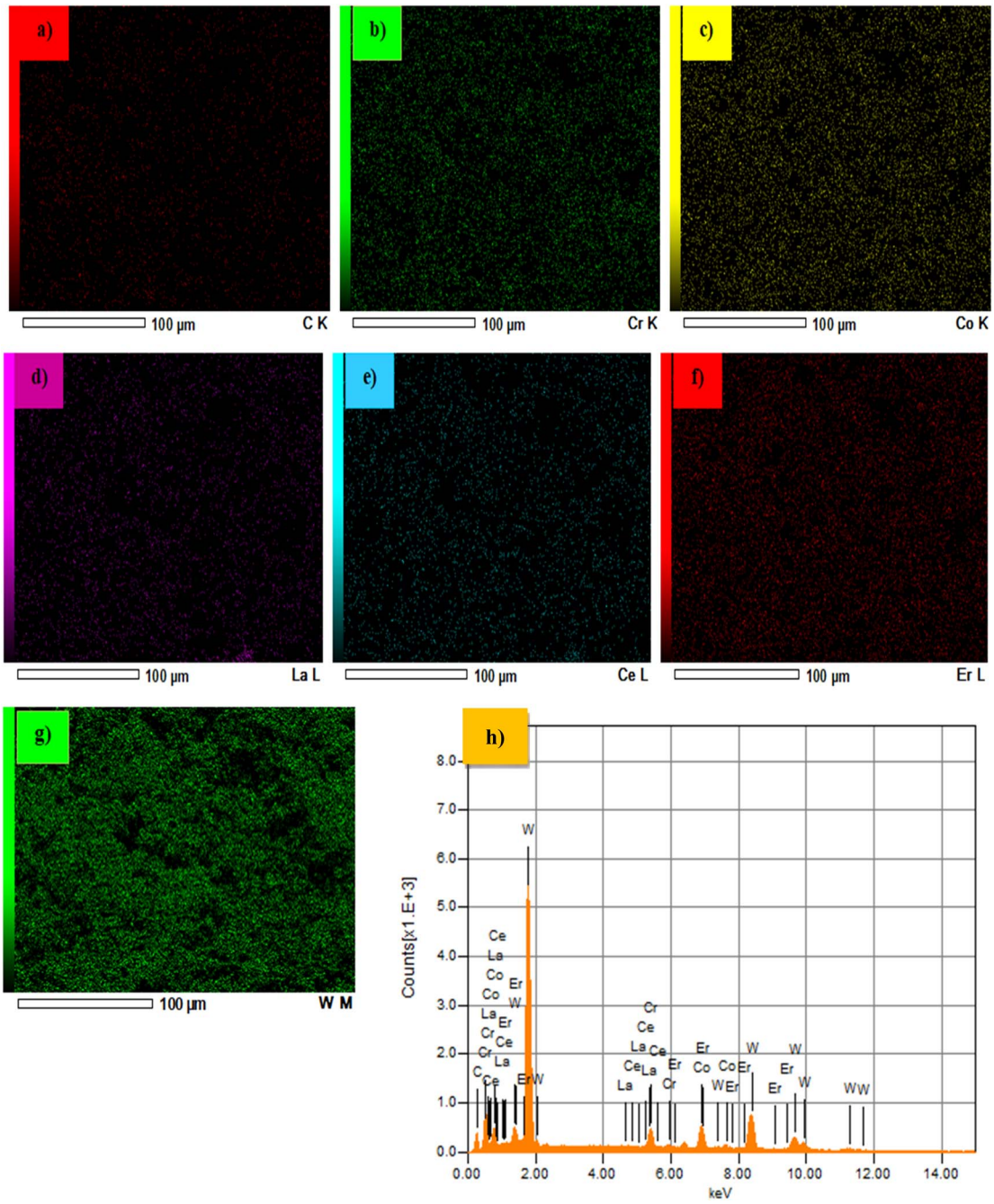


Figure 4. Energy dispersive x-ray analysis (a) Carbon (C), (b) Chromium (Cr), (c) Cobalt (Co), (d) Lanthanum (La), (e) Cerium (Ce), (f). Erbium (Er), (g) Tungsten (W), and (h) Peaks of elements.

Table 3. Operational parameters.

Acquisition condition of EDS	
Volt	20.00 kV
Live Time	196.60 s.
Real Time	202.56 s.
Dead Time	2.00%
Count Rate	1774.00 CPS

The Er_2O_3 is an allotropic phase oxide that generally exists in three forms such as (i) Cubic (C), (ii) Monoclinic (B), and (iii) Hexagonal (H) at elevated temperatures. Under atmospheric pressure, the most stable phase of erbium oxide (Er_2O_3) at elevated temperatures is the cubic phase, known as C-Er_2O_3. The monoclinic

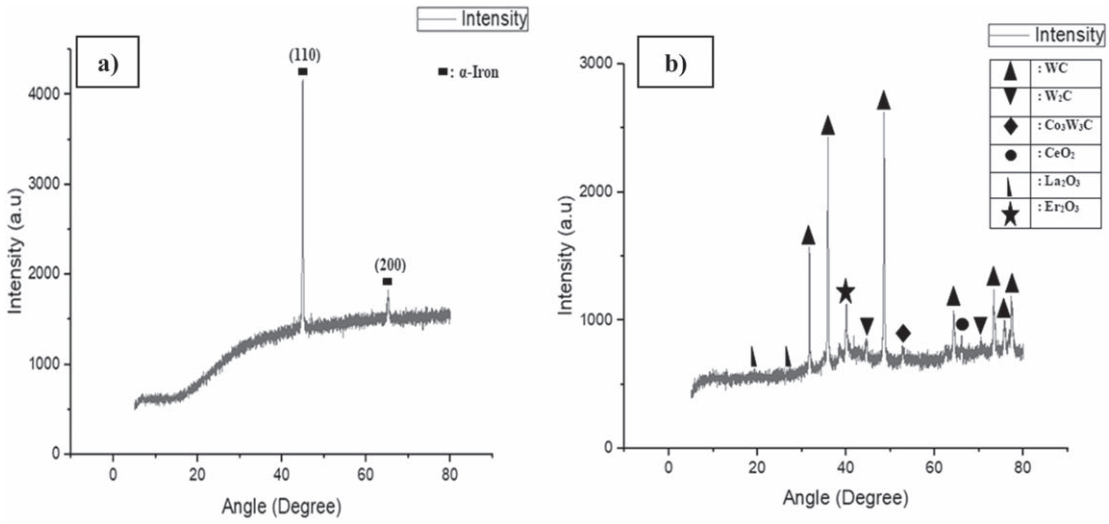


Figure 5. X-ray diffraction peaks of (a) Bared and; (b) RE doped coating.

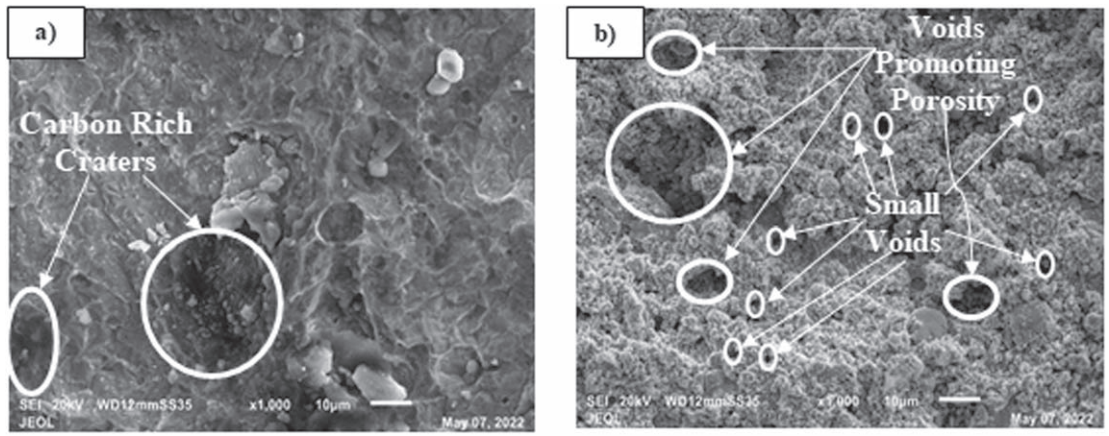


Figure 6. Micrographs of (a) Substrate and; (b) Coated sample at 1000X.

phase (B-Er_2O_3) is metastable and the hexagonal (H-Er_2O_3) phases can be seen at higher temperatures and disappear when cooled to room temperature [36, 37].

Figures 5(a) and (b) display the x-ray diffraction (XRD) plots for the substrate and coated samples, respectively. In figure 5(a), the XRD analysis reveals the identification of alpha phases of iron at different crystallographic planes. On the other hand, figure 5(b) demonstrates the identification of various rare earth oxides, including different cerium and lanthanum oxides, erbium oxide, as well as compounds of Co_3W_3C and WC resulting from doping. It is important to note that the XRD analysis performed in this study aligns with the findings reported by other researchers in their respective works [38–42].

The high-temperature decarburization process leads to the conversion of tungsten carbide into tungsten semi-carbide (W_2C) as described by reactions (a) and (b) [43]

$$2WC \rightarrow W_2C + C \quad \text{(a)}$$

$$W_2C \rightarrow 2W + C \quad \text{(b)}$$

Figure 6 depicts the x-ray graph of the coating, which reveals the presence of two distinct phases: tungsten carbide (WC) and tungsten semi-carbide (W_2C). The formation of the decarburized brittle W_2C phase occurs during the deposition process.

Figures 6(a) and (b) show micrographs of the substrate and coated sample, respectively, captured at a magnification of 1000X using a scanning electron microscope. The substrate has many carbon rich craters due to the presence of carbon in steel was abundant in quantity. The coated sample shows many small voids and these

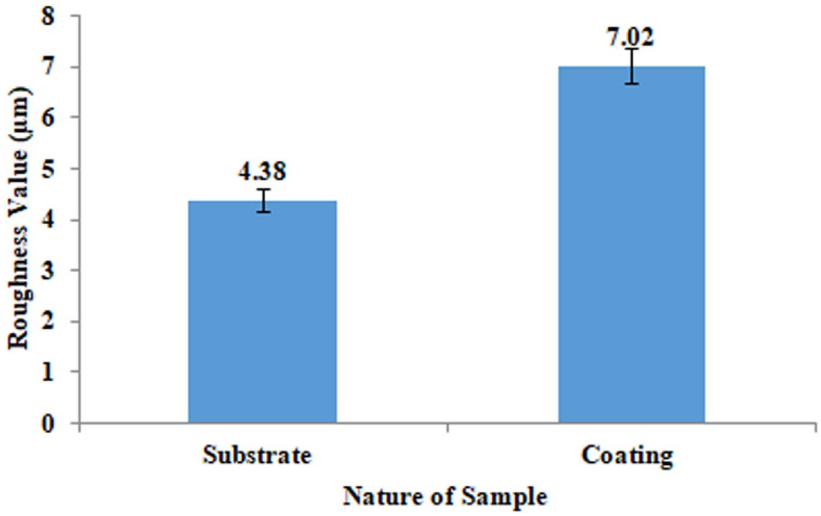


Figure 7. Roughness value of bare and coated samples.

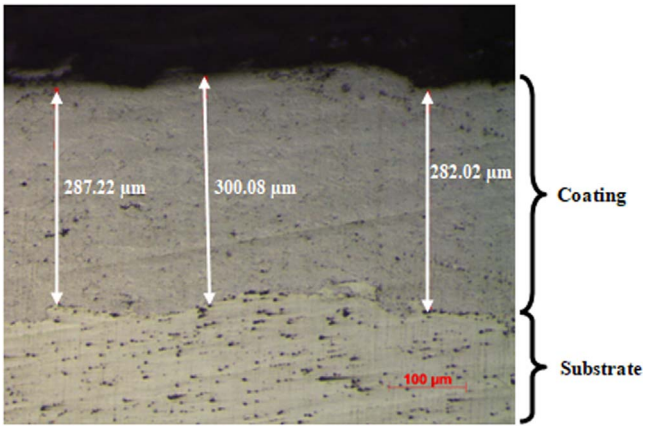


Figure 8. Microscopic image at 200X to measure the porosity level and thickness of the coating deposited.

voids promote the porosity as shown in figure 6(b). All elements were melted and formed a coating over the substrate as shown in figure 6(b).

The roughness of the substrate and the coated surface was measured using a roughness tester (Mitutoyo Surtest, model: SJ301), the roughness of coated sample has a higher roughness value as compared to the substrate. The figure 7 presents the average roughness value with a standard deviation of approximately ±5% for both samples. The coated surface exhibited high roughness due to the presence of significant voids and porosity on its surface as shown in figure 6(b).

In order to gain further insights into the coated surface, porosity and thickness measurements were conducted on the coating deposited via the HVOF process. This analysis, depicted in figure 8, followed the standard ASTM E 2L09-01 (reapproved 2007), Method A [44].

The 24 mm polished area was considered to measure the porosity level using microscopic images at 200X magnification and observed as in the range of $\geqslant 1$ to $\leqslant 2\%$. The presence of the porosity has already been revealed using SEM micrographs as shown in figure 6(b).

Hang *et al* (2018) developed a coating using HVOF and considered doping of rare earth (CeO_2 −1 wt%) in WC-12Co. They claimed the porosity level is 2.1% for nano WC-12Co and 1.6% for micro WC-12Co [14]. Sharma *et al* (2008 and 2009) conducted a study on coating deposited using a flame spray process. They were doped the NiCrBSi powder by rare earth-like ((CeO_2) with varying percentages as 0%, 0.4%, 0.8%, 1.2%, 1.6%, and 2% and (La_2O_3)− 0%, 0.4%, 0.8%, 1.2%, 1.6%, and 2%) respectively and claimed that the porosity level in the range of 5 to 8% [11]. The current work showed a less porosity level which was better than most of the available investigations.

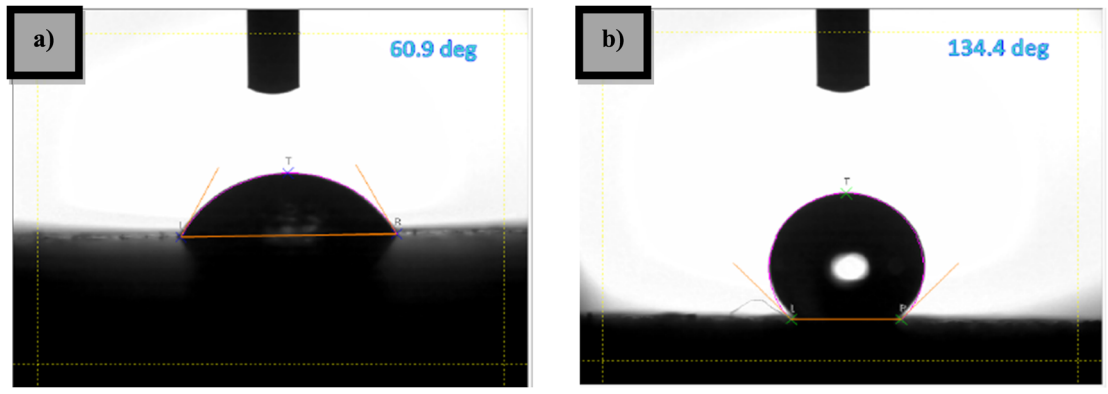


Figure 9. Contact angle of (a) Bared and; (b) Coated samples.

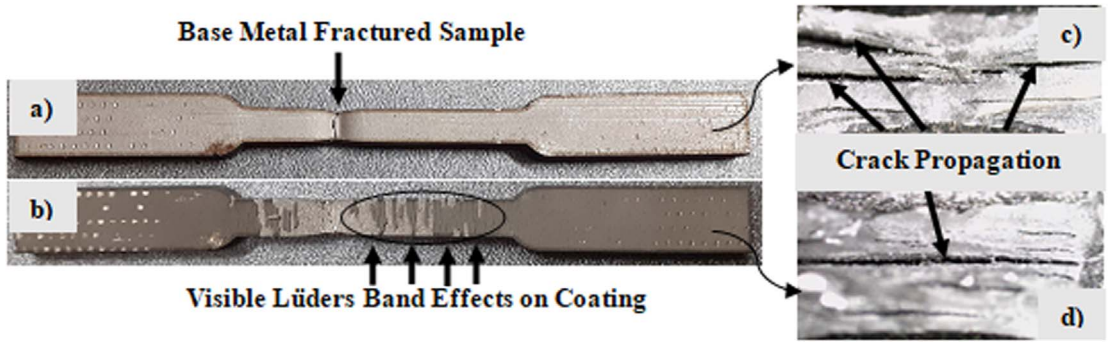


Figure 10. Fractured samples images of tensile test, (a) Substrate; (b) Coated and, Fractography of, (c) Substrate; (d) Coated samples.

The sessile drop method was employed to analyze the wettability behavior of both the bared and coated surfaces. Surface roughness is a critical parameter in assessing the wetting characteristics of a surface. The Wenzel and Cassie Baxter state models were used to explain the impact of surface roughness on wettability [45, 46]. As surface roughness increases, the surface tends to become more hydrophobic due to air trapping in the rough areas [47, 48].

When a liquid droplet is placed on a solid surface, it can either spread or remain confined to a specific area, depending on the interactions between the solid surface and the liquid droplet. The static water contact angle (WCA) is commonly used to measure the wettability behavior of a surface. A water contact angle less than 90° indicates a hydrophilic surface, while an angle greater than 90° suggests a hydrophobic surface [49–51]. Moreover, if the WCA of a surface is greater than 150°, it exhibits super-hydrophobic behavior. This is similar to the behavior observed in natural surfaces, such as lotus and colrabi leaves [52, 53].

The coated sample showed a higher contact angle (range: 130.2–134.4 degrees) than the substrate (range: 60.9–62.6 degrees) as shown in figures 9(a) and (b) because the porosity offers the surface roughness which turns the surface to be hydrophobic.

3.2. Mechanical behavior of bared and coated samples

The specimens depicted in figures 10(a) and (b) were fabricated in accordance with the ASTM-E8 standard. To ensure accuracy, three separate experiments were carried out, and the average values were determined, accounting for a deviation of approximately ±5%.

Figures 10(a) and (b) show the fractured substrate and coated samples. The presence of Lüders bands on the coating was observed in figure 10(b). The damaged coating exhibited multiple bands as a result of the Lüders band effect. It is worth noting that the substrate, which is inherently brittle, demonstrated enhanced ductility after the deposition of the coating. This enhanced ductility supported the formation of Lüders bands.

The substrate has a large number and opening cracks as shown in figure 10(c) as compared with the coated sample as shown in figure 10(d).

Figures 11(a) and (b), the stress–strain diagram for the substrate and the coated sample was obtained from a universal tensile testing machine.

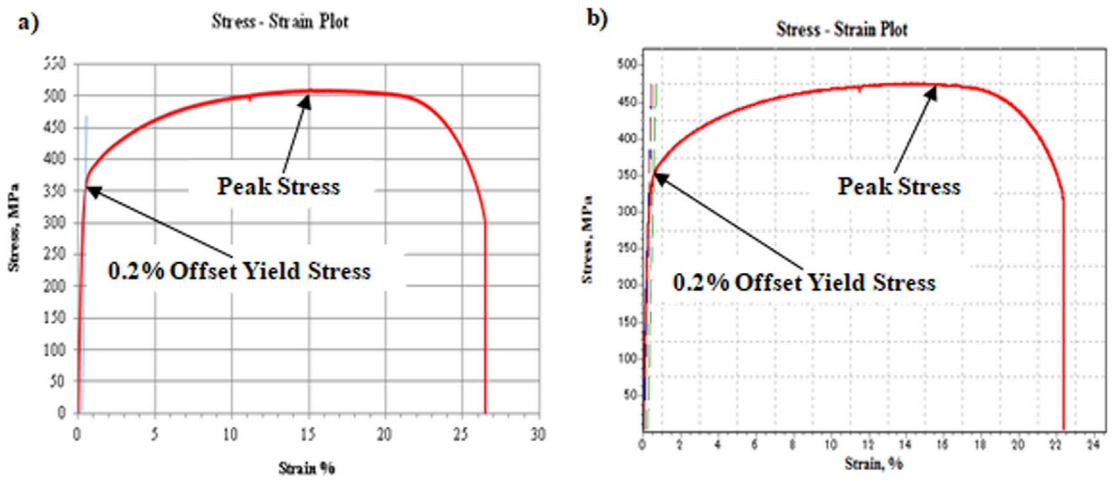


Figure 11. Stress–strain curve from the tensile test, (a) Bared and; (b) Coated sample.

Table 4. Tensile test results.

Tensile test properties	Substrate	Coated sample	Percentage change
Peak Stress (MPa)	508.72	475.32	6.57 (Decrease)
Peak Load (kN)	12.82	13.12	2.33 (Increase)
0.2% Offset Yield Stress (MPa)	355.00	344.58	2.94 (Decrease)
Yield Strain (%)	0.48	0.53	9.77(Increase)
Yield Load (kN)	8.95	9.51	6.28 (Increase)
0.02% Offset Yield Stress (MPa)	250.00	264.65	5.86 (Increase)
Modulus (GPa)	116.00	109.66	5.47 (Decrease)
Elongation at Break (Using Strain) (%)	26.47	22.37	15.48 (Decrease)

Table 5. Flexural (three-point bend) test results.

Flexural test properties	Substrate	Coated sample	Percentage change
Flexural Peak Stress (MPa)	781.80	908.73	16.24 (Increase)
Peak Load (kN)	1.79	2.37	32.89 (Increase)
Offset Flexural Yield Stress (MPa)	0.72	0.95	32.82 (Increase)
Offset Flexural Yield Load (kN)	0.72	0.95	32.82 (Increase)
Flexural Stress at Break (MPa)	921.68	860.72	6.61 (Decrease)

Table 4 showed that the various mechanical properties were obtained from the stress–strain curve. The ceramic coating offers a more brittle behavior than the substrate. This implies a decrease in the yield and tensile strength value of the coated samples as shown in table 4. Initially, the 0.02% offset yield stress value for the coated sample is more than the substrate but at 0.2%, the offset yield stress value for the substrate is more than the coated sample. The modulus value is decreased for the coated sample than for the substrate.

In accordance with the ASTM-D790 standard, the specimen was prepared, and a total of three experiments were carried out. The average values, along with a deviation of approximately ±5%, were taken into account and are presented in table 5.

Figures 12(a) and (b) showed the stress–strain curve of the substrate and coated samples after the bend test.

Table 5 presents the compiled values from the stress–strain diagram. It is observed that the flexural peak stress value of the coated sample exceeds that of the substrate. This finding indicates that the coated sample exhibits a greater ability to withstand bending forces prior to yielding, highlighting its enhanced resistance to bending.

The Vickers (HV0.3) hardness test was conducted to determine the hardness values of both the substrate and coated samples followed by the ASTM-A270 standard. The test involved applying a 300 g load with a dwell time of 10 seconds. Figure 13 represents the average values along with a deviation of approximately ±5% for the three readings recorded during the test. The coated samples exhibited a higher hardness value (approximately HV798)

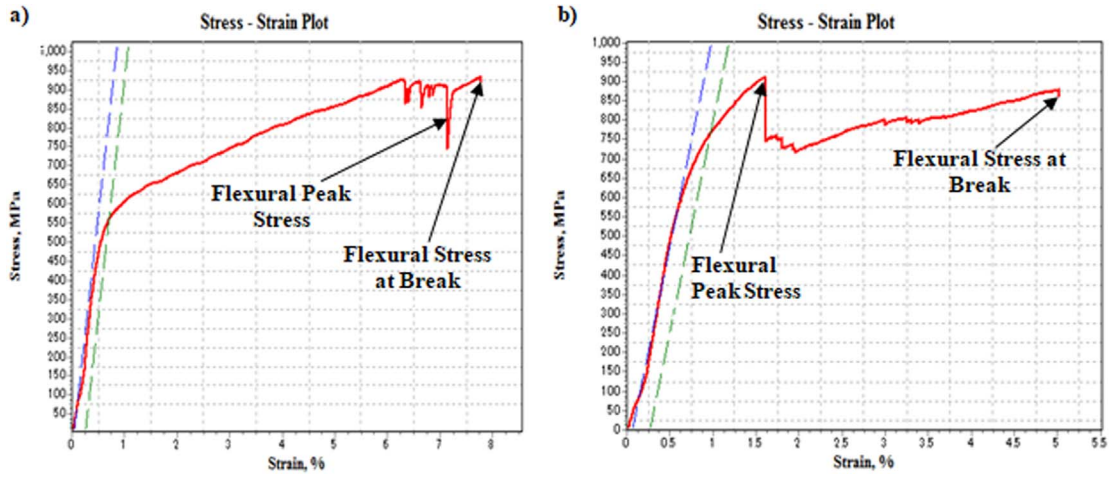


Figure 12. Stress–strain curve from the flexural test (3-point bend test), (a) Bared and; (b) Coated sample.

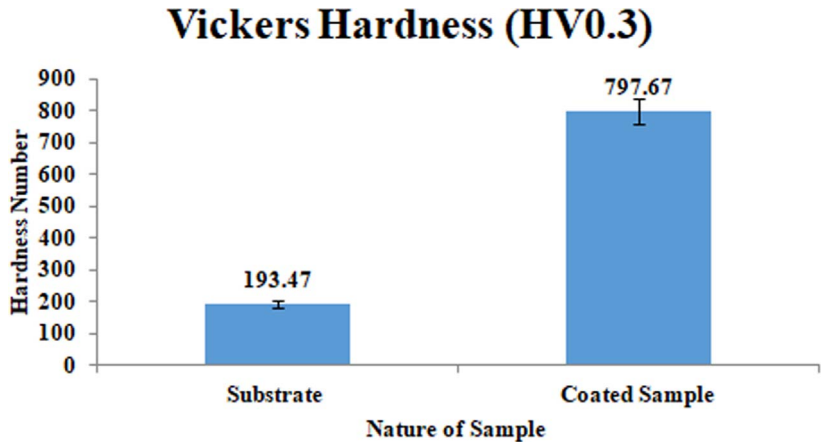


Figure 13. Hardness of the bare and coated sample.

in comparison to the substrate (≈HV193). The hardness value (%) exhibited a significant increase of approximately 312 compared to the substrate. Rare earth-doped carbide coating has an extremely high hardness value as compared to the substrate. The tungsten carbide particles are present in the coating, which enhanced the hardness value. The distributions of tungsten particles are uniform in the coating as shown in figure 4(g).

4. Conclusions

This study aimed to deposit a rare earth doped tungsten carbide molten powder on substrate (grade SS410) using the high velocity oxy fuel (HVOF) process. The following conclusions have been made:

1. The hardness of the coated samples (HV0.3≈797) was considerably higher than the substrate (HV0.3≈193).
2. There were considerable increases in the flexural strength of the coated surface compared to the substrate.
3. The contact angle of the coated samples was notably higher, measuring 134.4°, in comparison to the substrate material's contact angle of 60.9°. This disparity suggests that the newly applied coating has resulted in a surface with hydrophobic properties.
4. The energy dispersive spectroscopy was utilized to identify the presence of multiple elements, including Tungsten (W), Cobalt (Co), Chromium (Cr), Erbium (Er), and others, within the coating. Additionally, the XRD pattern analysis indicated that the high-temperature conditions caused the decarburization of tungsten carbide, resulting in its transformation into tungsten semi-carbide (W_2C).

5. The number of voids was identified on the coated surface using a scanning electron microscope. The porosity level was identified on the coated surface in the range of $\geqslant 1$ to $\leqslant 2\%$.
6. The surface roughness of the coated sample ($\approx 7.02\ \mu$m) was significantly higher than the substrate ($\approx 4.38\ \mu$m).

Acknowledgments

The author expresses gratitude to the Precision Manufacturing Lab at DTU, Delhi, and the MEC Pvt. Ltd Jodhpur for their assistance in fabricating the coating. Additionally, the author would like to thank the IIT, Kanpur for their support in characterizing (mechanical and surface) the samples.

Data availability statement

The data cannot be made publicly available upon publication because no suitable repository exists for hosting data in this field of study. The data that support the findings of this study are available upon reasonable request from the authors.

Future scope

The tribological analysis (such as slurry erosion, air jet erosion, pin-on disc, scratch test, etc) can also be explored for the same mixture composition with some optimization techniques.

Funding information

No funding has been involved in this work.

Author contribution

Mohit Vishnoi: Conceptualization, Investigation, Data Curation, Writing - Original Draft.
Qasim Murtaza: Review & Editing, Supervision.
Paras Kumar: Review & Editing, Supervision.

Conflict of interest

The authors confirm that they do not have any known financial interests or personal relationships that could have influenced the objectivity of the research presented in this paper.

ORCID iDs

Mohit Vishnoi https://orcid.org/0000-0002-8455-6334
Qasim Murtaza https://orcid.org/0000-0001-8784-2304

References

[1] Sheng D 1995 Situation and prospect of application of rare earth in cast iron *Chin. Rare Earths.* **16** 36
[2] Wang K L, Zhu Y M, Zhang Q B and Sun M L 1997 Effect of rare earth cerium on the microstructure and corrosion resistance of laser cladded nickel-base alloy coatings *J. Mater. Process. Technol.* **63** 563–7
[3] Liu Y, Gou G, Wang X, Jia Q, Chen H and Tu M 2014 Effects of rare earth elements on the microstructure and mechanical properties of HVOF-sprayed WC-Co coatings *J. Therm. Spray Technol.* **23** 1225–31
[4] Zhou H X 2008 Preparation and characterization of rare-earth modified thermal sprayed WC/12Co coatings *Master Thesis* Harbin Institute of Technology
[5] Zhang Z, Lu X, Han B and Luo J 2007 Rare earth effect on the microstructure and wear resistance of Ni-based coatings *Materials Science and Engineering:* A **454** 194–202
[6] Wang Y, Liu J J and Yu Z H 1993 Effect of rare earth elements on microstructure and wear resistance of laser remelted iron alloy coatings containing metalloids *Surf. Eng.* 151–5
[7] Wang Y, Su M, Dai X M and Chen C N 1994 Influence of CeO 2 on Corrosion Resistance of B Laser-alloyed layer *Laser Tech.* **18** 9–11
[8] Wang Y, Su M, Dai X M and Chen C N 1994 The influence of corrosion resistance of laser-alloyed b alloy coatings *Laser Technozl.* **18** 6–16

[9] Kasuya A and Suezawa M 1997 Resonant excitation of visible photoluminescence from an erbium-oxide overlayer on Si *Appl. Phys. Lett.* **71** 2728–30
[10] Wang Y, Kovacevic R and Liu J 1998 Mechanism of surface modification of CeO2 in laser remelted alloy spray coatings *Wear* **221** 47–53
[11] Sharma S P, Dwivedi D K and Jain P K 2008 Effect of CeO2 addition on the microstructure, hardness, and abrasive wear behaviour of flame-sprayed Ni-based coatings *Proc. Inst. Mech. Eng. Part J J. Eng. Tribol.* **222** 925–33
[12] Branagan D J, Swank W D, Haggard D C and Fincke J 2001 Wear-resistant amorphous and nanocomposite steel coatings *Metallurgical and Materials Transactions* A **32** 2615–21
[13] Ni H S, Liu X H, Chang X C, Hou W L, Liu W and Wang J Q 2009 High performance amorphous steel coating prepared by HVOF thermal spraying *J. Alloys Compd.* **467** 163–7
[14] Hang Z Q, Xi N Y, Liu Y, Liu Y and Chen H 2018 High-temperature oxidation behavior of HVOF-sprayed rare earth-modified WC–12Co coating *Rare Met.* 1–8
[15] Liu Y, Hang Z, Chen H, Ceng S, Gou G, Wang X, Tu M and Wu X 2016 Erosion–corrosion property of CeO 2-modified HVOF WC-Co coating *J. Therm. Spray Technol.* **25** 815–22
[16] Wang W, Lu Z C, Chen Z H, Zeng M Q, Wang H and Zhu M 2016 Properties of WC–8Co hardmetals with plate-like WC grains prepared by plasma-assisted milling *Rare Met.* **35** 763–70
[17] Jing Q F and Tan Y F 2013 Tribological properties of cobalt-based alloy coating with different cobalt contents by electro-spark deposition *Rare Met.* **32** 40–6
[18] Guo C Y, He X B, Ren S B and Qu X H 2016 Thermal properties of diamond/Al composites by pressure infiltration: comparison between methods of coating Ti onto diamond surfaces and adding Si into Al matrix *Rare Met.* **35** 249–55
[19] Kim J H, Yang H S, Baik K H, Seong B G, Lee C H and Hwang S Y 2006 Development and properties of nanostructured thermal spray coatings *Curr. Appl Phys.* **6** 1002–6
[20] de Villiers Lovelock H L, Kinds J and Young P M 1998 Characterisation of WC—12Co thermal spray powders and HPHVOF wear resistant coatings *Powder Metall.* **41** 292–9
[21] Sawada A, Suzuki A, Maier H, Koch F, Terai T and Muroga T 2005 Fabrication of yttrium oxide and erbium oxide coatings by PVD methods *Fusion Eng. Des.* **75** 737–40
[22] Lian Y, Yu L and Xue Q 1995 The effect of cerium dioxide on the friction and wear properties of flame praying nickel-based alloy coating *Wear* **181** 436–41
[23] Vishnoi M, Kumar P and Murtaza Q 2020 A review on coating for hydro-turbine application by HVOF process *Advances in Materials Engineering and Manufacturing Processes* Select Proceedings of ICFTMM 2019 79–103
[24] Bhati P, Shrama V, Jha R, Gupta S, Vishnoi M and Mamatha T G 2021 The Coatings Developed by thermal spraying technique for hydroturbine applications—a brief review *Advances in Fluid and Thermal Engineering* Select Proceedings of FLAME 2020 559–74
[25] Tripathi A, Pandey A, Rathore A, Singh A and Vishnoi M 2021 Study on tribological behavior of HVOF developed coatings especially for hydroturbine runner application—a concise review *Advances in Fluid and Thermal Engineering* Select Proceedings of FLAME 2020 575–95
[26] Pradeep D G, Venkatesh C V and Nithin H S 2022 Review on tribological and mechanical behavior in HVOF thermal-sprayed composite coatings *Journal of Bio-and Tribo-Corrosion.* **8** 30
[27] Ding X, Cheng X D, Shi J, Li C, Yuan C Q and Ding Z X 2018 Influence of WC size and HVOF process on erosion wear performance of WC-10Co4Cr coatings *Int. J. Adv. Manuf. Technol.* **96** 1615–24
[28] Vishnoi M, Murtaza Q and Kumar P 2021 Effect of rare earth elements on coatings developed by thermal spraying processes (TSP)–a brief review *Mater. Today Proc.* **44** 4053–8
[29] Schwetzke R and Kreye H 1999 Microstructure and properties of tungsten carbide coatings sprayed with various high-velocity oxygen fuel spray systems *J. Therm. Spray Technol.* **8** 433–9
[30] Ding Z, Zhang Y and Zhao H 2007 Resistance of HVOF nanostructured WC-12Co coatings to cavitation erosion *Thermal Spray 2007 Global Coating Solutions.* **1** 633–7
[31] Hou G, Zhao X, Zhou H, Lu J, An Y, Chen J and Yang J 2014 Cavitation erosion of several oxy-fuel sprayed coatings tested in deionized water and artificial seawater *Wear* **311** 81–92
[32] Korobov Y, Alwan H, Soboleva N, Makarov A, Lezhnin N, Shumyakov V, Antonov M and Deviatiarov M 2022 Cavitation resistance of WC-10Co4Cr and WC-20CrC-7Ni HVAF coatings *J. Therm. Spray Technol.* **31** 234–46
[33] Kanno A, Takagi K and Arai M 2020 Influence of chemical composition, grain size, and spray condition on cavitation erosion resistance of high-velocity oxygen fuel thermal-sprayed WC cermet coatings *Surf. Coat. Technol.* **394** 125881
[34] Wang Y Y, Li C J and Ohmori A 2005 Influence of substrate roughness on the bonding mechanisms of high velocity oxy-fuel sprayed coatings *Thin Solid Films* **485** 141–7
[35] Vishnoi M, Murtaza Q and Kumar P 2023 Characterization of erbium oxide doped HP-HVOF deposited carbide ceramic coating on martensitic steel *Proc. of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science* 09544062231164299
[36] Tang M, Lu P, Valdez J A and Sickafus K E 2006 Ion-irradiation-induced phase transformation in rare earth sesquioxides (Dy 2 O 3, Er 2 O 3, Lu 2 O 3) *J. Appl. Phys.* **99** 063514
[37] Gy G Y and Imanaka N 1998 The binary rare earth oxides *Chem. Rev.* **98** 1479–514
[38] Li X, Wu P, Qiu H, Chen S and Song B 2012 Crystallization behavior and mechanical properties of erbium oxide coatings fabricated by pulsed magnetron sputtering *Thin Solid Films* **520** 2316–20
[39] Wang K L, Zhang Q B, Sun M L, Wei X G and Zhu Y M 2001 Rare earth elements modification of laser-clad nickel-based alloy coatings *Appl. Surf. Sci.* **174** 191–200
[40] Wang Y and Northwood D O 2007 An investigation of the electrochemical properties of PVD TiN-coated SS410 in simulated PEM fuel cell environments *Int. J. Hydrogen Energy* **32** 895–902
[41] Jeyaprakash N, Yang C H, Ramkumar K R and Sui G Z 2020 Comparison of microstructure, mechanical and wear behaviour of laser cladded stainless steel 410 substrate using stainless steel 420 and Colmonoy 5 particles *J. Iron. Steel Res. Int.* **27** 1446–55
[42] Zhang Y, Yu G, He X, Ning W and Zheng C 2012 Numerical and experimental investigation of multilayer SS410 thin wall built by laser direct metal deposition *J. Mater. Process. Technol.* **212** 106–12
[43] Koutsomichalis A, Vardavoulias M and Vaxevanidis N 2017 HVOF sprayed WC-CoCr coatings on aluminum: tensile and tribological properties *IOP Conf. Ser.: Mater. Sci. Eng.* **174** 012062 IOP Publishing
[44] E04 Committee 2014 ASTM E2109-01. Standard test methods for determining area percentage porosity in thermal sprayed coatings https://astm.org/e2109-01r21.html
[45] Wenzel R N 1936 Resistance of solid surfaces to wetting by water *Industrial & Engineering Chemistry.* **28** 988–94
[46] Cassie A B and Baxter S 1944 Wettability of porous surfaces *Trans. Faraday Soc.* **40** 546–51

[47] De Foggi C C, Machado A L, Zamperini C A, Fernandes D, Wady A F and Vergani C E 2016 Effect of surface roughness on the hydrophobicity of a denture-base acrylic resin and C andida albicans colonization *Journal of Investigative and Clinical Dentistry.* **7** 141–8
[48] Kittu A T, Bulut R and Puckette J 2014 Effects of surface roughness on contact angle measurements on a limestone aggregate *Recent Developments in Evaluation of Pavements And Paving Materials Geo-Hubei 2014 (Hubei, China, July 20-22, 2014)* pp. 1–8
[49] Liu W, Luo Y, Sun L, Wu R, Jiang H and Liu Y 2013 Fabrication of the superhydrophobic surface on aluminum alloy by anodizing and polymeric coating *Appl. Surf. Sci.* **264** 872–8
[50] Vishnoi M, Muthupandi V and Murugan S S 2017 Characterization of hydrophobic coating developed by micro arc oxidation on AA2014 alloy *Indian Journal of Engineering & Materials Sciences.* **24** 325–30
[51] Vishnoi M, Kumar P and Murtaza Q 2021 Surface texturing techniques to enhance tribological performance: a review *Surfaces and Interfaces.* **27** 101463
[52] Barthlott W and Neinhuis C 1997 Purity of the sacred lotus, or escape from contamination in biological surfaces *Planta.* **202** 1–8
[53] Zhang Y 2007 The effect of surface roughness parameters on contact and wettability of solid surfaces *Retrospective Theses and Dissertations* Iowa State University https://core.ac.uk/download/pdf/38911929.pdf

Hindawi
Advances in Civil Engineering
Volume 2023, Article ID 9064303, 10 pages
https://doi.org/10.1155/2023/9064303

Research Article

Modeling the Influence of Ground-Granulated Blast Furnace Slag on Hydration of Cement

Andualem Yadeta,[1,2] **Pradeep Goyal**,[1] **and Raju Sarkar**[1]

[1]*Department of Civil Engineering, Delhi Technological University, Delhi 110042, India*
[2]*Department of Construction Technology and Management, Madda Walabu University, Bale Robe 242, Ethiopia*

Correspondence should be addressed to Andualem Yadeta; andu1781@gmail.com

Received 22 April 2023; Revised 6 September 2023; Accepted 27 September 2023; Published 19 October 2023

Academic Editor: Heng-Long Zhang

Copyright © 2023 Andualem Yadeta et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Portland cement is usually substituted with granulated blast furnace slag to make a slag-blended cement. A slag-incorporating cement has a more complicated hydration process than Portland cement due to the interactions between the slag reaction and the hydration of Portland cement in the cementitious systems. Understanding the effect of slag substitution on the hydration of cement is still challenging; hence, this research is aimed at predicting the hydration of slag-incorporating cement. To achieve this, the extended CEMHYD3D model was employed to predict the hydration of a slag-blended cement. The simulation was done with cement paste samples made with various water-to-cement ratios and different slag substitution levels, in which the interaction between the hydration of Portland cement and the reaction of slag was considered. The prediction model has been validated with experimental results and verified to be successful in predicting the hydration of slag-blended cement.

1. Introduction

Supplementary cementitious materials (SCMs) such as ground-granulated blast furnace slag, often called slag cement, fly ash (FA), and silica fume (SF) are widely used in the cementitious composites due to their durability, environmental friendliness, and minimal cost. When added to Portland cement, SCMs improve the characteristics of a cementitious material via either their pozzolanic activity and/or hydraulic activity. They reduce the energy used and the CO_2 released while producing Portland cement [1], and they are widely used in cement-based composites. When compared to the composites made only with Portland cement, composites incorporating SCMs have better durability regarding several deterioration processes, such as chloride penetration and sulfate attack [2]. Slag is among the most frequently used SCMs and is made from iron blast-furnace slag during the reduction of iron ore with coke at a temperature of about 1,500°C. Slag shows high activity during hydration and comprises over 90% of glass because of quick water quenching [3]. Studies conducted by Choi et al. [4], Lee et al. [5], and Cahyani and Rusdianto [6] explained that the substitution of slag for Portland cement enhances the workability and hydration of the process.

At the microscale level, cement particles are typically made up of tricalcium silicate (C_3S), dicalcium silicate (C_2S), tricalcium aluminate (C_3A), and tetra-calcium aluminate ferrite (C_4AF) mineral phases. Cementitious materials require a supply of water for hydration, and moisture exchange can shift the water balance within the material. The volume of water in the component changes due to these two processes. Hydration is highly sensitive to the water-to-cement (w/c) ratio of the cementitious system immediately after it has been prepared. Once the cement and water are mixed, a number of chemical reactions can be performed: C_3S and C_2S produce calcium hydroxide (CH) and calcium–silicate–hydrates (C—S—H), while C_3A and C_4AF in the hydration process contribute to the formation of C_3AH_6 and FH_3 [7].

The substitution of slag for Portland cement can alter the rate of hydration and the microstructural properties of the cementitious materials. The effect of slag on the hydration of cement has been experimentally studied by Kolani et al. [3], Binici et al. [8], and Zhao et al. [9]. In other studies, advanced characterization methods like scanning electron microscopy (SEM) [10] and X-ray diffraction (XRD) [11] were employed to determine the hydration and microstructure of various

blended cements. But techniques like SEM and XRD analysis can also be time-consuming and costly. Computer models are helpful to simulate the hydration of cement. As long as water and space are available, the volume of cement and hydration change over time and continue to do so during the hydration process. Computer models simulate these changes using volume stoichiometries based on the material properties. On the basis of hydration models for cementitious systems that use solely Portland cement, a number of computer-based models such as CEMHYD3D [12], HYMOSTRUC3D [13, 14], μic [15], and DuCOM [16] were developed to predict the hydration of cement and how its microstructure changes over time. Several multiscale models for the hydration of Portland cement have been developed and are now in use throughout the last few decades [17–19]. The existing models of hydration were developed to predict the hydration of composites made of only Portland cement. The rate of hydration in a single cement system was predicted previously by Park [20], in which the rate of hydration that results from the free water reduction and reduced interaction area within the hydration products was considered. The hydration of cements incorporating SCMs like slag is more complex than that of cementitious systems involving only Portland cement owing to the mineral reactions involved in SCMs. The mineral reactions during the hydration of cements involving SCMs such as slag, FA, and limestone are considered distinct from those involving regular Portland cement, and some reactions are considered by looking at the concentration of calcium hydroxide and free water.

A general hydration model was developed by Maekawa et al. [16] for concrete containing FA and slag. A hydration model for limestone containing cement was developed and validated by Mohamed et al. [21]. As described by Swaddiwudhipong et al. [22], the study predicted the hydration of cements containing SF in concrete with a computer model. A multiphase hydration model was proposed by Merzouki et al. [19] and Meinhard and Lackner [23]. The heat generation and the various microstructural characteristics of slag-incorporating cement are all dependent on the hydration of the minerals involved in each cement in the composite. However, there are currently no models available to predict the rate of hydration of slag-incorporating cements. Furthermore, the prediction models that were developed in the previous studies did not consider how slag influences the rate of hydration of a cement system, considering the w/c ratio and the slag replacement level in the system. Therefore, there is a need to consistently predict the hydration of slag-incorporating cementitious systems with the computer models.

2. Hydration Modeling

A modeling software called CEMHYD3D was created for the first time by the National Institute of Standards and Technology to describe the hydration and microstructure of cement [12]. In this model, the microstructure is explained as a grid of volume pixels, each of which stands for a different type of hydration phase. CEMHYD3D is an effective tool for simulating cement hydration that can also involve SCMs [24]. Because the model simulates the cement particles based on actual cement pictures and particle size distributions, it can accurately model the compositions of cement particles and distributions of phases. So, the extended CEMHYD3D version 3.0 modeling package is proposed to predict the hydration of slag-incorporating cements in this study.

2.1. Hydration Model for Portland Cement. The hydration of a single cement particle can be calculated using the basic equation expressed in Equation (1) for each chemical composition in the cement [25]. The activation of chemical reaction and the diffusion process during hydration of a single cement are all taken into account in this model:

$$\frac{d\alpha}{\mathrm{dt}} = \left[\frac{3\left(\frac{S_w}{S_0}\right)\rho_w C_{w\text{-}fre}}{\left(v + w_g\right) r_0 \rho_c}\right] x \left[\frac{1}{\left(\frac{1}{k_d} - \frac{r_0}{D_e}\right) + \frac{r_0}{D_e}(1-\alpha)^{\frac{-1}{3}} + \frac{1}{k_r}(1-\alpha)^{\frac{-2}{3}}}\right], \tag{1}$$

where α is the degree of hydration of cement, v is the mass ratio of water to cement, w_g is the physically bound water in the C—S—H gel, S_w is the effective surface area of the cement particles that are in contact with water, S_0 is the total surface area, $C_{w\text{-}fre}$ is the physically free water in the C—S—H gel, r_0 is the radii of an unhydrated cement particle, ρ_c is the density of cement, and ρ_w is the density of water, k_d is the reaction coefficient in the dormant period, k_r is the coefficient of the rate of reaction for each mineral, and D_e is the effective diffusion coefficient of water in the hydration products for each mineral constituents.

The reaction coefficient, k_d in Equation (2) is assumed to be proportional to the rate of hydration of cement, whereas the values of coefficients B and C govern the initial rates of shell formation and decay, respectively.

$$k_d = \frac{B}{\alpha^{1.5}} + C\alpha^3. \tag{2}$$

In regard to hydrates, the effective diffusion coefficient of water changes based on the radii of the gel pore in the cementitious system. The following Equation (3) presents an expression for this phenomenon in terms of the rate of hydration.

$$D_{e0} = \ln\left(\frac{1}{\alpha}\right). \tag{3}$$

Moreover, when the hydration of the minerals in each cement continues, the free water in the capillary pores reduces. Therefore, it is important to consider that some amount of water can be trapped in the gel pores and may be used for further hydration at every stage of hydration [20]. Hence, as shown in Equation (4), the amount of water in the capillary pore varies with the rate of hydration.

Table 1: Reaction of slag [33].

Quantity measured	Values (g/g of slag)
Calcium hydroxide	0.22
Chemically bound water	0.30
Gel water	0.15

$$C_{w\text{-}free} = \frac{W_0 - 0.4\alpha C_0}{W_0}, \quad (4)$$

where C_0 and W_0 represent the initial percentages by mass of cement and water in the mixture.

Hydration modeling covers the chemical effects of the cement minerals in the hydration reactions. According to the theories put forth by Papadakis [26], Lu et al. [27], and Maciel et al. [28], the reactions that the minerals in Portland cement undergo during the hydration stage can be described with Equations (5)–(8) as follows:

$$2C_3S + 6H \rightarrow C_3S_2H_3 + 3\text{CH}, \quad (5)$$

$$2C_2S + 4H \rightarrow C_3S_2H_3 + \text{CH}, \quad (6)$$

$$C_3A + C\overline{S}H_2 \rightarrow 10H + C_4\text{A}\overline{S}\overline{H}_{12}, \quad (7)$$

$$C_4\text{AF} + 2\text{CH} \rightarrow 10H + C_6\text{AFH}_{12}. \quad (8)$$

Both CH and C—S—H are formed when C_3S and C_2S are hydrated with water. As long as C_3A or C_4AF aluminate clinker phases are present, aluminate is absorbed by C—S—H which produces a C—A—S—H product. If no remaining aluminate reactants are left, the C_3S and C_2S reactions further yield CSH without aluminum substitution. However, using Equations (5)–(8), the mass per volume of the CH and C—S—H products in Portland cement can be calculated.

2.2. Hydration Model for Slag-Incorporating Cement. There are five major hydration products that occur when slag is added to cement: C—S—H, C—S—A—H, hydrotalcite (M_5AH_{13}), ettringite ($C_6A\overline{S}_3H_{32}$), and AFm phase (C_4AH_{13}) are usually formed [29]. While both slag and Portland cement produce similar hydration products, part of the portlandite produced when cement clinker hydrates is actually consumed when slag is hydrated [30–32]. Nevertheless, a high rate of slag reaction requires a lower w/c ratio [26]. The reaction of slag during hydration is appropriately summarized by the approximate critical values shown in Table 1, as reported by Maekawa et al. [33], based on his study conducted on the blended cements.

In accordance with the stoichiometry and hydration model of the slag reaction [33], the mass of calcium hydroxide, the chemically bound water, and the gel water may all be computed with Equations (9)–(11) in which the hydration of slag-incorporating cements can be described as follows:

$$\text{CH} = \text{RCH}_{ce}\, x C_0 x\alpha - \text{RCH}_{sg} x\alpha_{sg} xP, \quad (9)$$

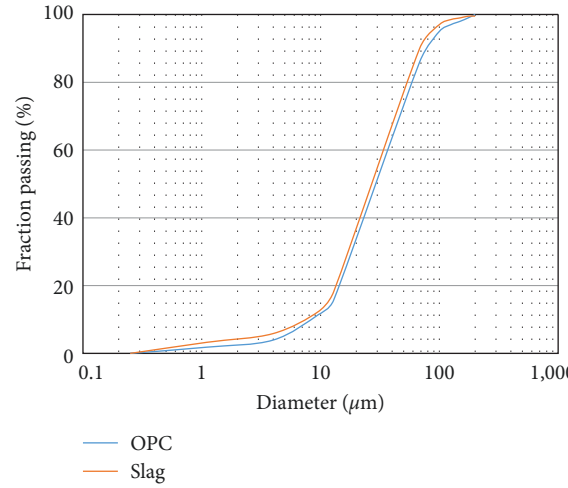


Figure 1: Particle size distribution of OPC and slag used in the model.

$$W_{\text{cap}} = W_0 - 0.4xC_0x\alpha - \text{RCW}_{sg}\, x\alpha_{sg}\text{xP} - \text{RPW}_{sg} x\alpha_{sg} xP, \quad (10)$$

$$W_{\text{cbw}} = vxC_0x\alpha + \text{RCW}_{sg} x\alpha_{sg} xP, \quad (11)$$

where CH, W_{cap}, and W_{cbw} are the mass of calcium hydroxide, capillary water, and chemically bound water, respectively. sg is the rate of the slag reaction, P is the mass of the slag in the mix, RCH_{ce} and RCH_{sg} are the mass of calcium hydroxide formed in the hydration of cement, and the slag reaction in that order, where as RCW_{sg} and RPW_{sg} are the mass of chemically bound water and the gel water in the slag reaction, respectively.

3. Experiments

An OPC as a reference cement confirmed to ASTM C 150 and a slag confirmed to ASTM C 989 standard requirements were used in this study. Before the simulation of the rate of hydration of slag-incorporating cement with the proposed model, the individual cements were first characterized according to their particle size distribution (PSD) and chemical compositions. The average PSD for OPC and slag was measured with a laser diffraction (LD) technique and plotted in Figure 1. The hydration model development needs the consideration of some reaction chemistry, for which standard abbreviations are used throughout the model: C stands for CaO, S for SiO_2, A for Al_2O_3, F for Fe_2O_3, M for MgO, and H for H_2O. The hydration equations of the four cement clinker phases (C_3S, C_2S, C_3A, and C_4AF) and slag reactions are considered in the model.

As a reference cement, an OPC with the chemical composition presented in Table 2 is used to determine the influence of the substitution of slag into cement. While simulating the hydration of the blended cement, the w/c ratio and the slag replacement level were considered.

The proposed CEMHYD3D model from open source offers the possibility of simulating the hydration of cement.

Table 2: Chemical composition of the OPC used in the model (mass%).

Clinker phase	Volume fraction	Surface area fraction
C_3S	0.7229	0.6689
C_2S	0.0961	0.1377
C_3A	0.1291	0.1346
C_4AF	0.0397	0.0391

Table 3: Compound composition of the slag used in the model (mass%).

SiO_2 (%)	CaO (%)	Al_2O_3 (%)	SO_3 (%)	MgO (%)	Fe_2O_3 (%)	K_2O (%)	Na_2O (%)
40.23	35.04	11.18	6.80	2.32	2.05	0.43	0.29

Moreover, the code in the program allows us to adjust the model for extension as necessary.

To extend the CEMHYD3D model and enhance its predictive capabilities, a critical refinement has been introduced by incorporating the reaction of slag into the program. This significant modification is applied because the rate of hydration in cementitious systems is intricately linked to the mineral composition of each specific type of cement.

The compound composition of the slag used in the model is presented in Table 3. In a slag-blended cement system, the simplest and most direct approach is to assume that the slag reacts with CH in the hydration process and yields a single product with a mixed phase. This process provides a C—S—H product that has a lower ratio of Ca/Si than the typical C—S—H product, amorphous silica (AFm), and some hydrotalcite products [34, 35].

Various samples of cement pastes were prepared with a w/c ratio of 0.3, 0.4, 0.5, and 0.6 at a slag replacement level of 10%, 20%, 30%, and 40% by mass of cement. The pastes were made by mixing the raw materials in a vertical axis pot mixer with water first, then slowly adding cementitious materials while the blades were rotating; the subsequent mixing process included a 3-min low speed mixing, a half-minute pause, and a 2-min high speed mixing. After being mixed, fresh mixtures were cast into plastic tubes of φ15 mm, and the two ends of the tubes were sealed to prevent moisture exchange. The sealed specimens were cured in a saturated condition at a standard temperature of 20°C until the age of testing. The specimens were tested at specific ages of 1, 3, 7, 28, and 90 days. The contents of hydrate water and CH in the Portland cement and blended cement pastes were determined using thermogravimetric analysis (TGA). The specimens were removed from the plastic tubes, ground to a fine powder with particles of size $<60\,\mu$m in a chamber, and flushed with methanol in a thistle tube connected to a vacuum to arrest the hydration. After drying, about 40 mg of the powder sample was analyzed by measuring the weight while heating the sample up from standard temperature to 550°C in TGA. Weight loss in the ranges of 450–550°C was used to compute the contents of hydrate water and CH, respectively [36–38]. The amount of hydrate water per gram of binder (w_H) was calculated by normalizing the hydrate water content by the mass loss of the sample between 450–550°C divided by the mass of the ignited sample, corrected for the measured loss-on-ignition of the unhydrated cement (or of the unhydrated cement divided by the percentage of slag for slag-blended cement). The values of w_H were converted to estimated degrees of hydration (α) based on the phase composition of the cement used in the study.

4. Simulation of the Influence of Slag on Hydration of Cement

Cement pastes with various values of w/c and replacement levels of OPC with slag were created using the extended CEMHYD3D version model to match the experimental mix. In all simulations of the proposed cement paste, a conversion factor of 0.00035 hr/cycle^2 as described by Bentz [12] has been applied to make conversions between the model hydration cycles and the real time. As discussed in previous sections, it was assumed in this study that the slag in a blended cement reacts with CH and provides a single product of hydration in a mixed phase. With the proposed model, the hydration of slag-incorporating cement was simulated using cement paste samples made with a replacement level of 10%, 20%, 30%, and 40% by mass of slag and a w/c of 0.3, 0.4, 0.5, and 0.6. The prediction has been validated with measured values from experiments based on the standard test method of ASTM C186. Previously, experimental studies were conducted by Zhao et al. [9], Durdziński et al. [39], and Hlobil et al. [40]. In slag-incorporating cement, the amount of CH and the rate of reaction of mineral components determine the rate of hydration of the cement [3, 9]. This proves that the hydration product of slag usually develops in close proximity to the slag particles [41–43]. Using the approach suggested by Saeki and Monteiro [44], the equation for the reaction of slag can be expressed with Equations (12)–(14) as follows, considering the aforementioned factors:

$$\frac{d\alpha_{sg}}{dt} = \left(\frac{m_{CH}(t)}{P}\right)\left(\frac{W_{cap}}{W_0}\right)\left(\frac{3P_W}{\nu_{sg} r_{sgo} \rho_{sg}}\right) x \frac{1}{\left(\frac{1}{k_{dsg}} - \frac{r_{SO0}}{D_{esg}}\right)\left(1-\alpha_{sg}\right)^{-\frac{1}{3}} + \left(\frac{1}{k_{rsg}}\right)\left(1-\alpha_{sg}\right)^{-\frac{2}{3}}}, \quad (12)$$

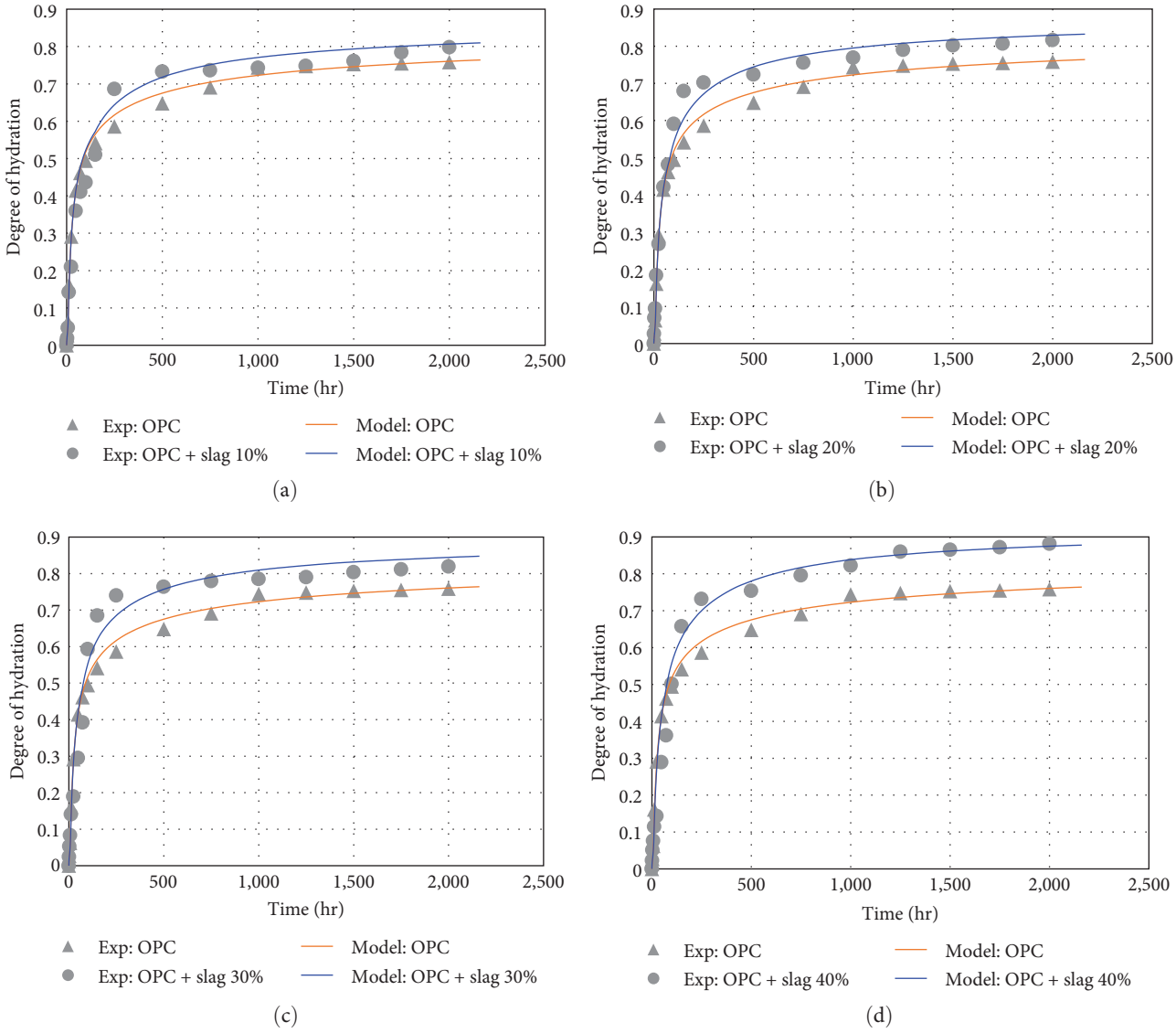


FIGURE 2: Simulated degree of hydration of cement with different % by mass slag substitution for w/c = 0.3. (a) Cement paste without and with 10% slag. (b) Cement paste without and with 20% slag. (c) Cement paste without and with 30% slag. (d) Cement paste without and with 40% slag.

$$k_{dsg} = \frac{B_{sg}}{(\alpha_{sg})^{1.5}} + C_{sg}(\alpha_{sg})^3, \tag{13}$$

$$D_{esg} = D_{esg0} x \ln\left(\frac{1}{\alpha_{sg}}\right), \tag{14}$$

where $m_{CH}(t)$ is the mass of calcium hydroxide formed in the hydrated slag-blended mixes obtained from Equation (9), D_{esg0} is the initial diffusion coefficient, ρ_{sg} is the density of slag, k_{dsg} is the rate of reaction coefficient of slag in the dormant phase (B_{sg} and C_{sg} are coefficients), and k_{rsg} is the rate of reaction coefficient of slag.

5. Analysis and Discussion

Figures 2–5 shows the hydration of slag-incorporated cements, which were simulated with the extended CEMHYD3D model for different w/c values and various replacement levels. The simulations with the computer model were carried out under the same conditions as the experiments. The model provided a good prediction of the influence of slag on the hydration of cement in all simulations, which agreed with the experimental results.

The hydration of slag-incorporated cement is simulated in Figure 2 separately as a, b, c, and d in volume fractions of each cement. The model predicted the hydration of blended cement pastes without and with 10%, 20%, 30%, and 40% slag substitution made with a w/c of 0.3 for all mixes under sealed conditions at a temperature of 20°C. It was observed

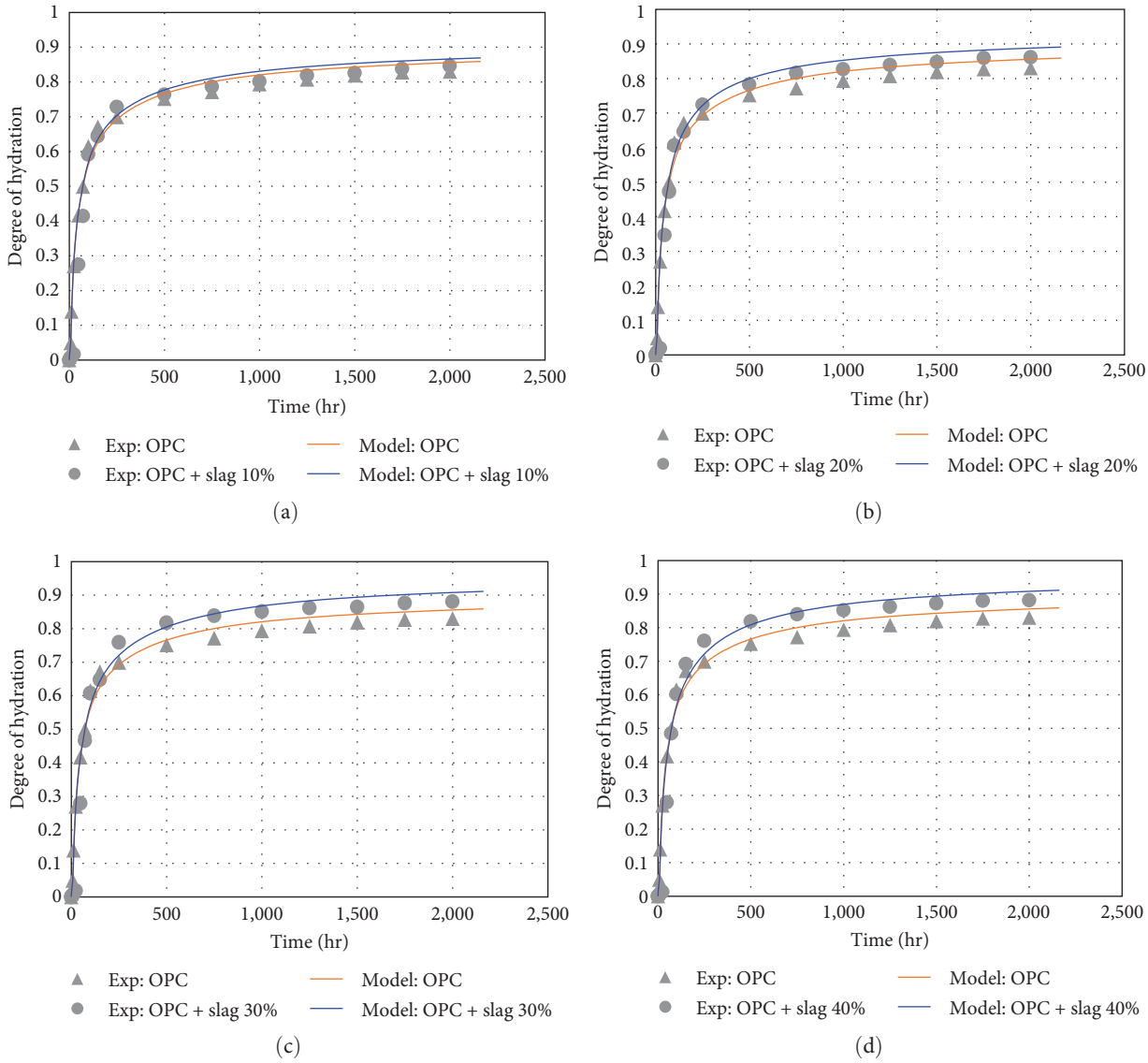


FIGURE 3: Simulated degree of hydration of cement with different % by mass slag substitution for w/c = 0.4. (a) Cement paste without and with 10% slag. (b) Cement paste without and with 20% slag. (c) Cement paste without and with 30% slag. (d) Cement paste without and with 40% slag.

that the cement clinker phase reacted much faster than the reaction of slag in the early ages. Slag can significantly reduce the heat needed for hydration when substituted for cement [45]. However, the slag continues to react at a later age while the clicker phase reaction decreases. In all mixes, the slag reaction increases as the substitution percentage of slag increases, and it was observed that the rate of hydration at a later age increases when a lower w/c of about 0.3 is used in the mix.

As shown in Figure 3 below, the model predicted the influence of slag on the rate of hydration of cement for pastes without and with 10%, 20%, 30%, and 40% slag substitution for a w/c of 0.4 for all mixes under sealed conditions at 20°C. The prediction indicated that no substantial effect of slag on the rate of hydration of cement was observed when up to 10% slag was substituted for a w/c of 0.4. Even though the reaction of clicker phases in the early age is faster than the reaction of slag, the simulation illustrates that the hydration of the blend increases as the substitution level of slag into cement increases in the later age, when a w/c of about 0.4 is used for the paste.

The prediction in a, b, c, and d of Figure 4 illustrates that the rate of reaction of slag at a replacement level of 10%, 20%, 30%, and 40% is similar in most cases with the paste with 0% slag cement when a w/c of 0.5 is utilized for the mix. From the model prediction, no significant effect was observed between the hydration of Portland cement and a slag-incorporating cement in the early ages, when about 10% slag was substituted. In contrast, using a w/c of 0.5 accelerates the rate of hydration in the later age at about 90 days and increases as the substitution percentages of slag in the blend increase.

The predicted values for hydration of a slag-blended cement made with a w/c of 0.6 and varying slag substitution

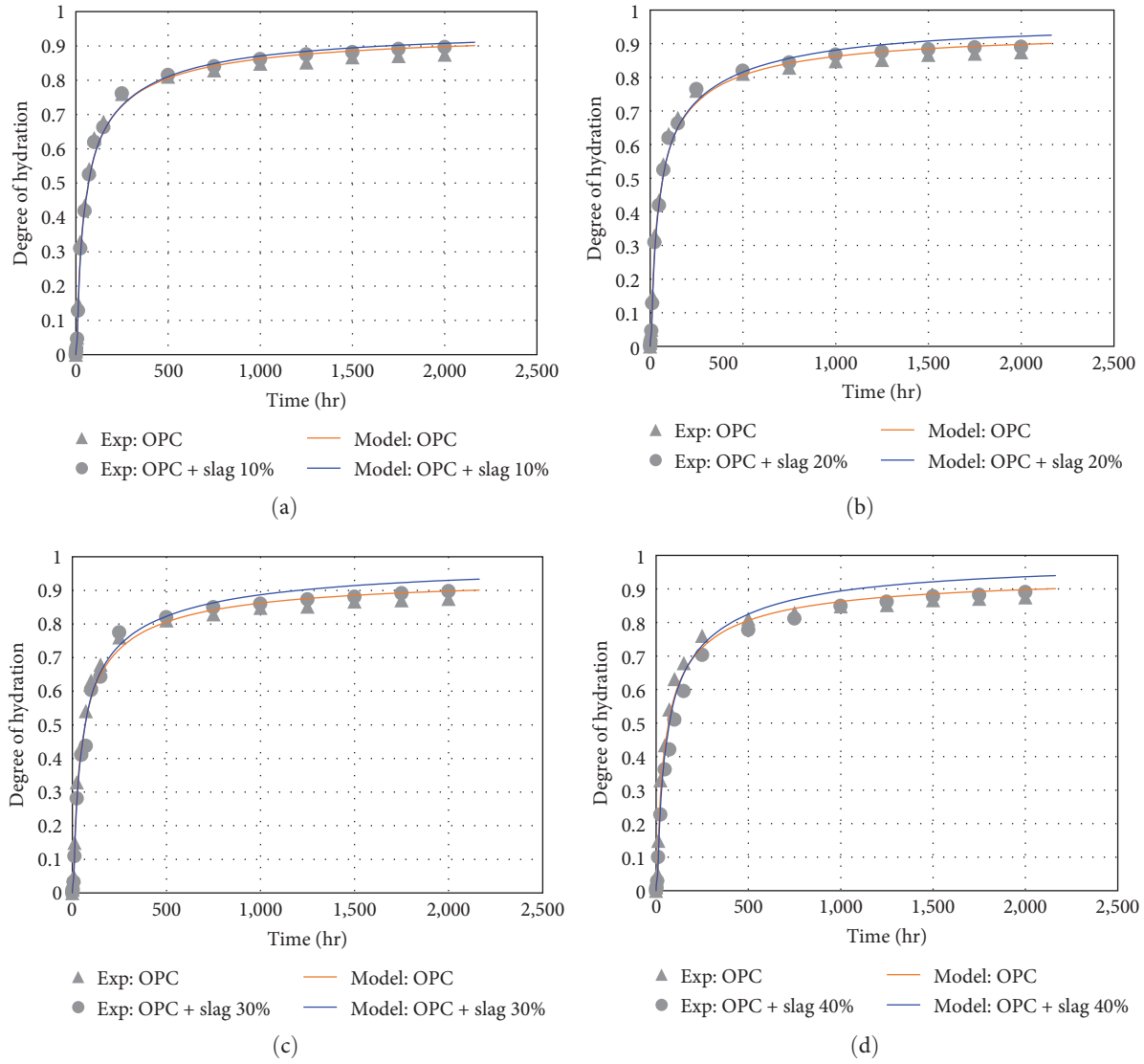


FIGURE 4: Simulated degree of hydration of cement with different % by mass slag substitution for w/c = 0.5. (a) Cement paste without and with 10% slag. (b) Cement paste without and with 20% slag. (c) Cement paste without and with 30% slag. (d) Cement paste without and with 40% slag.

levels are also plotted in a, b, c, and d of Figure 5. The rate of hydration of Portland cement and up to 10% by mass slag substituted cement paste is similar when w/c is around 0.6. When compared with the clicker phase, the reaction of slag slightly increases as the slag substitution level increases, even at the later age of 90 days when the same w/c is used in both mixes.

The hydration of a slag-incorporating cement was accurately predicted using the extended CEMHYD3D model. The hydration of cement pastes made with a w/c ratio of 0.3, 0.4, 0.5, and 0.6 and with a slag replacement level of 10%, 20%, 30%, and 40% was simulated for the age of 90 days cured in a saturated condition at a standard temperature of 20°C. The model prediction and the measured values from experiments were in good agreement in all mixes. The simulation shows that the slag-incorporating cement was predicted to have greater hydration than the Portland cement in the later ages. The rate of hydration of the blend with slag is found to be higher for a w/c of 0.3 and negligible when a w/c of about 0.6 is used. Hence, a greater slag reaction for hydration is preferred at a higher w/c. This makes sense given that there is more space and water in the paste available for reactions and the formation of products. Moreover, there is more water accessible for cement hydration at the paste level. The simulation clearly shows that the rate of reaction of the slag in the blend increases proportionally as the slag substitution level increases. According to Ahmad et al. [46], when the fraction of slag is increased up to 60% in cement, it reduces the amount of time required to reach a maximal rate of hydration. However, no substantial change was observed in the rate of hydration for a slag substitution up to 10% in Portland cement.

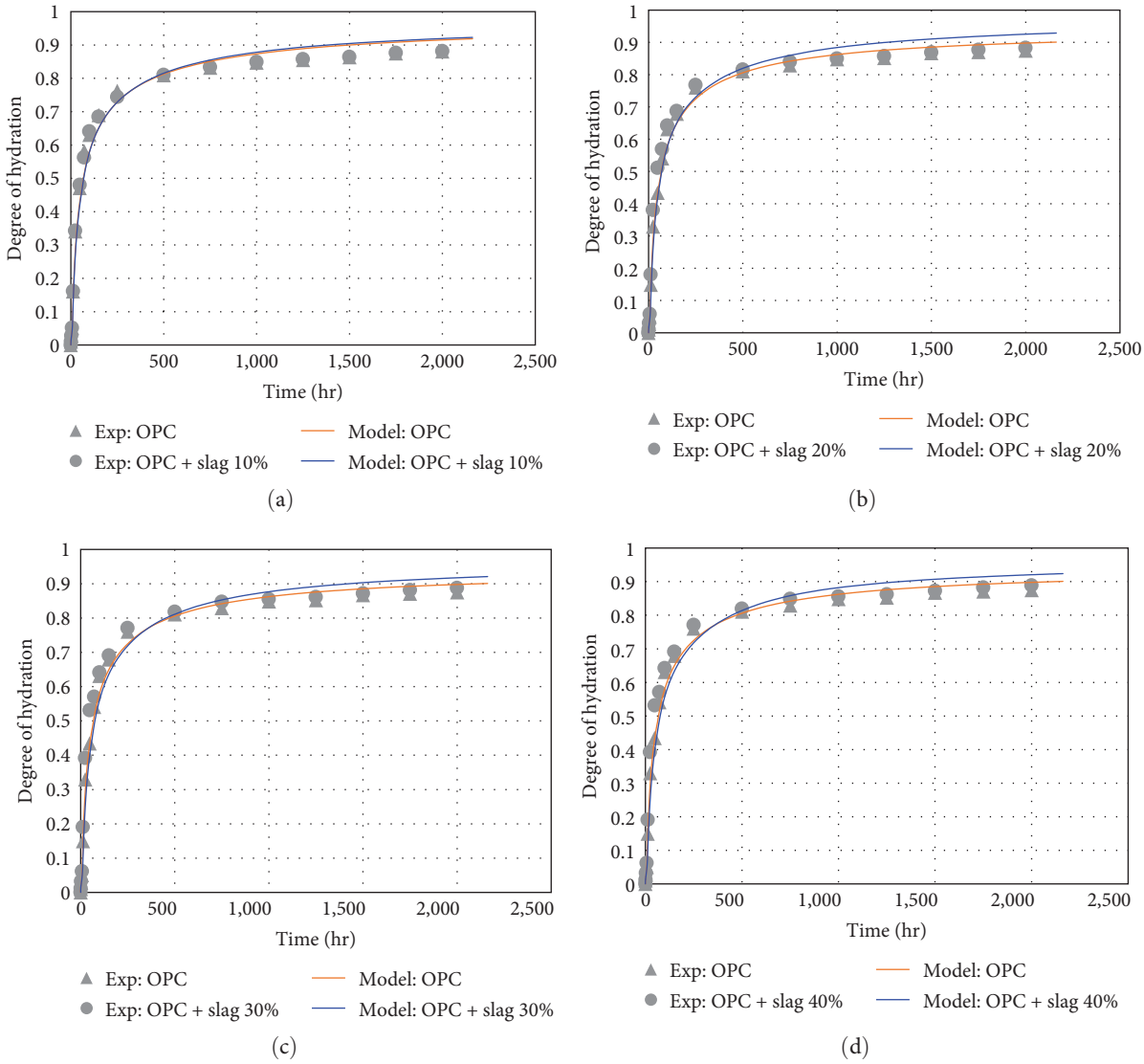


FIGURE 5: Simulated degree of hydration of cement with different % by mass slag substitution for w/c = 0.6. (a) Cement paste without and with 10% slag. (b) Cement paste without and with 20% slag. (c) Cement paste without and with 30% slag. (d) Cement paste without and with 40% slag.

6. Conclusions

The paper presented the simulation of the influence of slag substitution on the hydration of cement using the extended CEMHYD3D version 3.0 model. The model predictions agreed with the experimental values regarding the w/c ratio and substitution percentage of slag in Portland cement. The possible influence of slag on the hydration of cement is higher when a lower w/c of about 0.3 is used in the mixture in the later ages as the replacement level of slag increases. The increase in w/c ratio to about 0.6 can accelerate the early hydration of slag-blended cement systems but it has a negligible effect in the later ages as the replacement level of slag increases. Thus, slag substitution in Portland cement is recommended with a higher w/c ratio to achieve an optimal hydration result. The rate of slag reaction in a blended cement system is variable and dependent on the substitution level of the slag. The consideration of the chemical composition of slag improves the predictability of hydration in cement, and the same method was applied in this model. Both in the early age and up to the age of 90 days, the hydration of slag-incorporating cement can be predicted with the proposed model.

Data Availability

The data supporting the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by Ministry of Education of the Federal Democratic Republic of Ethiopia.

References

[1] D. Wang, X. Zhou, Y. Meng, and Z. Chen, "Durability of concrete containing fly ash and silica fume against combined freezing-thawing and sulfate attack," *Construction and Building Materials*, vol. 147, pp. 398–406, 2017.

[2] S.-W. Yoo and S.-J. Kwon, "Effects of cold joint and loading conditions on chloride diffusion in concrete containing GGBFS," *Construction and Building Materials*, vol. 115, pp. 247–255, 2016.

[3] B. Kolani, L. Buffo-Lacarrière, A. Sellier, G. Escadeillas, L. Boutillon, and L. Linger, "Hydration of slag-blended cements," *Cement and Concrete Composites*, vol. 34, no. 9, pp. 1009–1018, 2012.

[4] Y. C. Choi, J. Kim, and S. Choi, "Mercury intrusion porosimetry characterization of micropore structures of high-strength cement pastes incorporating high volume ground granulated blast-furnace slag," *Construction and Building Materials*, vol. 137, pp. 96–103, 2017.

[5] J.-Y. Lee, J.-S. Choi, T.-F. Yuan, Y.-S. Yoon, and D. Mitchell, "Comparing properties of concrete containing electric arc furnace slag and granulated blast furnace slag," *Materials*, vol. 12, no. 9, Article ID 1371, 2019.

[6] R. A. T. Cahyani and Y. Rusdianto, "Concrete performance with ground granulated blast furnace slag as supplementary cementitious materials," *IOP Conference Series: Materials Science and Engineering*, vol. 771, no. 1, Article ID 012062, 2020.

[7] H. Minard, S. Garrault, L. Regnaud, and A. Nonat, "Mechanisms and parameters controlling the tricalcium aluminate reactivity in the presence of gypsum," *Cement and Concrete Research*, vol. 37, no. 10, pp. 1418–1426, 2007.

[8] H. Binici, H. Temiz, and M. M. Köse, "The effect of fineness on the properties of the blended cements incorporating ground granulated blast furnace slag and ground basaltic pumice," *Construction and Building Materials*, vol. 21, no. 5, pp. 1122–1128, 2007.

[9] Y. Zhao, J. Gao, Z. Xu, S. Li, X. Luo, and G. Chen, "Long-term hydration and microstructure evolution of blended cement containing ground granulated blast furnace slag and waste clay brick," *Cement and Concrete Composites*, vol. 118, Article ID 103982, 2021.

[10] H. Zeng, S. Qu, and Y. Qin, "Microstructure and transport properties of cement-based material enhanced by graphene oxide," *Magazine of Concrete Research*, vol. 73, no. 19, pp. 1011–1024, 2021.

[11] M. Zhang, "Pore-scale modelling of relative permeability of cementitious materials using X-ray computed microtomography images," *Cement and Concrete Research*, vol. 95, pp. 18–29, 2017.

[12] D. P. Bentz, *CEMHYD3D: A Three-Dimensional Cement Hydration and Microstructure Development Modeling Package*, National Institute of Standards and Technology, Gaithersburg, MD, Version 3.0., NIST Interagency/Internal Report (NISTIR), 2005.

[13] K. Van Breugel, "Numerical simulation of hydration and microstructural development in hardening cement-based materials (I) theory," *Cement and Concrete Research*, vol. 25, no. 2, p. 319, 331, 1995.

[14] P. Gao, G. Ye, J. Wei, and Q. Yu, "Extension of the Hymostruc3D model for simulation of hydration and microstructure development of blended cements," *Heron*, vol. 64, no. 1-2, pp. 125–148, 2019.

[15] S. Bishnoi and K. L. Scrivener, "μic: A new platform for modelling the hydration of cements," *Cement and Concrete Research*, vol. 39, no. 4, pp. 266–274, 2009.

[16] K. Maekawa, T. Ishida, and T. Kishi, "Multi-scale modeling of concrete performance integrated material and structural mechanics," *Journal of Advanced Concrete Technology*, vol. 1, no. 2, pp. 91–126, 2003.

[17] J. Carette and S. Staquet, "Monitoring and modelling the early age and hardening behaviour of eco-concrete through continuous non-destructive measurements: Part I. Hydration and apparent activation energy," *Cement and Concrete Composites*, vol. 73, pp. 10–18, 2016.

[18] F. Lin and C. Meyer, "Hydration kinetics modeling of Portland cement considering the effects of curing temperature and applied pressure," *Cement and Concrete Research*, vol. 39, no. 4, pp. 255–265, 2009.

[19] T. Merzouki, M. Bouasker, N. E. H. Khalifa, and P. Mounanga, "Contribution to the modeling of hydration and chemical shrinkage of slag-blended cement at early age," *Construction and Building Materials*, vol. 44, pp. 368–380, 2013.

[20] K.-B. Park, "Prediction of temperature and moisture distributions in hardening concrete by using a hydration model," *Architectural Research*, vol. 14, no. 4, pp. 153–161, 2012.

[21] A. R. Mohamed, M. Elsalamawy, and M. Ragab, "Modeling the influence of limestone addition on cement hydration," *Alexandria Engineering Journal*, vol. 54, no. 1, pp. 1–5, 2015.

[22] S. Swaddiwudhipong, H. Wu, and M. H. Zhang, "Numerical simulation of temperature rise of high-strength concrete incorporating silica fume and superplasticiser," *Advances in Cement Research*, vol. 15, no. 4, pp. 161–169, 2003.

[23] K. Meinhard and R. Lackner, "Multi-phase hydration model for prediction of hydration-heat release of blended cements," *Cement and Concrete Research*, vol. 38, no. 6, pp. 794–802, 2008.

[24] N. Holmes, "Introducing a new cement hydration and microstructure model," *Journal of Material Science & Engineering*, vol. 7, no. 4, Article ID 1000480, 2018.

[25] J. J. Thomas, J. J. Biernacki, J. W. Bullard et al., "Modeling and simulation of cement hydration kinetics and microstructure development," *Cement and Concrete Research*, vol. 41, no. 12, pp. 1257–1278, 2011.

[26] V. G. Papadakis, "Effect of supplementary cementing materials on concrete resistance against carbonation and chloride ingress," *Cement and Concrete Research*, vol. 30, no. 2, pp. 291–299, 2000.

[27] X. Lu, Z. Ye, S. Wang, P. Du, C. Li, and X. Cheng, "Study on the preparation and properties of belite-ye'elimite-alite cement," *Construction and Building Materials*, vol. 182, pp. 399–405, 2018.

[28] M. H. Maciel, G. S. Soares, R. C. de O. Romano, and M. A. Cincotto, "Monitoring of Portland cement chemical reaction and quantification of the hydrated products by XRD and TG in function of the stoppage hydration technique," *Journal of Thermal Analysis and Calorimetry*, vol. 136, no. 3, pp. 1269–1284, 2019.

[29] W. Chen and H. J. H. Brouwers, "The hydration of slag, Part 2: reaction models for blended cement," *Journal of Materials Science*, vol. 42, no. 2, pp. 444–464, 2007.

[30] K. Luke and E. Lachowski, "Internal composition of 20-year-old fly ash and slag-blended ordinary Portland cement pastes," *Journal of the American Ceramic Society*, vol. 91, no. 12, pp. 4084–4092, 2008.

[31] R. Taylor, I. G. Richardson, and R. M. D. Brydson, "Composition and microstructure of 20-year-old ordinary Portland

cement-ground granulated blast-furnace slag blends containing 0–100% slag," *Cement and Concrete Research*, vol. 40, no. 7, pp. 971–983, 2010.

[32] J. J. Thomas, A. J. Allen, and H. M. Jennings, "Density and water content of nanoscale solid C–S–H formed in alkali-activated slag (AAS) paste and implications for chemical shrinkage," *Cement and Concrete Research*, vol. 42, no. 2, pp. 377–383, 2012.

[33] K. Maekawa, T. Ishida, and T. Kishi, *Multi-Scale Modeling of Structural Concrete*, Taylor & Francis, New York, 1st edition, 2009.

[34] J. Xing, Y. Zhao, J. Qiu, and X. Sun, "Microstructural and mechanical properties of alkali activated materials from two types of blast furnace slags," *Materials*, vol. 12, no. 13, Article ID 2089, 2019.

[35] Q. Fu, M. Bu, Z. Zhang, W. Xu, Q. Yuan, and D. Niu, "Hydration characteristics and microstructure of alkali-activated slag concrete: a review," *Engineering*, vol. 20, pp. 162–179, 2023.

[36] J. Zhang, F. Zheng, Z. Liu, S. Hong, B. Dong, and F. Xing, "Nondestructive monitoring on hydration behavior of cement pastes via the electrochemical impedance spectroscopy method," *Measurement*, vol. 185, Article ID 109884, 2021.

[37] S. A. Bernal, M. C. G. Juenger, X. Ke et al., "Characterization of supplementary cementitious materials by thermal analysis," *Materials and Structures*, vol. 50, no. 1, Article ID 26, 2017.

[38] S. Adu-Amankwah, M. Zajac, C. Stabler, B. Lothenbach, and L. Black, "Influence of limestone on the hydration of ternary slag cements," *Cement and Concrete Research*, vol. 100, pp. 96–109, 2017.

[39] P. T. Durdziński, M. B. Haha, M. Zajac, and K. L. Scrivener, "Phase assemblage of composite cements," *Cement and Concrete Research*, vol. 99, pp. 172–182, 2017.

[40] M. Hlobil, V. Šmilauer, and G. Chanvillard, "Micromechanical multiscale fracture model for compressive strength of blended cement pastes," *Cement and Concrete Research*, vol. 83, pp. 188–202, 2016.

[41] C. Çetin, S. T. Erdoğan, and M. Tokyay, "Effect of particle size and slag content on the early hydration of interground blended cements," *Cement and Concrete Composites*, vol. 67, pp. 39–49, 2016.

[42] X.-Y. Wang, "Analysis of hydration kinetics and strength progress in cement-slag binary composites," *Journal of Building Engineering*, vol. 35, Article ID 101810, 2021.

[43] X. Feng, E. J. Garboczi, D. P. Bentz, P. E. Stutzman, and T. O. Mason, "Estimation of the degree of hydration of blended cement pastes by a scanning electron microscope point-counting procedure," *Cement and Concrete Research*, vol. 34, no. 10, pp. 1787–1793, 2004.

[44] T. Saeki and P. J. M. Monteiro, "A model to predict the amount of calcium hydroxide in concrete containing mineral admixtures," *Cement and Concrete Research*, vol. 35, no. 10, pp. 1914–1921, 2005.

[45] S. Mengxiao, W. Qiang, and Z. Zhikai, "Comparison of the properties between high-volume fly ash concrete and high-volume steel slag concrete under temperature matching curing condition," *Construction and Building Materials*, vol. 98, pp. 649–655, 2015.

[46] J. Ahmad, K. J. Kontoleon, A. Majdi et al., "A comprehensive review on the ground granulated blast furnace slag (GGBS) in concrete production," *Sustainability*, vol. 14, no. 14, Article ID 8783, 2022.

Proceedings of the Fourth International Conference on Smart Electronics and Communication (ICOSEC-2023)
IEEE Xplore Part Number : CFP23V90-ART ; ISBN : 979-8-3503-0088-8

Modelling and Manufacturing of Abrasive Water Jet Machining of Reinforced AI Alloy Cast Hybrid Nanocomposites with Mechanical and Metallurgical Behavior through Artificial Neural Network

[1]Kingshuk Das Bakshi,
Department of Computer Science and Engineering, Ramgovind Institute of Technology Koderma, Jharkhand-825409, India, kingshukdasbakshi@gmail.com

[2]Thilagham K T,
Department of Metallurgical Engineering, Government College of Engineering, Salem, 636011,Tamil nadu, India,
thilagham.met@gmail.com

[3]R. Yamini,
Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur -603203,Tamil Nadu, India,
yaminir@srmist.edu.in

[4]C Vijaya Bhaskar Reddy,
Department of Mechanical, Sri Venkateswara College of Engineering and Technology, RVS Nagar, Andhra Pradesh, India,
bhaskarreddychalla@gmail.com

[5]Anjali Sandeep Gaikwad,
Bharati Vidyapeeth (Deemed to be University) Institute of Management,Kolhapur,India,
anjali.gaikwad@bharatividyapeeth.edu

[6]Anshika Gupta,
Research Scholar QIP, Department of Mechanical Production and Automobile Engineering, Delhi Technological University, Delhi, India,
anshikaguptaji@gmail.com

2023 4th International Conference on Smart Electronics and Communication (ICOSEC) | 979-8-3503-0088-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICOSEC58147.2023.10276082

***Abstract*— This study presents a comprehensive investigation on the abrasive water jet machining of reinforced AI 6081 alloy cast hybrid nanocomposites with mechanical and metallurgical behavior using artificial neural networks (ANNs). Al2O3 and SiC nano-particles were added to the mixture using a stir-casting process. The abrasive jet machining experiments were conducted based on Taguchi L9 array of experiments, where surface roughness and rate of material removal were the responses measured. Optimization was accomplished by means of Taguchi method and ANN was used to predict the results. The results showed that the addition of nano-particles significantly improved the mechanical and metallurgical properties of the AI 6081 alloy cast hybrid nanocomposites. Furthermore, the outcomes indicated that the ANN prototypical was capable of accurately predicting the surface roughness and rate of material removal throughout abrasive water jet machining, providing a valuable tool for the optimization and improvement of machining processes. The findings of this study provide valuable insights into the abrasive water jet machining of AI 6081 alloy cast hybrid nanocomposites and will be useful for further research in the field of nanocomposite processing and machining.**

Keywords—ANN, nanoparticles, Taguchi, abrasive machining

I. INTRODUCTION

The use of abrasive water jet machining (AWJM) has gained considerable attention in recent years due to its ability to machine a wide range of materials with high precision and low thermal damage. Reinforced Al alloy cast hybrid nanocomposites have emerged as a promising material for various industrial applications due to their unique combination of metallurgical and tribological characteristics. [1] However, the machining of these materials using conventional methods has proven to be challenging due to their high hardness, wear resistance, and poor thermal conductivity. Recently, researchers have focused on the use of AWJM to machine-reinforced Al alloy cast hybrid nanocomposites. A study by investigated the AWJM of reinforced Al alloy cast hybrid nanocomposites and found that the totalling of nano-particles significantly improved the mechanical properties of the material. The authors also noted that the surface roughness increased with the addition of nano-particles, which resulted in a lower rate of material removal [2]. Another study by investigated the effect of process parameters on the AWJM of reinforced Al alloy cast hybrid nanocomposites and found that the rate of material removal increased with increasing abrasive water jet pressure and standoff distance [3]. The authors also noted that the surface roughness decreased with increasing abrasive water jet pressure and standoff distance [4].

The optimization of AWJM processes is a crucial aspect of the machining of reinforced Al alloy cast hybrid nanocomposites. A study used Taguchi method to optimize the AWJM process parameters and found that the optimal process limits resulted in improved surface roughness and rate of material removal [5]. Artificial neural networks (ANNs) have been widely used in the optimization of machining processes due to their ability to accurately predict the response of complex systems. A study used ANNs to predict the roughness of the surface and removal of the material during the AWJM of reinforced Al alloy cast hybrid nanocomposites. The authors found that the ANN prototypical was accomplished of precisely forecasting the surface roughness and rate of material removal, providing a valuable instrument for the optimization and development of the machining process [6], [7].

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:28:26 UTC from IEEE Xplore. Restrictions apply.

Additionally, research has also been done on the use of different nano-particles and their effect on the machining of reinforced Al alloy cast hybrid nanocomposites. A study by investigated the use of Al2O3 and SiC nano-particles and found that the addition of SiC nano-particles resulted in higher wear resistance and improved the mechanical properties of the material compared to the addition of Al2O3 nano-particles. This indicates that the choice of nano-particles has a significant impact on the machining performance of reinforced Al alloy cast hybrid nanocomposites [8], [9].

Furthermore, the use of hybrid nanocomposites has also been investigated for other applications such as corrosion resistance. A study by found that the addition of nano-particles improved the corrosion resistance of the reinforced Al alloy cast hybrid nanocomposites, which could have potential applications in harsh corrosive environments. In terms of the AWJM process itself, research has also been done on the use of different abrasives and their effect on the machining of reinforced Al alloy cast hybrid nanocomposites [10], [11]. A study investigated the use of garnet and alumina abrasives and found that garnet abrasives resulted in lower surface roughness and improved the rate of material removal compared to alumina abrasives. This highlights the importance of the choice of abrasive in the AWJM process and its effect on the machining performance of reinforced Al alloy cast hybrid nanocomposites [12], [13].

In conclusion, the literature suggests that the use of AWJM to machine reinforced Al alloy cast hybrid nanocomposites is a promising approach, with the totalling of nano-particles improving the mechanical and metallurgical properties of the material. The optimization of the AWJM process parameters is crucial for improving the surface roughness and rate of material removal, and ANNs have been found to be a valuable device for the prediction and optimization of machining procedures. Further research is needed to fully understand the behavior of reinforced Al alloy cast hybrid nanocomposites during the AWJM process and to develop new approaches for the optimization and improvement of the machining procedure.

II. METHODOLOGY

A. Material Preparation

The process of manufacturing the reinforced AI alloy cast hybrid nanocomposite begins with heating the aluminium alloy 6081 in an electric crucible. The temperature is maintained at 750°C for 40 minutes to ensure complete melting of the alloy. Once the alloy is melted, 0.5% of Al2O3 and 0.5 % SiC nanoparticles are added to the mixture. Stirring is initiated and continued for 10 minutes to ensure proper dispersion of the nanoparticles in the melted alloy. After this, the melted Al 6081 alloy with reinforcement is subjected to the ultrasonic vibrator to further enhance the dispersion of Al2O3 and SiC nanoparticles. This ensures that the nanoparticles are uniformly distributed throughout the melted alloy. The resulting semi-liquid mixture is then poured into a die with dimensions of 100mm x 150mm x 25mm and allowed to cool in atmospheric conditions.

The arrangement of the Al 6081 alloy used in this investigation is shown in Table 1. It is important to note that the composition of the alloy and the addition of nanoparticles greatly influence the mechanical and metallurgical behavior of the final nanocomposite product. The combination of Al2O3 and SiC nanoparticles in the present study is aimed at enhancing the hardness and strength of the final product, making it more suitable for applications that require high resistance to wear and tear. In summary, the manufacturing process of the reinforced AI alloy cast hybrid nanocomposite involves heating the aluminium alloy 6081, adding Al2O3 and SiC nanoparticles, stirring the mixture, subjecting it to ultrasonic vibration, and finally, pouring the semi-liquid mixture into a die and allowing it to cool in atmospheric conditions. The final product has the potential to exhibit improved mechanical and metallurgical properties due to the presence of Al2O3 and SiC nanoparticles.

TABLE I. EXPERIMENTAL OUTCOME

Elements	Percentage (wt%)
Al	96.00
Mg	0.80
Si	0.70
Fe	0.70
Mn	0.50
Cr	0.20
Zn	0.25
Cu	0.10
T	0.10
Others	0.35

B. Design of experiments

The Al6081 alloy was subject to abrasive water jet machining and enhanced with the addition of 0.5% SiC and 0.5% Al2O3 nanoparticles. To carry out the experiment, Granet 80 mesh abrasive material was utilized. The nozzle used was unthreaded and set at a 90-degree angle to ensure a smooth delivery of the abrasive through water as the medium. The Taguchi L9 orthogonal array was applied in the study, with St, Fr, and Pw as the input parameters. The results of the experiment were evaluated based on the surface roughness and material removal rate as the output parameters. Table 2 provides details on the various process parameters that were adjusted during the experiment, and Table 3 showcases the project of the L9 orthogonal array. A 15*15 mm incision was created on the final composite material to conduct the experiment.

TABLE II. EXPERIMENTAL PARAMETERS

Level	Pw (Bar)	Fr(gm/min)	St (mm/min)
1	2500	250	100
2	3000	300	200
3	3500	350	250

The utilization of abrasive water jet machining on Al6081 alloy has proven to be a successful method for enhancing its properties through reinforcement with nanoparticles. The addition of SiC and Al2O3 has increased the strength and durability of the alloy, making it ideal for use in various applications where high levels of performance are required. The Taguchi L9 orthogonal array has provided a systematic approach to the optimization of the machining process, ensuring that the results are precise and accurate. The utilization of water as the medium for the abrasive has allowed for a smoother and more efficient machining process.

The combination of abrasive water jet machining and reinforcement with SiC and Al2O3 has resulted in a high-

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:28:26 UTC from IEEE Xplore. Restrictions apply.

performance Al6081 alloy that is ideal for use in various applications. The use of the Taguchi L9 orthogonal array and the precise adjustment of the process parameters has ensured that the machining process is optimized for the best results. The utilization of water as the medium for the abrasive has also contributed to the success of the experiment, providing a smoother and more efficient machining process.

TABLE III. DESIGN OF THE EXPERIMENT USED FOR ABRASIVE MACHINING

L9 array	Pw (Bar)	Fr (gm/min)	St (mm/min)
1	2500	250	100
2	2500	300	200
3	2500	350	250
4	3000	250	200
5	3000	300	250
6	3000	350	100
7	3500	250	250
8	3500	300	100
9	3500	350	200

C. Signal to noise ratio

In the field of manufacturing, optimizing process parameters plays a crucial role in achieving the desired output characteristics. In the case of abrasive water jet machining of reinforced Al6081 alloy, the responses are the two important response characteristics that require optimization. In this scenario, the Taguchi method is employed for optimizing the process parameters.

In the Taguchi method, the signal-to-noise (S/N) ratio is used to measure the quality of the output characteristics. The S/N ratio provides a quantitative measure of how well the output characteristics are aligned with the desired characteristics. The smaller the S/N relation, the better the excellence of the output. In the case of surface roughness, a smaller value is desired, and thus, a smaller-is-better (SIB) S/N ratio is used. On the other hand, for the material removal rate, a larger value is desired, and thus, a larger-is-better (LIB) S/N ratio is used.

The formula for S/N ratio in the Taguchi method is given by:

For SIB: S/N = 10log(1/mean square deviation)

For LIB: S/N = 10log(mean square deviation)

Here, the mean square deviation is calculated as the average of the squares of the deviations between the measured response and the target value. The target value is the desired characteristic for a particular output.

The S/N ratio provides a relative amount of the quality of the output characteristics. The best mixture of input limits can be obtained by finding the maximum value of S/N ratio for the respective response characteristics. Once the optimal combination is found, the procedure parameters can be adjusted accordingly to achieve the desired output characteristics.

The optimization of parameters through the use of the Taguchi technique and the S/N ratio provides a systematic and efficient approach to achieve the desired output characteristics. By using the S/N ratio, it is possible to quantify the quality of the output and make adjustments to the process parameters until the desired output is achieved. The S/N ratio is an essential tool for optimizing process parameters and achieving the desired output in abrasive water jet machining of reinforced Al6081 alloys.

III. RESULT AND DISCUSSION

The experiment is conducted as per the parameter shown in table III. The following results are obtained as shown in the table IV.

TABLE IV. RESULTS OF THE EXPERIMENT

Pw (Bar)	Fr (gm/min)	St (mm/min)	RM (gm/min)	Rs (µm)
2500	250	100	30.32	0.82
2500	300	200	24.23	1.32
2500	350	250	31.57	2.92
3000	250	200	27.47	0.95
3000	300	250	31.76	4.22
3000	350	100	34.23	4.65
3500	250	250	35.45	6.64
3500	300	100	33.94	8.18
3500	350	200	30.88	9.22

The result of the analysis obtained from the MINITAB software is depicted in Figure 1. The aim of this investigation was to control the optimal parameters for diminishing the response surface roughness (RS). The MINITAB software was used to perform this analysis and the optimal parameters were identified and displayed in Figure 1. The analysis revealed that the optimal condition of parameters for minimum surface roughness was achieved with a Pw of 2500 bar, Fr of 250 gm/min, and St of 200 mm/min. The results of this analysis were used to optimize the machining process and achieve a desired level of surface roughness.

It is worth mentioning that the optimization of surface roughness is an important aspect of the abrasive water jet machining process. A smoother surface finish can improve the performance and overall quality of the machined part. In addition, it can also enhance the aesthetic appearance of the part, making it more desirable for certain applications. To obtain the optimal parameters for the abrasive water jet machining process, several trials were conducted using different combinations of process parameters. The MINITAB software was used to analyze the results and determine the optimal combination of parameters for minimum surface roughness.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:28:26 UTC from IEEE Xplore. Restrictions apply.

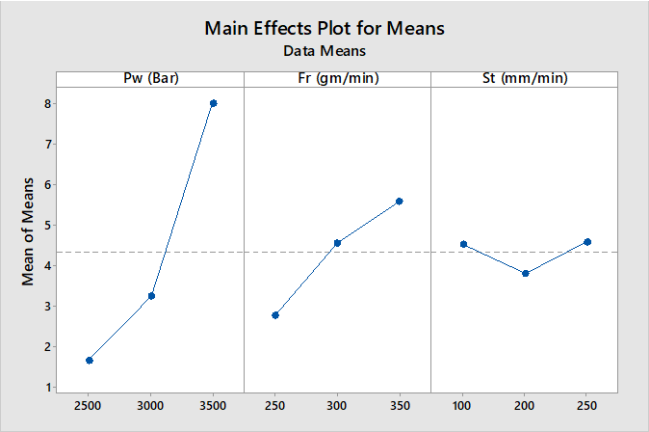


Fig. 1. Optimal condition for minimising response

In addition to minimizing surface roughness, the other key objective of the abrasive water jet machining experiment was to maximize the material removal rate (RM). Figure 2 also displays the results obtained from the analysis done using MINITAB software for this objective. The examination was conducted to identify the optimal parameters that lead to maximum RM. As seen in the figure, the optimal combination for maximizing RM was found to be Pw = 3500 bar, Fr = 350 gm/min and St = 250 mm/min. These parameters were the result of the systematic optimization process, which involved the application of the Taguchi L9 orthogonal array and the evaluation of the performance of the abrasive water jet machining process based on the signal-to-noise ratio method.

This result highlights the significance of each of the process parameters in determining the removal rate of material. It can be observed that a Pw =3500 bar of water was required for maximum RM, along with a Fr = 350 gm/min and a St = 250 mm/min. This information could be useful in the development of future abrasive water jet machining processes for similar materials. Furthermore, it demonstrates the potential of this optimization methodology in finding the optimal conditions for maximizing different performance criteria in abrasive water jet machining processes.

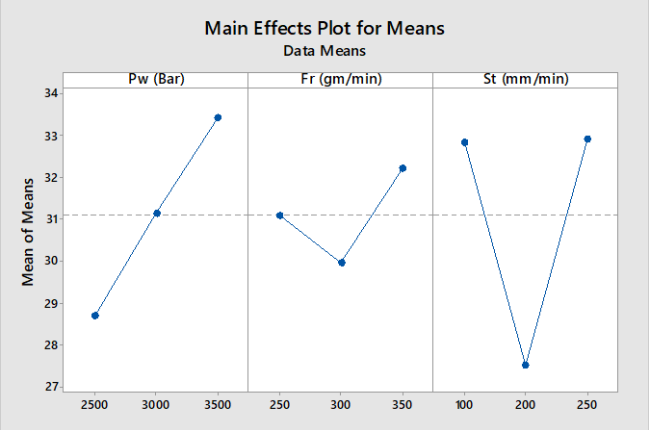


Fig. 2. Optimal condition for maximising response

The contour plots of material removal rate (RM) as shown in Figures 3, 4, and 5 provide a visual representation of the relationship among the input parameters and the resulting material removal rate. The contour plots serve as a confirmation of the optimal grouping of parameters obtained from the signal-to-noise (S/N) ratio analysis. The results from the contour plot clearly demonstrate that an increase in the input parameters directly leads to a higher rate of material removal. This information is valuable for industrial professionals who are looking to machine aluminum nano composites efficiently and effectively. With this knowledge, they can adjust the parameters to achieve maximum material removal, which can save time and increase production speed. The contour plots provide a visual understanding of the correlation between the input parameters and the resulting material removal rate, making it easier for industrial professionals to make informed decisions.

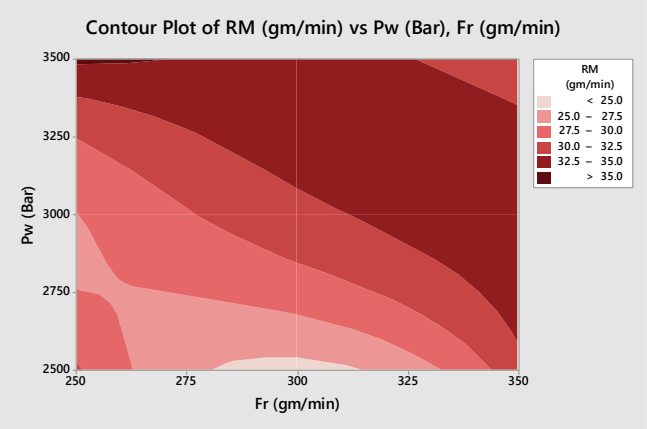


Fig.3. Contour plot of RM with Pw and Fr

Furthermore, the contour plots provide a comprehensive representation of the surface, allowing for an understanding of how the input parameters affect the material removal rate across different areas of the surface. This information helps industrial professionals to make informed decisions on how to optimize the process to achieve maximum material removal rate, without sacrificing the quality of the surface.The contour plots of material removal rate as shown in Figures 3, 4, and 5 provide important perceptions into the association among the input parameters and the resulting material removal rate. The information obtained from these contour plots can be used to make informed decisions and optimize the process to achieve maximum material removal rate without sacrificing the quality of the surface. This information is critical for industrial professionals who are looking to machine aluminum nano composites efficiently and effectively.

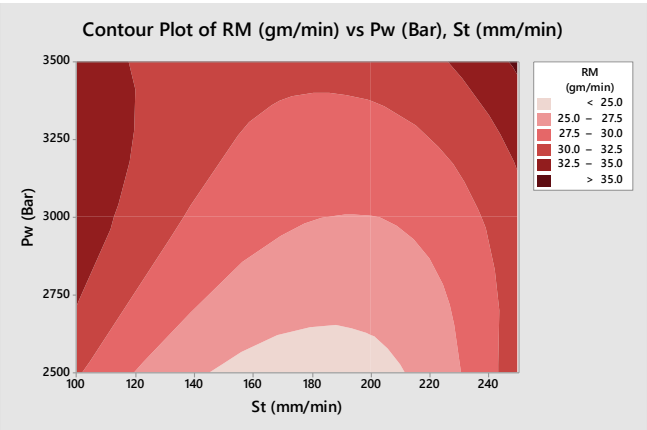


Fig.4. Contour plot of RM with Pw and St

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:28:26 UTC from IEEE Xplore. Restrictions apply.

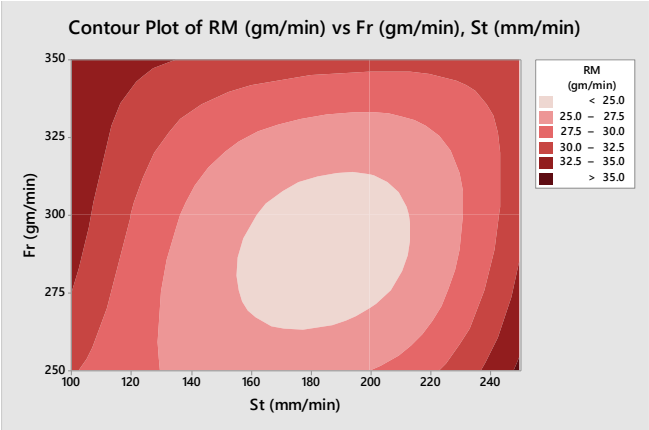


Fig. 5. Contour plot of RM with Pw and Fr

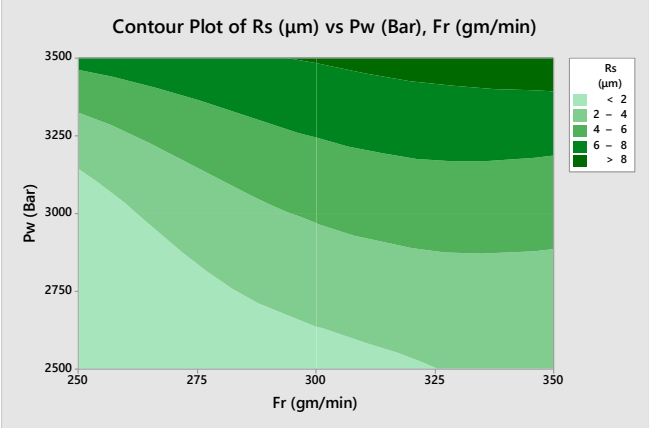


Fig. 6. Contour plot of Rs with Pw and Fr

The contour plots of Figure 6, 7, and 8 provide a graphical representation of the results obtained from the signal-to-noise ratio analysis for the minimum surface roughness (Rs) of the Al6081 nano-composite material. The plots show that the optimal combination of process parameters, such as Pw, Fr, and St, results in the lowest level of surface roughness. This information is critical for industrial applications, as lower surface roughness generally results in better material properties, such as improved strength and corrosion resistance. The results obtained from the analysis are similar to those obtained in previous studies, which confirms the accuracy of the experiment. The contour plots show that the higher the values of Pw, Fr, and St, the lower the level of surface roughness. This pattern has been observed in many other studies, which further supports the validity of the results obtained in this study. In industrial applications, it is vital to obtain the optimal mixture of process limits that results in the lowest possible surface roughness. The information obtained from the contour plots can be used by engineers and technicians to adjust the process parameters accordingly, thus ensuring that the Al6081 nano-composite material is machined to the desired level of roughness.

Furthermore, the information obtained from the analysis can be used to optimize the material's surface roughness for specific applications. For example, in applications where low surface roughness is critical, such as in the aerospace and biomedical industries, the best grouping of process parameters can be used to ensure that the Al6081 nano-composite material meets the desired level of roughness. In conclusion, the results obtained from the contour plots of Figure 6, 7, and 8 provide valuable information for optimizing the surface roughness of the Al6081 nano-composite material. The results confirm the accuracy of the experiment and provide a roadmap for industrial applications. The information obtained from the analysis can be used to adjust the process parameters and obtain the optimal combination for minimum surface roughness.

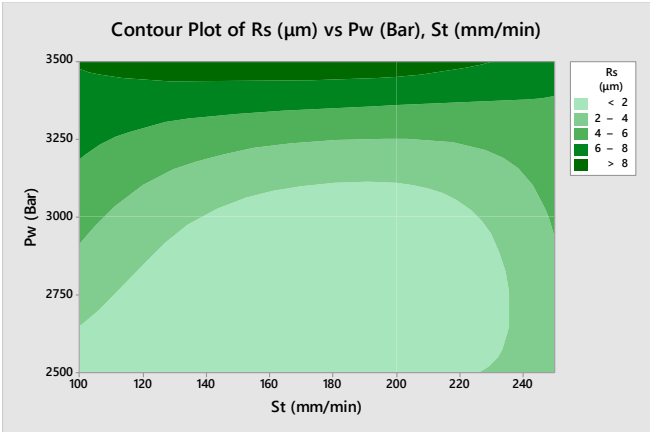


Fig. 7. Contour plot of Rs with Pw and St

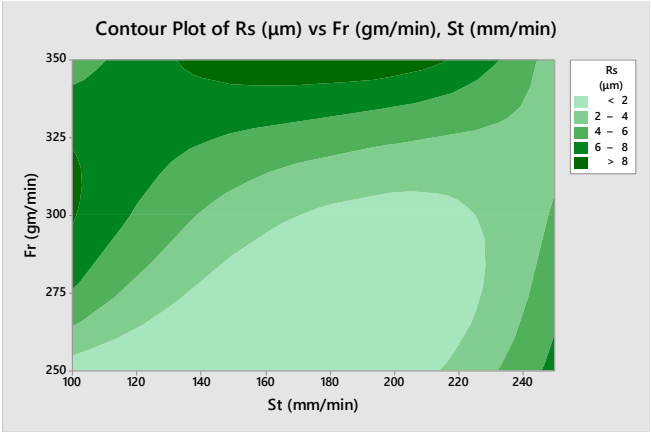


Fig.8. Contour plot of Rs with Fr and St

The figures 3, 4, and 5 demonstrate that the incorporation of nano gold nanoparticles (GNP) into a composite material leads to a reduction in the friction coefficient (Cf). The most effective wear resistance characteristics were observed at a composition of 0.9%. It was also found that an increase in the applied load resulted in an increase in friction for the prepared composite. The results of the experiment suggest that the friction coefficient does not significantly change with changes in sliding distance. Overall, it can be concluded that a 0.9% composition of nano GNP effectively reduces friction and improves wear resistance properties.

ANN is a influential machine learning tool that has been widely used to predict and model complex associations amid input and output strictures in various applications, including industrial processes and materials engineering. The main advantage of ANNs is that they can handle non-linear and complex relationships, making them ideal for predicting results in experiments where there are multiple interacting factors. In this study, ANNs are used to predict the experimental results of the Al6081 nano-composite machining process. The inputs to the ANN model are the process parameters and the outputs are the surface roughness and material removal rate. To train the ANN model, the experimental data collected from the machining process is

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:28:26 UTC from IEEE Xplore. Restrictions apply.

used. The data is divided into two parts, with 80% used for training and 20% for testing. The ANN model is designed using a feed-forward neural network with multiple hidden layers and activation functions. The activation function is used to introduce non-linearities into the model and allow it to capture complex relationships between the inputs and outputs. The number of hidden layers and the number of neurons in each layer are determined through trial and error. Once the ANN model is trained, it is tested on the 20% of the data set that was set aside for testing.

The performance of the ANN model is evaluated using various performance metrics, such as mean squared error, root mean squared error, and coefficient of determination. These metrics are used to quantify the accuracy of the model in predicting the experimental results. The results of the ANN model show that it can accurately predict the experimental results of the Al6081 nano-composite machining process. The mean squared error and root mean squared error are both low, indicating that the model has a good fit to the data. The coefficient of determination is also high, indicating that the model can explain a large proportion of the variance in the data. The results of the ANN model are also compared to those obtained from other regression models, such as multiple linear regression and support vector regression. The results show that the ANN model outperforms these other models in terms of accuracy and ability to handle complex relationships between the inputs and outputs.

In conclusion, the use of ANNs in this study has shown that they can be used to accurately predict the experimental results of the Al6081 nano-composite machining process. The ability of ANNs to handle non-linear and complex relationships makes them ideal for predicting results in experiments where there are multiple interacting factors. The results of this study can be useful for industrial applications, where accurate predictions of machining results are important for improving processes and reducing costs. Figure 9 illustrates the result of ANN with the prediction of result at 100 percentage accuracy.

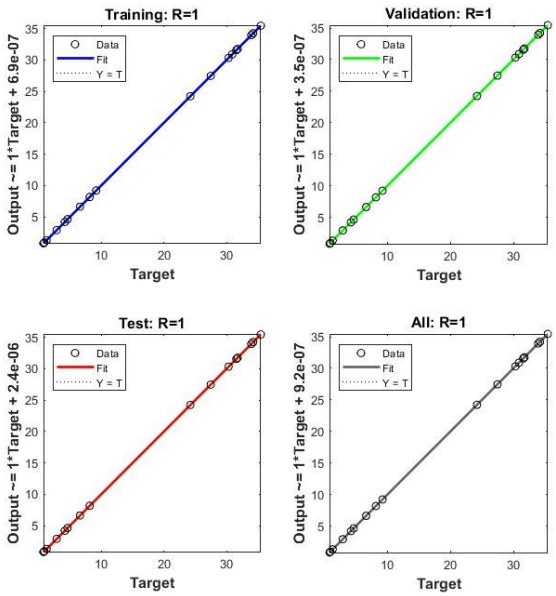


Fig.9. Result of ANN

IV. CONCLUSION

In conclusion, the research article focuses on the modelling and manufacturing of abrasive water jet machining of reinforced aluminum alloy cast hybrid nanocomposites. The study highlights the significance of understanding the mechanical and metallurgical behavior of the nanocomposites to predict the effectiveness of abrasive water jet machining. The researchers applied Artificial Neural Network (ANN) for this purpose and were able to achieve accurate predictions of the machining parameters. This result highlights the importance of using advanced analytical tools such as ANN for forecasting the efficacy of machining processes and optimizing the manufacturing process. The study also provides a useful contribution to the field of advanced materials and machining by demonstrating the potential of hybrid nanocomposites in various industrial applications. Overall, the study highlights the potential of abrasive water jet machining in the production of high-quality nanocomposite components and the importance of understanding the mechanical and metallurgical behavior of these materials to ensure optimal performance.

REFERENCES

[1] C. J. Nicholls, B. Boswell, I. J. Davies, and M. N. Islam, "Review of machining metal matrix composites," *International Journal of Advanced Manufacturing Technology*, vol. 90, no. 9–12, pp. 2429–2441, 2017, doi: 10.1007/s00170-016-9558-4.

[2] Z. W. Zhong and G. Lin, "Ultrasonic assisted turning of an aluminium-based metal matrix composite reinforced with SiC particles," *International Journal of Advanced Manufacturing Technology*, vol. 27, no. 11–12, pp. 1077–1081, 2006, doi: 10.1007/s00170-004-2320-3.

[3] P. Kishore, P. M. Kumar, and D. Dinesh, "Wear analysis of Al 5052 alloy with varying percentage of tungsten carbide," *AIP Conference Proceedings*, vol. 2128, no. 1, p. 040003, Jul. 2019, doi: 10.1063/1.5117965.

[4] M. Guden and I. W. Hall, "Quasi-static and dynamic compression behaviour of an FP™ alumina-reinforced aluminium metal matrix composite," *Journal of Materials Science*, vol. 33, no. 13. pp. 3285–3291, 1998, doi: 10.1023/A:1013272910939.

[5] T. M. Yue, Y. X. Wu, and H. C. Man, "On the role of CuAl2 precipitates in pitting corrosion of aluminum 2009/SiCW metal matrix composite," *Journal of Materials Science Letters*, vol. 19, no. 11, pp. 1003–1006, 2000, doi: 10.1023/A:1006745028307.

[6] B. Vishnu Vardhana Naidu, M. Natarajan, G. M. Surendranatha, and K. C. Varaprasad, "Fabrication of aluminium hybrid metal matrix composites and development multiple regression models for wire electrical discharge machining," *International Journal on Interactive Design and Manufacturing*, 2022, doi: 10.1007/s12008-022-00920-2.

[7] N. R. J. Hynes and R. Kumar, "Electrochemical Machining of Aluminium Metal Matrix Composites," *Surface Engineering and Applied Electrochemistry*, vol. 54, no. 4, pp. 367–373, 2018, doi: 10.3103/S1068375518040087.

[8] A. Hemalatha, V. D. Reddy, and K. Prasad, "Evolution of regression and ANFIS models for wire spark erosion machining of aluminium metal matrix composites for aerospace applications," *International Journal on Interactive Design and Manufacturing*, 2022, doi: 10.1007/s12008-022-01012-x.

[9] A. H. V Diwakar and R. S. Hemachandra, "Development of empirical models and machinability comparison on wire electrical discharge machining of aluminium based hybrid and metal matrix composites," 2023, doi: 10.1007/s12008-023-01200-3.

[10] P. Kishore, A. K. Dash, A. Pragallapati, D. Mugunthan, A. Ramesh, and K. D. Kumar, "A Tripod-Type Walking Assistance for the Stroke Patient BT - Congress on Intelligent Systems,"

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:28:26 UTC from IEEE Xplore. Restrictions apply.

2021, pp. 151–160.

[11] T. Panneerselvam, T. K. Kandavel, and P. Kishore, "Experimental Investigation on Cutting Tool Performance of Newly Synthesized P/M Alloy Steel Under Turning Operation," *Arabian Journal for Science and Engineering*, vol. 44, no. 6, pp. 5801–5809, Jun. 2019, doi: 10.1007/s13369-019-03763-4.

[12] T. S. Senthilkumar and R. Muralikannan, "Enhancing the geometric tolerance of aluminium hybrid metal matrix composite using EDM process," *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, vol. 41, no. 1, pp. 1–15, 2019, doi: 10.1007/s40430-018-1553-2.

[13] D. S. E. J. Dhas, C. Velmurugan, and K. L. D. Wins, "Investigations on the Effect of Tungsten Carbide and Graphite Reinforcements during Spark Erosion Machining of Aluminium Alloy (AA 5052) Hybrid Composite," *Silicon*, vol. 10, no. 6, pp. 2769–2781, 2018, doi: 10.1007/s12633-018-9817-0.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:28:26 UTC from IEEE Xplore. Restrictions apply.

Multifunctional Sm^{3+} induced bismuth lithium tungstate phosphor for flexible anti-counterfeiting and encryption

Indrajeet Maurya[1], Mukesh K. Sahu[1,2], Vikas[1], A. K. Vishwakarma[3], M. Jayasimhadri[1]*

[1] Luminescent Materials Research Lab (LMRL), Department of Applied Physics, Delhi Technological University, Delhi-110 042, India
[2]School of Computing, Graphic Era Hill University (GEHU), Haldwani, Uttarakhand, 263 136, India
[3]Department of Physics, ARSD College, University of Delhi, 110021, India

*Corresponding Author e-mail- jayaphysics@yahoo.com

Abstract: In recent times, one of the challenging issues is the counterfeiting of goods which has major implications for governments, corporations, and consumers everywhere in the world [1, 2]. Since counterfeiting is an increasingly high-tech crime, there must be high-tech solutions to stop and discourage the practice. With this motive, the current research emphasized on strategic flexible anticounter fitting. The crystalline monoclinic phase of Sm^{3+} induced bismuth lithium tungstate (BLW) phosphor has been synthesized via solid state reaction technique to examine the direct utility in anti-counterfeiting and encryption. Diffraction patterns and micrographs have confirmed the monoclinic crystalline and non-uniform micron size particles, respectively. Defuse reflectance spectra revealed the numerous absorption peaks in UV and visible range. The photoluminescence (PL) emission peak observed at 603 nm via pumping through UV/n-UV and blue light is corresponding to the $^4G_{5/2} \rightarrow {}^6H_{7/2}$ transition. CIE coordinates of BLW: Sm^{3+} phosphors are located in the orange-red region of the chromaticity diagram. Above mentioned results of BLW phosphor suggest the prolific properties of the as-synthesized phosphor for the luminescent security inks and their anti-counterfeiting applications.

References

[1] Pawan Kumar, Satbir Singh, B. K. Gupta, *Nanoscale*, 8 (2016) 14297-14340. DOI: https://doi.org/10.1039/C5NR06965C.

[2] Jia Wang, Jun Ma, Jiachi Zhang, Yu Fan, Wenxiang Wang, Jika Sang, Zhidong Ma, Huihui Li, *ACS Appl. Mater. Interfaces* 11 (2019) 35871-35878. DOI: https://doi.org/10.1021/acsami.9b10870.

Trans Indian Inst Met
https://doi.org/10.1007/s12666-023-03096-9

ORIGINAL ARTICLE

Multi-response Optimization of Friction Stir Welded Reinforced Joints of Dissimilar Aluminum Alloys

Sumit Jain[1] · **R. S. Mishra**[1]

Received: 5 March 2023 / Accepted: 27 August 2023
© The Indian Institute of Metals - IIM 2023

Abstract In the present research, an attempt has been made to develop friction stir welded dissimilar reinforced joints of AA6061 and AA7075 using micro-sized Al_2O_3 particles. The influence of varying volume fractions of Al_2O_3 particles along with traverse and rotational speeds on microstructural evolution and mechanical behavior was investigated using response surface methodology. At a 95% confidence level, quadratic mathematical models were developed to predict the optimum value of output responses such as microhardness, tensile strength, and %elongation. According to the developed models, rotational speed was observed as the most influential variable, whereas traverse speed was observed as the least influential variable. The optimum values of rotational speed, traverse speed, and volume fraction of Al_2O_3 particles were found to be 971 rpm, 40 mm/min, and 10%, respectively. The optimum values of output responses, such as tensile strength, %elongation, and microhardness, were found to be 226.2 MPa, 17.4%, and 144.3 HV, respectively.

Keywords Reinforced joints · Tensile strength · %elongation · Microhardness · ANOVA

1 Introduction

Joining aluminum alloys using fusion welding techniques such as TIG welding is a tremendous challenge for researchers. Aluminum oxide, microcracks, coarse grain structure, and solidification shrinkage are the main issues that arise with welded joints of aluminum alloys, which diminish the mechanical characteristics of the joint, including poor tensile strength and low hardness [1]. To overcome these limitations of fusion weld techniques, friction stir welding (FSW) is most widely utilized these days to join aluminum and its alloys. FSW is a fascinating solid-state welding approach that can weld thin and thick plates and offers numerous advantages over other welding approaches, including slight distortion, low residual stress, and a more homogeneous grain structure [2]. FSW has been employed in shipbuilding, automobiles, and aviation parts to lighten their structural weight [3–5]. Surface and bulk composites comprised of metal matrix composites (MMCs) with micro/nanoparticles are developed by FSP to enhance their tribological and mechanical characteristics [6–8]. Several investigators have studied the distribution of reinforcing particles (RPs) by friction stir processing (FSP) [9–14].

Improved weld quality in friction stir welded (FSWed) dissimilar joints can have significant implications for various industrial applications, including aerospace, automotive, shipbuilding, etc. The improved weld quality achieved through FSW in dissimilar joints can lead to stronger, more durable, and more efficient industrial components. It can decrease the need for post-weld treatments and increase overall operating dependability, which will eventually lead to cost savings, an extension of the service life, and enhanced performance across a variety of industrial sectors [15]. FSW was also employed by several researchers to develop aluminum metal matrix (AMM) composites by incorporating various RPs, including AA6061/TiC, AA2124/SiC, AA7005/Al_2O_3, and AA7075 + AA2024/SiC and AA1350 + AA6101/Al_2O_3/SiC [16–20].

To optimize the processing variables of FSW, several researchers have used a variety of experimental design

✉ Sumit Jain
sumitjain4@gmail.com

[1] Department of Mechanical Engineering, Delhi Technological University (DTU), Delhi, India

Published online: 26 September 2023

methodologies [21–23]. RSM is a statistical approach employed to model and analyze the correlations between several input parameters and output responses like hardness, tensile strength, etc. RSM is frequently used for identifying the most optimal parameter combinations for FSW and permits the simultaneous investigation of process factors and their interactions, both of which are essential for the development of composite joints. It decreases the overall number of experiments necessary to find the optimal parameter setting [24]. RSM aids in selecting the best set of process variables for FSWed composite joints in order to fabricate defect-free joints with adequate mechanical characteristics [25].

RSM approach was employed to determine the optimal setting for input parameters to increase the corrosion resistance of dissimilar FSWed parts AA6082 and AA5052 using chromium and magnesium powder as filler materials. Response optimization reveals that the TS of 125 mm/min, filler ratio of 95% magnesium and 5% chromium, RS of 1000 rpm, and hole spacing of 2 mm were the welding parameters that resulted in the lowest corrosion rate [26, 27]. By combining B_4C and Al_2O_3 particles, aluminum hybrid surface composites were developed via FSP. The TS, axial force, RS, and reinforcement ratio were shown to be the most important process variables [28]. However, there is a paucity of studies related to the multi-optimization of the response variables by optimizing the volume fraction of RPs and FSW processing parameters. To the current study, the microhardness, ultimate tensile strength, and % elongation of reinforced joints were multi-optimized using the desirability function to determine the optimum variable settings.

The RS can have an impact on the workpiece's temperature, shear strain, and material flow. Selecting a certain RS will maximize the produced material's mechanical properties. Lower RS can reduce heat input, whereas higher RS can increase heat input and the degree of deformation. The rate of material flow and deformation during processing is governed by the TS. Lower TS can enhance the interfacial bonding and homogeneity of the composite, whereas greater TS can boost production rates [29]. The percentage of the composite material made up of the reinforcing phase is determined by the volume fractions of the RPs. The mechanical characteristics of the composite material may be best achieved by choosing a certain volume fraction [30]. In general, the selection of these variables is influenced by the specific research goals, the properties of the materials being welded, and the specific conditions of the processing equipment. The optimization of these parameters may improve the weld quality of dissimilar reinforced joint of AA6061 and AA7075.

Al_2O_3 particles are most widely used as reinforcement candidate in MMCs fabrication. Their use is expanding as a result of their low melting point temperature, great wettability, and cheaper cost. Al_2O_3 was used for improving mechanical and wear characteristics by numerous researchers while creating surface composites (SCs) using FSP [31, 32]. A lower content of Al_2O_3 might lead to minimal changes in mechanical properties. Excessive Al_2O_3 content can lead to increased hardness but may also increase the possibilities of cracks and defects within the joint, which reduce the joint's strength. Therefore, an optimal content range should be carefully determined based on the desired properties and the requirements of the specific application, while considering factors like hardness, strength, wear resistance, and brittleness. AA6061 and AA7075 are most widely utilized in various structural components for the automobile, aerospace, and marine sectors [33, 34]. The improved weld quality of these alloys would enhance the versatility of structural designs and broaden the range of their applications [35]. Thus, in the present work, the influences of TS and RS on the distribution pattern of RPs were examined. Furthermore, the impact of distribution pattern of Al_2O_3 microparticles on the mechanical characteristics of reinforced joints of AA6061 and AA7075 was examined. As per the author's knowledge, no previous research has been carried out on the optimization of the volume fraction of Al_2O_3 microparticles, RS, and TS for considered responses of dissimilar reinforced joints of AA6061 and AA7075.

2 Material and Methodology

Plates of AA6061 and AA7075 of size 150×40 x 6 mm were utilized for developing Al_2O_3mp incorporated FSWed reinforced joints. To incorporate the Al_2O_3mp in the stir zone as per the design of experiment, grooves of depth—3 mm and different widths of 0.24, 0.42, and 0.60 mm, were milled on the adjoining faces of the base plates, as depicted in Fig. 1a. Equation No. 1 was used to compute the volume percentage of RPs at 4%:

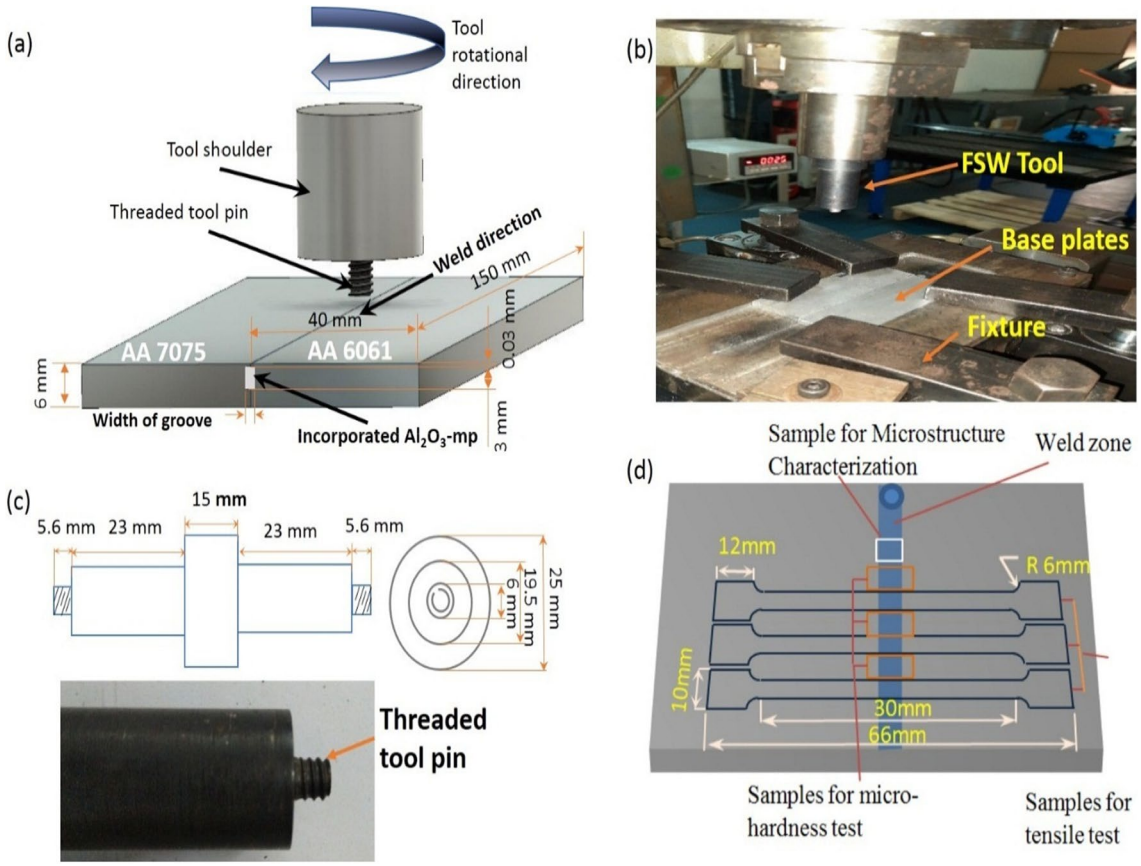


Fig. 1 **a** Systematic representation of dissimilar FSWed joint; **b** FSW experimentation; **c**, Dimensions of the tool used; **d** Dimensions of several test samples

$$\text{Volume fraction of RPs (\%)} = \frac{\text{Area of the groove manufactured}}{\text{Area of the pin}} \times 100$$
$$= \frac{\text{width of groove} \times \text{depth of groove}}{\text{diameter of pin} \times \text{length of pin upto the groove depth}} \times 100$$
$$= \frac{0.24 \times 3}{6 \times 3} \times 100 = 4\% \quad (1)$$

Before incorporating Al_2O_3mp, the milled grooves were adequately cleaned using acetone. For experimentation, numerical controlled friction stir welding (Fig. 1b) was used. A slurry was produced by the mixture of Al_2O_3mp and ethanol to embed the Al_2O_3mp adequately in the grooves. The slurry was then manually filled into the grooves and allowed to dry [36, 37]. In the current study, base plates of AA7075 were clamped on the retreating side, and a tilt angle of 2° was employed, as per the recommendations from previous studies [38, 39]. On the basis of preliminary tests and the capabilities of the FSW machine configuration, the levels of process parameters were chosen.

RSM is a statistical approach for optimizing the influence of independent variables on dependent variables. According to RSM's center composite design (CCD) strategy, a series of experiments were conducted [40]. In total, 20 trial runs were performed as suggested by the face-centered full-factorial CCD approach, as illustrated in Table 1. The values of the output responses of FSWed

Table 1 Design of experimentation and output responses of FSWed reinforced joints

Specimen no	RS (rpm)	TS (mm/min)	VF-Al_2O_3mp (%)	Mean tensile strength (MPa)	Strain (%)	Microhardness (HV)
1	700	40	10	185.5	14.9	122.2
2	700	60	10	169.5	14.7	127.5
3	700	60	4	172.0	11.9	106.7
4	700	40	4	178.6	13.6	110.9
5	700	50	7	177.5	13.8	115.5
6	900	50	7	204.1	15.7	128.2
7	900	50	7	205.5	15.1	126.7
8	900	40	7	209.2	16.1	134.4
9	900	50	7	203.7	15.8	128.8
10	900	50	7	201.5	15.7	125.5
11	900	50	10	208.1	16.3	133.0
12	900	50	7	204.4	15.1	126.4
13	900	60	7	201.7	15.3	131.2
14	900	50	7	202.6	15.1	126.6
15	900	50	4	194.6	13.1	113.4
16	1100	50	7	222.6	18.3	142.6
17	1100	40	4	214.6	16.7	140.7
18	1100	60	10	231.4	20.3	153.8
19	1100	40	10	241.4	19.4	157.5
20	1100	60	4	210.6	16.2	130.5

reinforced joints are also tabulated in Table 1. Batool et al. [41] revealed that threaded pin-profiled cylindrical tool produced dissimilar defect-free joints of AA7075 and AA6061 with higher strength compared to tapered cylindrical and straight cylindrical tools using FSW. As a result, the various composite joints were produced using a threaded pin-profiled H13 tool made of tool steel under various parametric settings. Figure 1c illustrates the dimensions of the tool used in this work.

Three test samples for various tests were taken from each reinforced joint, as specified in Fig. 1d, and the average of the three results was listed. A Vickers microhardness tester was used to assess the microhardness profile in the weld zone. The indents used for measuring microhardness employed a dwell period of 15 s and a load of 100 g. The tensile samples were extruded in accordance with the ASTM-E8 standard, as depicted in Fig. 1d. For microstructural examination, the samples were polished using various emery sheets of grades 200 to 2000. After that, Keller's reagent was applied to etch the polished samples for 10 s. A scanning electron microscope (SEM) with an energy-dispersive X-ray spectroscopy (EDX) detector was used to evaluate the dispersion pattern of Al_2O_3mp in the stir zone, while an optical microscope (OM) was used to study the evolution of grain structure. The micro-sized Al_2O_3 particles (50 to 60 µm) were used as reinforcing material. Figure 2 (a, b) lists the SEM image of Al_2O_3mp and the EDX analysis. Figure 2 (c–f) illustrates the OM micrograph and the elemental composition of AA7075 and AA6061 as determined by EDX analysis.

3 Results and Discussion

3.1 Effect on Tensile Properties

In accordance with Fig. 2(c) and (e), the average grain size of the as-received AA6061 and AA7075 was found to be 36 and 47 µm, respectively. The grain size was evaluated using ImageJ software. The base material (BM) deforms significantly during FSW via dynamic recrystallization (DRX), causing a succession of refined grains to develop inside the stir zone, as depicted in Fig. 2g. Three distinct zones are visible in the FSWed joints: the thermo-mechanically affected zone (TMAZ), the heat-affected zone (HAZ), and the stir zone (SZ) [42, 43]. The grain size, size, and distribution pattern of RPs, and the bond strength between the AMM and the RPs are the key factors impacting the mechanical properties the reinforced joints [44, 45]. The base materials AA7075 and AA6061 demonstrated tensile strength of 495 and 288 MPa, respectively, with corresponding %elongation of 21.6 and 18.4%. The tensile characteristics of all the reinforced joints are listed in Table 1. The FSWed reinforced joints fabricated at RS of 700 rpm revealed the clustering of Al_2O_3mp and poor bonding between Al_2O_3mp and the AMM owing to insufficient material mixing [46]. Figure 3a

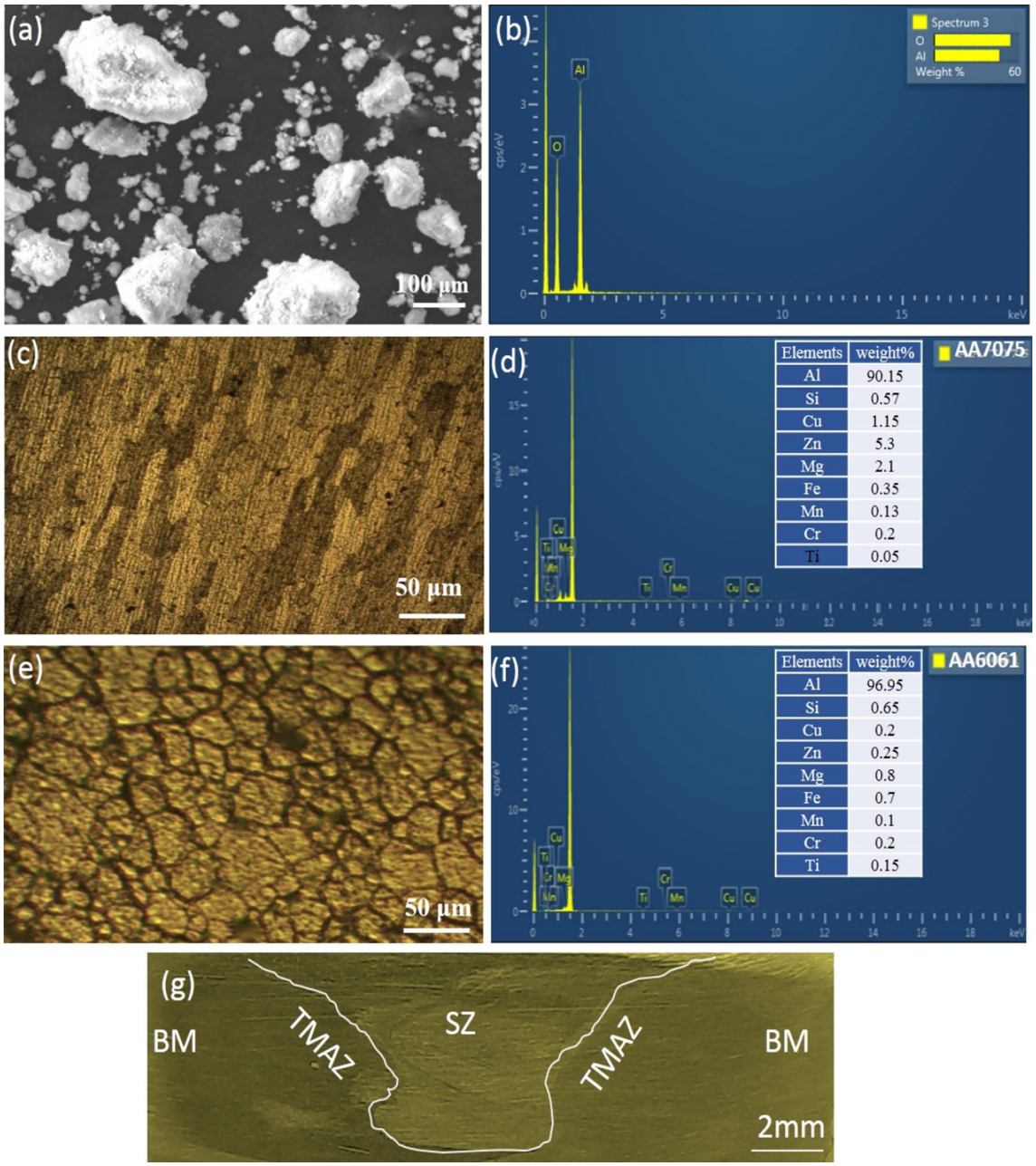


Fig. 2 **a** SEM micrograph of Al_2O_3mp, **b** EDX analysis of Al_2O_3mp; **c–f** Grain structure and elemental composition; **c**, **d** AA 7075; **e**, **f** AA6061; **g** Weld bead geometry of specimen no-19

delineates the SEM micrograph of the reinforced joint (specimen no-2) fabricated at 700 rpm, 60 mm/min, and 10% VF-Al_2O_3mp. This micrograph clearly attests to the clustering of Al_2O_3mp (Fig. 3a). The average grain size of 18.4 µm was found in the SZ (Fig. 3b).

Homogeneously dispersed Al_2O_3mp and adequate material mixing were observed due to significant strain developed at an enhanced RS of 1100 rpm. The FSWed reinforced joint (specimen no-19) fabricated at 1100 rpm, 40 mm/min, and 10% VF-Al_2O_3mp revealed the uniform dispersion of Al_2O_3mp due to effective material mixing, as cited in Fig. 3c. The absence of clustering of Al_2O_3mp indicates excellent bonding between AMM and Al_2O_3mp. The mobility of RPs got increased by higher values of RS [47]. The reduced grain structure of size 7.2 µm in the SZ of specimen no-19 (Fig. 3d) can be attributed to the DRX and the pinning effect of homogeneously dispersed Al_2O_3mp. The Al_2O_3 particles' pinning action improves the nucleation sites and inhibits grain development [48]. Grain refinement was increased by the enhanced pinning effect with an increased

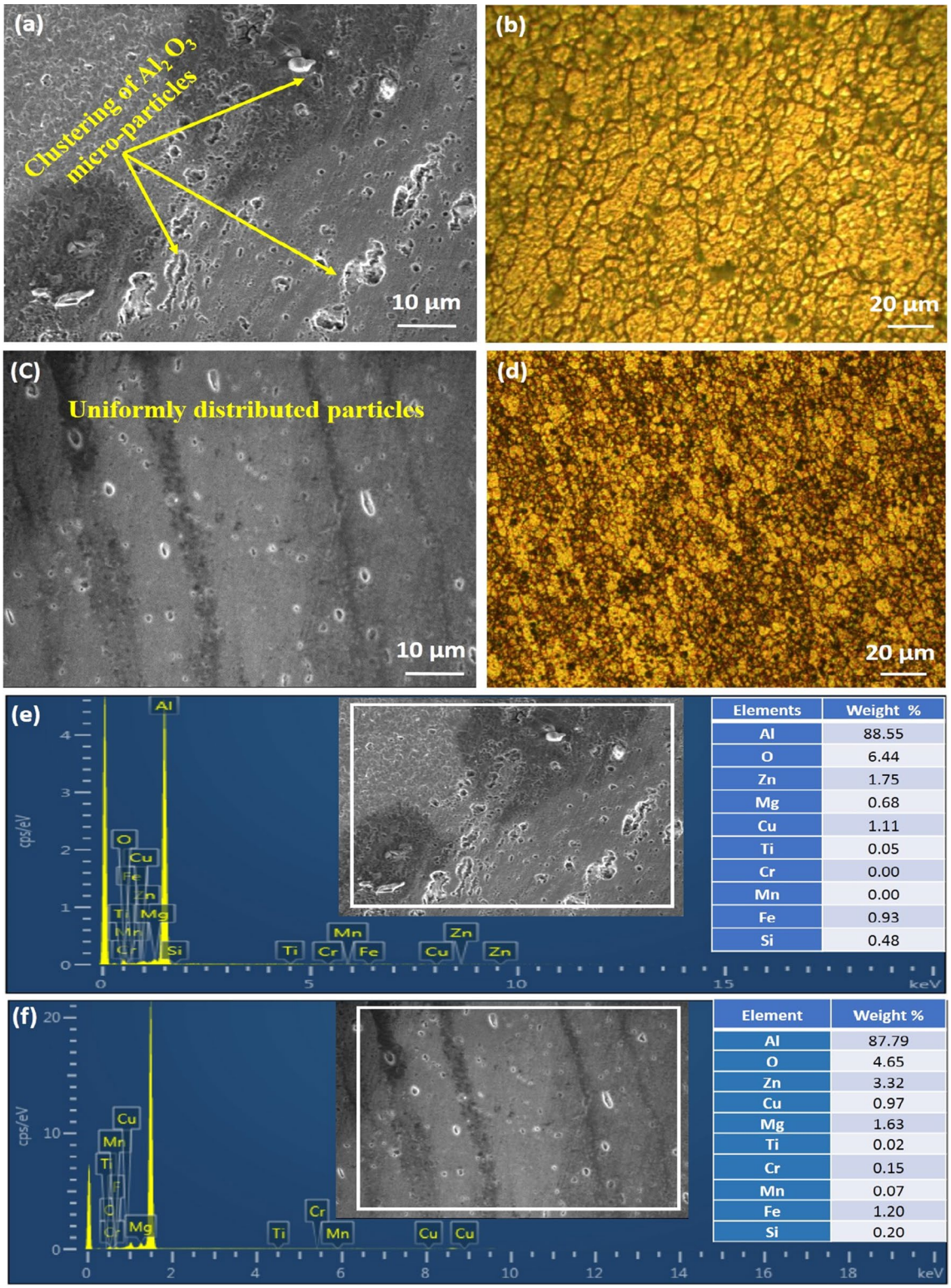


Fig. 3 **a**–**f** SEM, OM micrograph and EDX analysis; **a**, **b**, **e** specimen no-2, **c**, **d**, **f** specimen no-19

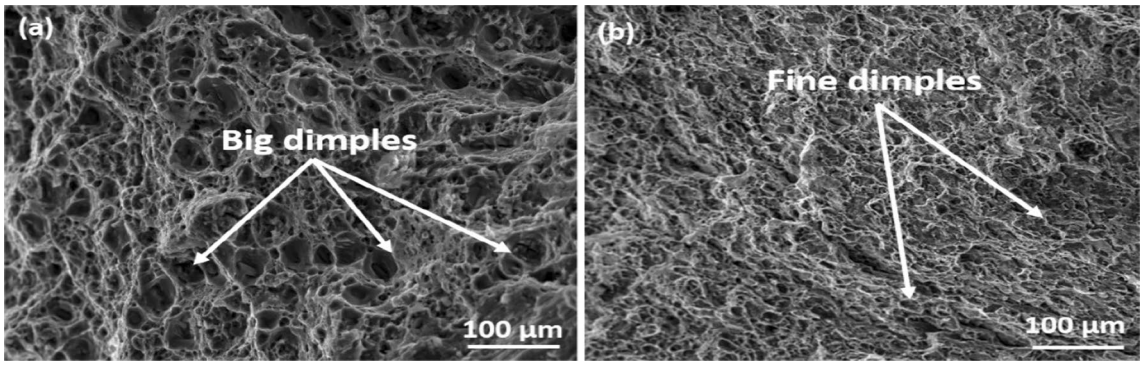


Fig. 4 Fractography of fractured samples; **a** specimen no-2; **b** specimen no-19

content of more uniformly distributed Al_2O_3mp [49]. The Hall–Petch equation indicates that tensile strength increases as grain size decreases [50, 51]. Consequently, specimen no-19 exhibited the highest tensile strength, with a value of 241.1 MPa. In contrast, specimen no-2 exhibited a minimal tensile strength of 169.5 MPa. The FSWed reinforced joint (specimen no-3) fabricated at 700 rpm, 60 mm/min, and 4% VF-Al_2O_3mp revealed a minimum %elongation of 11.9%, which was lower than both the base materials. However, specimen no-18 fabricated at 1100 rpm, 60 mm/min, and 10% VF-Al_2O_3mp revealed a maximum %elongation of 20.3%, which was higher than that of AA6061 but lower than that of AA7075. The EDX evaluation of the SZ of specimen no-2 and specimen no-19 verified the presence of Al_2O_3mp. Figure 3e and f cites the peaks of oxygen and confirms the presence of Al_2O_3mp. The EDX analysis of specimen no-2 and specimen no-19 revealed the oxygen percentages of 6.44% and 4.65%, respectively. In specimen

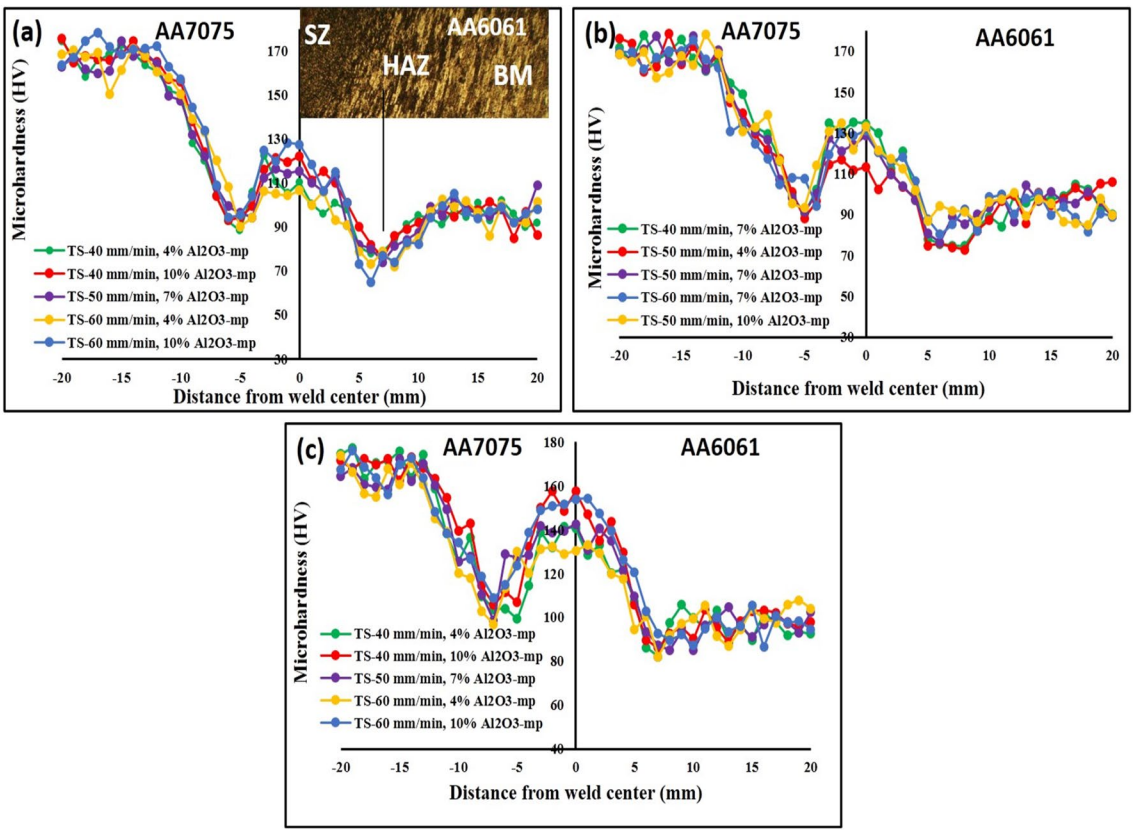


Fig. 5 Microhardness variation of the FSWed reinforced joints; **a** 700 rpm, **b** 900 rpm, **c** 1100 rpm

Table 2 Results of ANOVA for tensile strength

Sources	SS	DF	MS	F value	P value	Remark
Model	6603.41	9	733.71	445.35	< 0.0001	Significant
RS	5640.62	1	5640.62	3423.73	< 0.0001	
TS	194.48	1	194.48	118.05	< 0.0001	
VF-Al_2O_3mp	429.02	1	429.02	260.41	< 0.0001	
RS*TS	9.25	1	9.25	5.61	0.0394	
RS* VF-Al_2O_3mp	233.28	1	233.28	141.60	< 0.0001	
TS* VF-Al_2O_3mp	29.64	1	29.64	17.99	0.0017	
RS^2	29.37	1	29.37	17.83	0.0018	
TS^2	12.50	1	12.50	7.59	0.0203	
VF-$Al_2O_3$$mp^2$	10.65	1	10.65	6.47	0.0292	
Residual	16.48	10	1.65			
Lack of fit	6.56	5	1.31	0.66	0.6691	Not significant
Pure-Error	9.91	5	1.98			
Cor Total	6619.89	19				
C.V. %	1.28				R^2	0.9976
Mean	201.95				Adjusted R^2	0.9914
Std. Dev	0.64				Predicted R^2	0.9872
					Adeq. Precision	79.08

Table 3 Results of ANOVA for %elongation

Sources	SS	DF	MS	F value	P value	Remark
Model	74.62	9	8.29	99.83	< 0.0001	Significant
RS	48.40	1	48.40	582.77	< 0.0001	
TS	0.53	1	0.53	6.37	0.0302	
VF-Al_2O_3mp	19.88	1	19.88	239.38	< 0.0001	
RS*TS	0.66	1	0.66	7.96	0.0181	
RS* VF-Al_2O_3mp	0.91	1	0.91	10.97	0.0078	
TS* VF-Al_2O_3mp	1.05	1	1.05	12.66	0.0052	
RS^2	1.58	1	1.58	19.08	0.0014	
TS^2	0.46	1	0.46	5.54	0.0404	
VF-$Al_2O_3$$mp^2$	0.96	1	0.96	11.56	0.0068	
Residual	0.83	10	0.08			
Lack of fit	0.22	5	0.04	0.37	0.8534	Not significant
Pure-Error	0.61	5	0.12			
Cor Total	75.45	19				
C.V. %	1.84				R^2	0.9891
Mean	15.65				Adjusted R^2	0.9791
Std. Dev	0.288				Predicted R^2	0.9690
					Adeq. Precision	41.80

no-2, the other main elements, such as zinc, magnesium, and iron, were found to be 1.75%, 0.68%, and 0.93%, respectively, whereas in specimen no-19, the elements were 3.32%, 1.63%, and 1.20%, respectively.

Fractography of the fracture surface of specimen no-2 and specimens no-19 is shown in Fig. 4(a, b). The rapture took place at the HAZ in the direction of AA6061, where there was minimal microhardness [52]. Along the edges of the raptured samples, the shear plane with a cup-cone structure was produced. In specimen no-2 (Fig. 4a), a ductile failure with honeycomb dimples was identified [53]. At higher RS, homogeneously distributed Al_2O_3mp and finer grain structure were noticed due to effective material mixing. According to the fractography of specimen no-19, the coalescence of the micro-voids resulted in equiaxed and tiny dimples (Fig. 4b) [54]. The microhardness variance validated the

Table 4 Results of ANOVA for microhardness

Sources	SS	DF	MS	F value	P value	Remark
Model	3120.11	9	346.68	320.97	< 0.0001	Significant
RS	2024.93	1	2024.93	1874.79	< 0.0001	
TS	25.60	1	25.60	23.70	0.0007	
VF-Al_2O_3mp	842.72	1	842.72	780.24	< 0.0001	
RS*TS	28.12	1	28.12	26.04	0.0005	
RS* VF-Al_2O_3mp	8.00	1	8.00	7.41	0.0215	
TS* VF-Al_2O_3mp	32.00	1	32.00	29.63	0.0003	
RS^2	12.18	1	12.18	11.28	0.0073	
TS^2	94.26	1	94.26	87.27	< 0.0001	
VF-$Al_2O_3mp^2$	38.58	1	38.58	35.72	0.0001	
Residual	10.80	10	1.08			
Lack of fit	3.27	5	0.65	0.43	0.8097	Not significant
Pure-Error	7.53	5	1.51			
Cor Total	3130.91	19				
C.V. %	1.04				R^2	0.9966
Mean	129.10				Adjusted R^2	0.9934
Std. Dev	0.80				Predicted R^2	0.9832
					Adeq. Precision	68.06

ruptured zone's location near the HAZ and led to an increase in ductility [55].

3.2 Effect on Microhardness

The microhardness variation in the weld zone of all the reinforced joints is shown in Fig. 5(a–c). The irregular plastic flow on the retreating and advancing sides induced an asymmetrical variation of microhardness in the weld zone [56].

The microhardness of the as-received AA7075 and AA6061 was found to be 170 and 96 HV, respectively. The mean value of microhardness at the mid-point of the weld zone of various reinforced joints is summarized in Table 1. The results revealed that specimen no-19 exhibited the highest hardness of 157.5 HV, while specimen no-3 exhibited the lowest hardness of 106.7 HV among all the reinforced joints. The microhardness in the SZ appeared to be higher in comparison with the HAZ and TMAZ due to the smaller grain size induced by the DRX and the pinning effect caused by Al_2O_3mp; however, the HAZ's hardness was noted to be less due to grain coarsening at higher temperatures and over-aging [57, 58]. The reinforced joints' microhardness was also enhanced due to the inclusion of hard Al_2O_3mp. Compared to other specimens, specimen no-19 demonstrated the maximum microhardness due to higher VF-Al_2O_3mp and improved grain refinement.

4 Development of Mathematical Model

To determine the relevance of processing factors for response variance, an ANOVA was used. Tables 2, 3, and 4 present the ANOVA findings for tensile strength, %elongation, and microhardness. Equations 2–4 outline the regression model equations that were designed for the output response. The developed models were evaluated using Fisher's F test, with a 95% confidence level [59].

The developed regression equations are presented below.

$$\begin{aligned}\text{Tensile strength} = {} & +203.51 + 23.75 \times \text{RS} \\ & - 4.41 \times \text{TS} + 6.55 \times \text{VF-}Al_2O_3\text{mp} \\ & + 1.09 \times \text{RS} \times \text{TS} + 5.38 \times \text{RS} \times \text{VF-} \\ & Al_2O_3\text{mp} - 1.92 \times \text{TS} \times \text{VF-}Al_2O_3\text{mp} \\ & - 3.27 \times \text{RS}^2 + 2.15 \times \text{TS}^2 \\ & - 1.92 \times \text{VF-}Al_2O_3\text{mp}^2 \end{aligned} \quad (2)$$

$$\begin{aligned}\%\text{elongation} = {} & +15.37 + 2.20 \times \text{RS} - 0.23 \times \text{TS} \\ & + 1.41 \times \text{VF-}Al_2O_3\text{mp} + 0.29 \times \text{RS} \times \text{TS} \\ & + 0.34 \times \text{RS} \times \text{VF-}Al_2O_3\text{mp} + 0.36 \times \text{TS} \\ & \times \text{VF-}Al_2O_3\text{mp} + 0.76 \times \text{RS}^2 \\ & + 0.41 \times \text{TS}^2 - 0.59 \times \text{VF-}Al_2O_3\text{mp}^2 \end{aligned} \quad (3)$$

$$\begin{aligned}\text{Microhardness} = {} & +127.01 + 14.23 \times \text{RS} - 1.60 \times \text{TS} \\ & + 9.25 \times \text{VF-Al}_2\text{O}_3\text{mp} - 1.96 \times \text{RS} \times \text{TS} \\ & + 0.91 \times \text{RS} \times \text{VF-Al}_2\text{O}_3\text{mp} + 1.91 \times \text{TS} \\ & \times \text{VF-Al}_2\text{O}_3\text{mp} + 2.07 \times \text{RS}^2 \\ & + 5.82 \times \text{TS}^2 - 3.78 \times \text{VF-Al}_2\text{O}_3\text{mp}^2 \end{aligned} \quad (4)$$

All three models are considered significant based on their respective F values of 445.35, 99.83, and 320.97. The processing variables (TS, RS, and VF-Al_2O_3mp) are also found to be significant, as their P value is less than 0.05 in all the models. Since the P value is more than 0.1, the lack of fit is not considered to be significant as revealed in Tables 2, 3, and 4. The model is sufficient to proceed as the value of adequate precision is greater than 4, as revealed in Tables 2, 3, and 4. The R^2 reflects the model's overall variability. The values of R^2 for tensile strength, %elongation, and microhardness were found to be 0.9976, 0.9891, and 0.9966, respectively, whereas the value of adjusted R^2 was found to be 0.9914, 0.9791, and 0.9934, respectively. The difference between R^2 and adjusted R^2 revealed that the models did not describe 0.62, 1, and 0.32% of the total variation.

4.1 Processing Variables Influence on the Output Responses

Figure 6 depicts the predicted vs. actual output responses of several reinforced joints. In the scattered plots, the output parameters were found to be extremely close to the straight

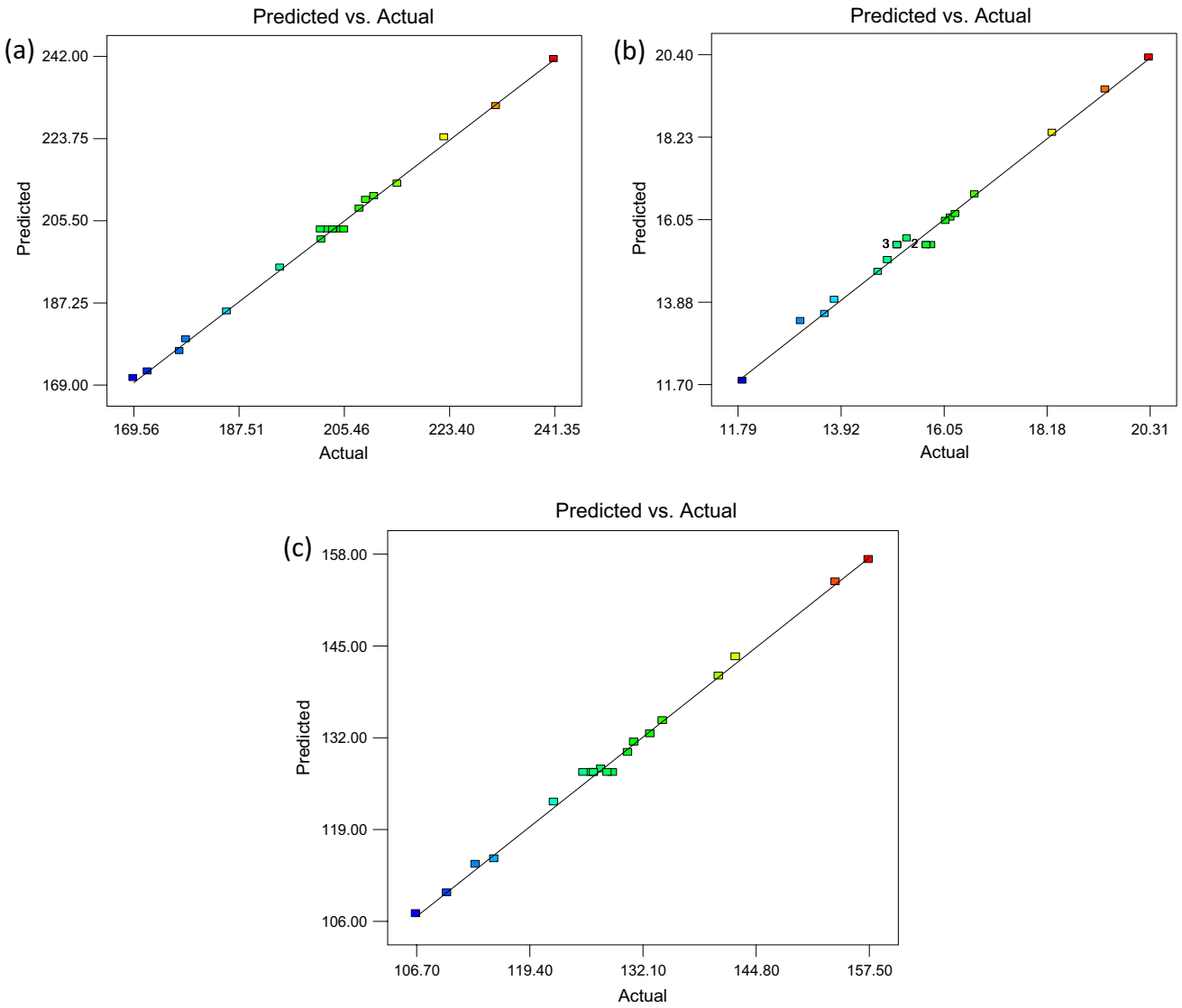


Fig. 6 **a**–**c** Actual versus predicted values plots; **a** tensile strength, **b** %elongation, **c** microhardness

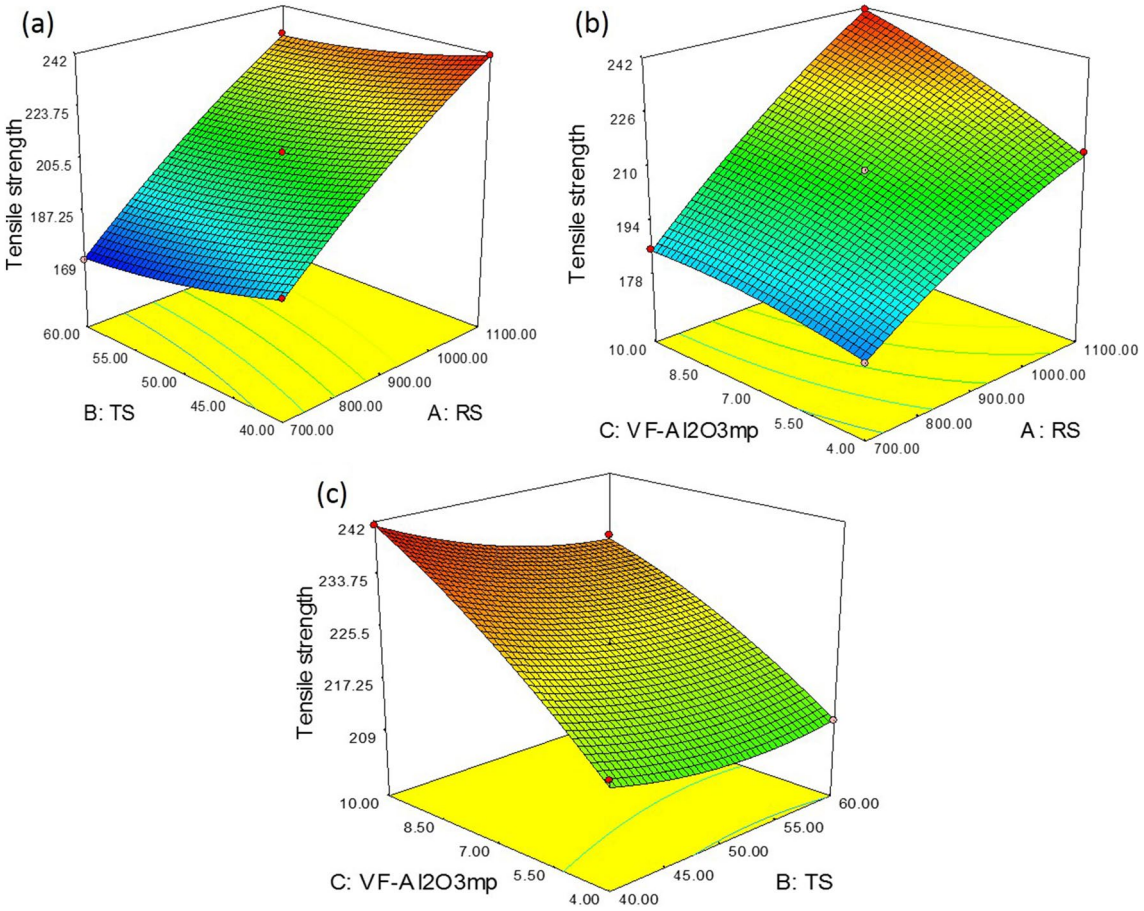


Fig. 7 Surface plots (3D) for tensile strength

lines, suggesting that the model's error was uniformly distributed. There is an excellent relationship between the actual and predicted response variables, showing that all of the regression models are acceptable.

The RSM was employed to optimize both output and input variables in the study. The resulting models enable the generation of 3D surface plots (Figs. 7, 8, and 9) using the output and input variables, and it was observed that the output responses were significantly influenced by the input variables. Specifically, the tensile strength increases as the RS and VF-Al_2O_3mp enhances, whereas the tensile strength exhibits decrements with increase in TS. This behavior can be attributed to the DRX and pinning effect of uniformly dispersed RPs at high RS and low TS. Furthermore, the %elongation and microhardness also increase with higher RS and VF-Al_2O_3mp, and when TS increases, the %elongation decreases, whereas microhardness decreases. A homogeneous microstructure and homogenous plastic flow were noticed at a higher RS of 1100 rpm. Consequently, improved tensile strength, % elongation, and microhardness were observed with increase in RS and VF-Al_2O_3mp [60], whereas the microhardness and tensile strength of the reinforced joints were enhanced with enhanced RS.

4.2 Multi-response Optimization: Desirability

RSM was employed to do multi-response optimization via the desirability function. The primary goal of the optimization process is to identify a variable setting that will accomplish each objective [61]. The objective is to minimize the %elongation and maximize the microhardness and tensile strength. To accomplish the desired responses within the constraints, a set of optimal variables with higher desirability is necessary [62]. The dot on the ramp graph (Fig. 10) represents the optimum input variable values for the best predicted response. The values of multi-optimized tensile strength, %elongation, and microhardness are 226.2 MPa, 17.4%, and 144.3 HV,

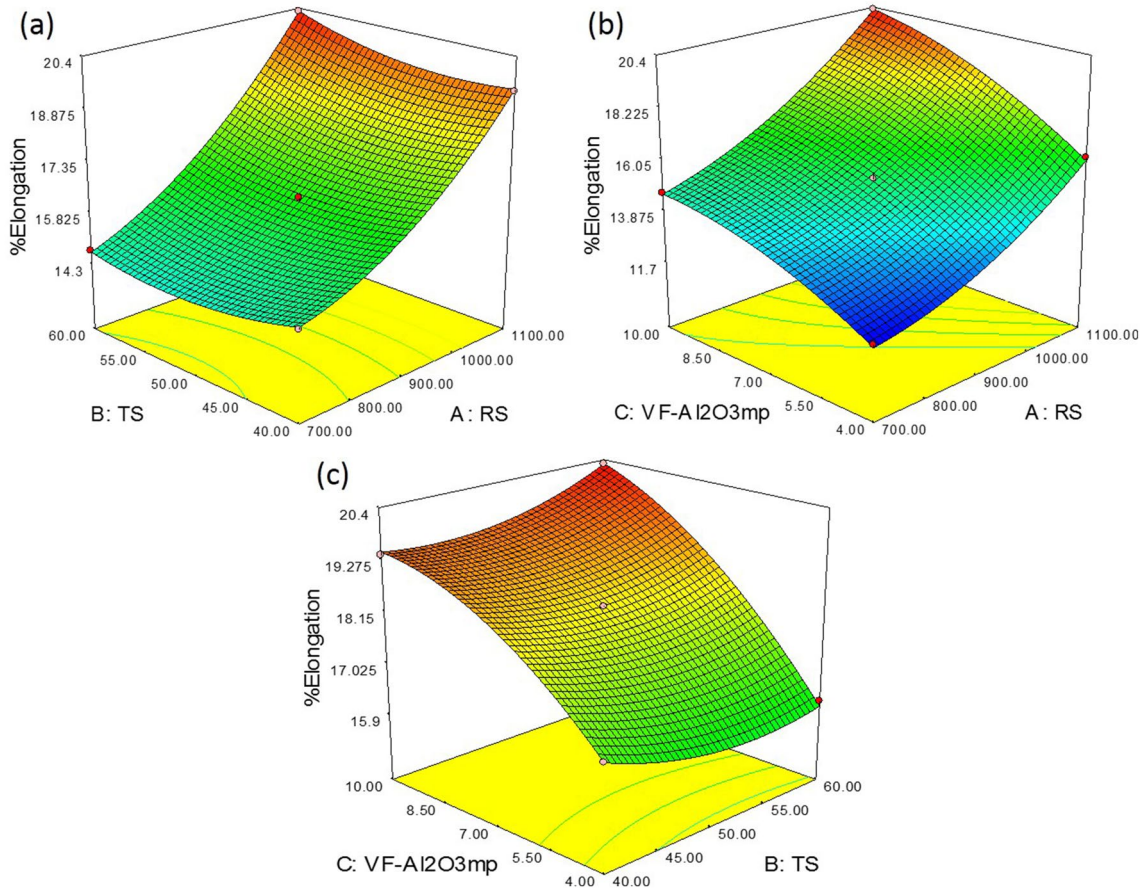


Fig. 8 Surface plots (3D) for %elongation

Table 5 Optimal variables setting and confirmatory test

Response variables	Optimum variables setting	Predicted optimum value	Experimental results	% Error
Tensile strength (MPa)	RS = 971 rpm,	226.2	234.4	3.6
%elongation (%)	TS = 40 mm/min,	17.4	18.1	4.0
Microhardness (HV)	VF-Al_2O_3mp = 10%	144.3	150.4	4.2

respectively. In contrast, the optimal values of process variables for RS, TS, and VF-Al_2O_3mp are 971 rpm, 40 mm/min, and 10%, respectively, as delineated in Fig. 10.

4.3 Confirmatory Tests

The goal of doing confirmatory tests was to check that the established mathematical models could appropriately represent each output response when the processing variables were adjusted to their optimal values. The results were averaged and compared to the predicted values of the response variables. Table 5 summarizes the results of the confirmatory tests and verifies the derived models' ability to predict responses.

5 Conclusions

In the current study, friction stir welding was utilized to develop reinforced joints of AA6061 and AA7075 incorporated with Al_2O_3mp. The following are some of the key findings from the research:

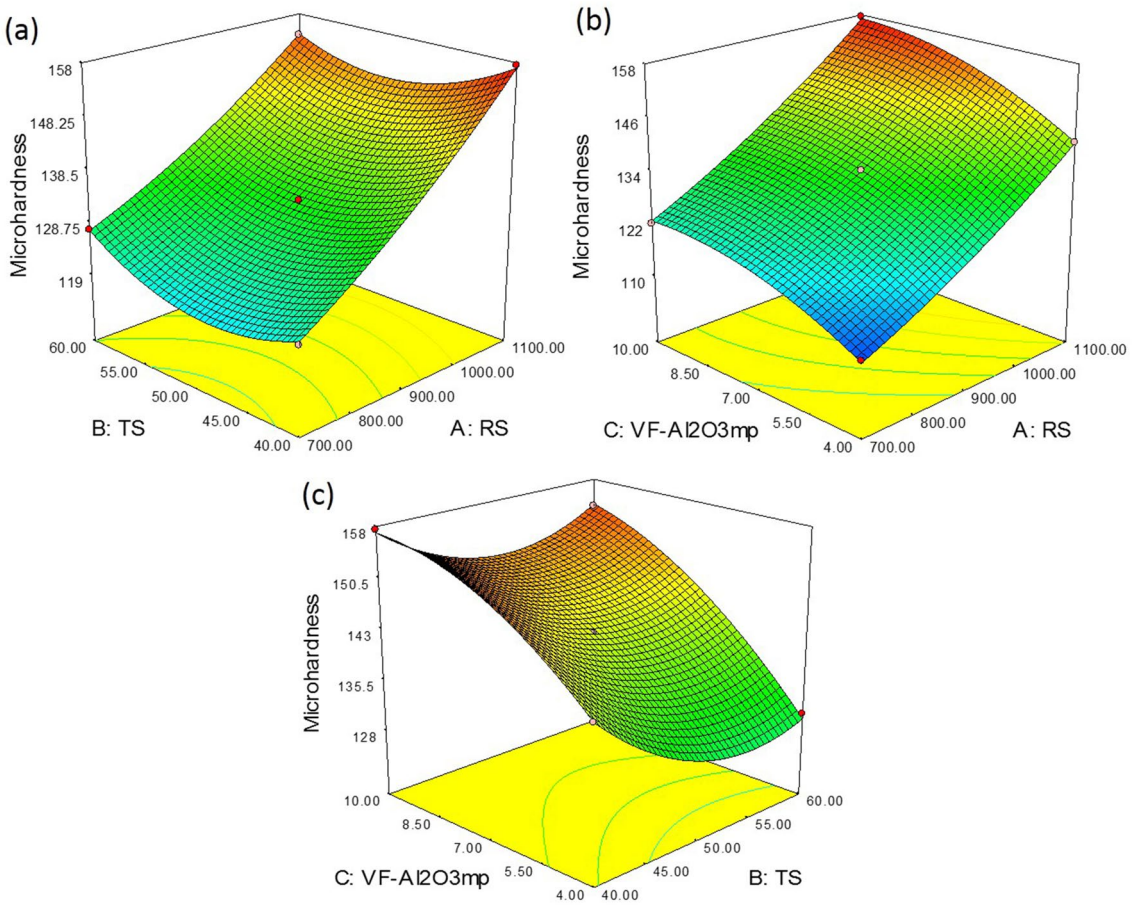


Fig. 9 Surface plots (3D) for microhardness

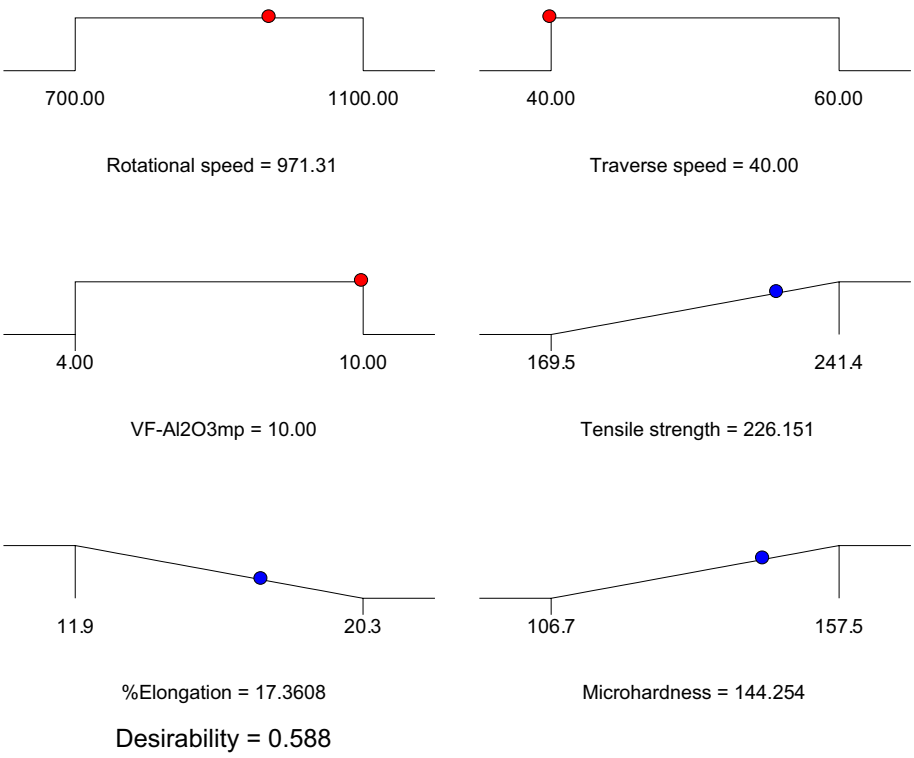


Fig. 10 Ramp graph for multi-response optimization

1. The homogenous dispersion of Al_2O_3mp was caused by an increase in RS and a decrease in TS, which led to smaller grains and enhanced mechanical properties.
2. The reinforced joints fabricated at RS of 1100 rpm, TS of 40 mm/min, and VF-Al_2O_3mp of 10% revealed a maximum tensile strength of 241.4 MPa and microhardness of 157.5 HV, whereas the reinforced joint fabricated at RS of 1100 rpm, TS of 60 mm/min, and 10% VF-Al_2O_3mp revealed the highest %elongation of 20.3%.
3. The developed quadratic modes of output responses were validated for suitability by conducting confirmatory tests. The findings showed that the model predictions were accurate within 5% of the observed values.
4. The optimum values of TS, RS, and VF-Al_2O_3mp were found to be 40 mm/min, 971 rpm, and 10%, respectively, whereas the multi-optimized response variables, including tensile strength, %elongation, and microhardness, were found to be 226.2 MPa, 17.4%, and 144.3 HV, respectively.

Funding Not applicable.

Declarations

Conflict of interest The author declares that they have no conflict of interest.

References

1. Fauzi M A N, Uday M B, Zuhailawati H, and Ismail A B, Microstructure and mechanical properties of alumina-6061aluminum alloy joined by friction welding. *Mater Des* **31** (2010) 670–676.
2. Mohan D G, and Wu C, A review on friction stir welding of steels. *Chin J Mech Eng* **34** (2021) 137. https://doi.org/10.1186/s10033-021-00655-3
3. Kuryntsev S, A review: laser welding of dissimilar materials (Al/Fe, Al/Ti, Al/Cu)—methods and techniques, microstructure and properties. *Materials* **15** (1), (2022) 122. https://doi.org/10.3390/ma15010122
4. Bohm M, and Kowalski M, Fatigue life estimation of explosive cladded transition joints with the use of the spectral method for the case of a random sea state. *Mar Struct* **71** (2020) 102739. https://doi.org/10.1016/j.marstruc.2020.102739
5. Agapiou J S, Filling friction stir welding in-process exit holes in copper squirrel cage rotors for electric motors. *Proc Manuf* **53** (2021) 802–813. https://doi.org/10.1016/j.promfg.2021.06.075
6. Hull D, and Clyne T, *An Introduction to Composite Materials*, Cambridge University Press, Great Britain (1996).
7. Hartley P, and Hartley D E, *FSW Implementation on the Space Shuttle's External Tank*, *Friction Stir Welding Technology for Defense Applications*, New Orleans, LA, (2002).
8. Chawla K K, *Composite Materials: Science and Engineering*, Springer, Heidelberg, NY (2012).
9. Alidokht S A, Abdollah-Zadeh A, Soleymani S, et al., Microstructure and tribological performance of an aluminium alloy based hybrid composite produced by friction stir processing. *Mater Des* **32** (5), (2011) 2727–2733.
10. Faraji G, and Asadi P, Characterization of AZ91/alumina nanocomposite produced by FSP. *Mater Sci Eng A* **527** (2011) 2320–2330.
11. Barmouz M, Asadi P, Besharati Givi M K, et al., Investigation of mechanical properties of Cu/SiC composite fabricated by FSP: effect of SiC particles' size and volume fraction. *Mater Sci Eng A* **528** (3), (2011) 1740–1749.
12. Tonelli L, Morri A, Toschi S, et al., Effect of FSP parameters and tool geometry on microstructure, hardness, and wear properties of AA7075 with and without reinforcing B C ceramic particles. *Int J Adv Manuf Technol* **102** (9–12), (2019) 391745–393945.
13. Deore H A, Mishra J, Rao A G, et al., Effect of filler material and post-process ageing treatment on microstructure, mechanical properties and wear behaviour of friction stir processed AA 7075 surface composites. *Surf Coat Technol* **374** (2019) 52–64.
14. Khodabakhshi F, Nosko M, and Gerlich A P, Effect of graphene nano-platelets (GNPs)on the microstructural characteristics and textural development of an Al-Mg alloy during friction-stir processing. *Surf Coat Technol* **335** (2018) 288–305.
15. Song Q, Wang H, Ji S, Ma Z, Jiang W, and Chen M, Improving joint quality of hybrid friction stir welded Al/Mg dissimilar alloys by RBFNN-GWO system. *J Manuf Process* **59** (2020) 750–759. https://doi.org/10.1016/j.jmapro.2020.10.037
16. Feng A H, Xiao B L, and Ma Z Y, Effect of microstructural evolution on mechanical properties of friction stir welded AA2009/SiC composite. *Compos Sci Technol* **68** (9), (2008) 2141–2148.
17. Ceschini L, Boromei I, Minak G, et al., Effect of friction stir welding on microstructure, tensile and fatigue properties of the AA7005/10 vol.%Al2O3 composite. *Compos Sci Technol* **67** (3–4), (2007) 605–615.
18. Uzun H, Friction stir welding of SiC particulate reinforced AA2124 aluminium alloy matrix composite. *Mater Des* **28** (5), (2007) 1440–1446.
19. Kumar K S, Murigendrappa S M, and Kumar H, Experimental investigation on effects of varying volume fractions of SiC nanoparticle reinforcement on microstructure and mechanical properties in friction stir welded dissimilar joints of AA2024-T351 and AA7075-T651. *J Mater Res* **455** (2019).
20. Ashok Kumar R, Ragav G R, Nagarajan K J, et al., Effect of hybrid reinforcement at stirred zone of dissimilar aluminum alloys during friction stir welding. *Metall Res Technol* **116** (2019) 631.
21. Abd Elnabi M M, El Mokadem A, and Osman T, Optimization of process parameters for friction stir welding of dissimilar aluminum alloys using different Taguchi arrays. *Int J Adv Manuf Technol* **121** (2022) 3935–3964. https://doi.org/10.1007/s00170-022-09531-3
22. Verma S, and Kumar V, Optimization of friction stir welding parameters of dissimilar aluminium alloys 6061 and 5083 by using response surface methodology. *Proc Inst Mech Eng Part C J Mech Eng Sci* **235** (23), (2021) 7009–7020. https://doi.org/10.1177/09544062211005804
23. Senthil S M, Parameshwaran R, Nathan R, Bhuvanesh Kumar M, and Deepandurai K, A multi-objective optimization of the friction stir welding process using RSM-based-desirability function approach for joining aluminum alloy 6063–T6 pipes. *Struct Multidiscipl Optim* (2020). https://doi.org/10.1007/s00158-020-02542-2
24. Mishra R S, and Jain S, Parametric optimization of FSWed dissimilar composite joints of AA7075 and AA6061 using RSM. *Trans Indian Inst Met* (2023). https://doi.org/10.1007/s12666-023-02996-0
25. John J, Shanmuganatan S P, and Kiran M B, Influence and optimization of input parameters on mechanical properties of friction

stir processed AA 2014–T6. *Mater Today Proc* **5** (2018) 25458–25467. https://doi.org/10.1016/j.matpr.2018.10.351
26. Sasikumar A, Gopi S, and Mohan D G, Prediction of filler added friction stir welding parameters for improving corrosion resistance of dissimilar aluminium alloys 5052 and 6082 joints. *Adv Mater Sci* **22** (2022) 79–95. https://doi.org/10.2478/adms-2022-0014
27. Sasikumar A, Gopi S, and Mohan D G, Effect of welding speed on mechanical properties and corrosion resistance rates of filler induced friction stir welded AA6082 and AA5052 joints. *Mater Res Express* **8** (6), (2022) 066531. https://doi.org/10.1088/2053-1591/ac0c9e
28. Anandha Kumar C J, Gopi S, Mohan D G, and Shashi K S, Predicting the ultimate tensile strength and wear rate of aluminium hybrid surface composites fabricated via friction stir processing using computational methods. *J Adhesion Sci Technol* **36** (16), (2022) 1707–1726. https://doi.org/10.1080/01694243.2021.1982237
29. Jain S, and Mishra R S, Microstructural and mechanical behavior of micro-sized SiC particles reinforced friction stir processed/welded AA7075 and AA6061. *Silicon* **14** (2022) 10741–10753. https://doi.org/10.1007/s12633-022-01716-5
30. Raj R, and Thakur D, Effect of particle size and volume fraction on the strengthening mechanisms of boron carbide reinforced aluminum metal matrix composites. *Proc Inst Mech Eng Part C J Mech Eng Sci* **233** (4), (2019) 1345–1356. https://doi.org/10.1177/0954406218771997
31. Shafiei-Zarghani S, Kashani-Bozor F, and Zarei-Hanzaki A, Microstructures and mechanical properties ofAl/Al2O3 surface nano-composite layer produced by friction stir processing. *Mater Sci Eng A* **500** (1–2), (2009) 84–91.
32. Raaft M, Mahmoud T S, Zakaria H M, and Khalifa T A, Microstructural, mechanical and wear behavior of A390/graphite and A390/Al2O3 surface composites fabricated using FSP. *Mater Sci Eng A* **528** (18), (2011) 5741–5746.
33. Bahemmat P, Haghpanahi M, Besharati M K, et al., Study on mechanical, micro-, and macrostructural characteristics of dissimilar friction stir welding of AA6061-T6 and AA7075-T6. *Proc Inst Mech Eng Part B J Eng Manuf* **224** (12), (2010) 1854–1865.
34. Çömez N, and Durmuş H, Cold metal transfer welding of AA6061 to AA7075: mechanical properties and corrosion. *ASME J Eng Mater Technol* **141** (3), (2019) 302.
35. Aherwar A, Patnaik A, and Pruncu C I, Effect of B4C and waste porcelain ceramic particulate reinforcements on mechanical and tribological characteristics of high strength AA7075 based hybrid composite. *J Mater Res Technol* **9** (5), (2020) 9882–9894. https://doi.org/10.1016/j.jmrt.2020.07.003
36. Rathee S, Maheshwari S, Siddiquee A N, et al., Effect of tool plunge depth on reinforcement particles distribution in surface composite fabrication via friction stir processing. *Def Technol* **13** (2), (2017) 86–91.
37. Rathee S, Maheshwari S, Siddiquee A N, et al., Analysis of microstructural changes in enhancement of surface properties in sheet forming of Al alloys via friction stir processing. *Mater Today Proc* **4** (2), (2017) 452–458.
38. Godhani P S, Patel V V, Vora J J, Chaudhary N D, and Banka R, *Effect Friction Stir Welding of Aluminum Alloys AA6061/AA7075 : Temperature Measurement , Microstructure , and Mechanical Properties*. Springer, Singapore
39. Cole E G, Fehrenbacher A, Duffie N A, and Zinn M R, Weld temperature effects during friction stir welding of dissimilar aluminum alloys 6061–t6 and 7075–t6. *Int J Adv Manuf Technol* **71** (2014) 643–652.
40. Mehdi H, and Mishra R S, An experimental analysis and optimization of process parameters of AA6061 and AA7075 welded joint by TIG+FSP welding using RSM. *Adv Mater Process Technol* (2020). https://doi.org/10.1080/2374068X.2020.1829952
41. Batool A, Anjum N A, and Khan M Y, Impact of pin tool designs on the mechanical properties and microstructure of aluminum alloys AA7075- AA6061. *J Chin Inst Eng* **44** (7), (2021) 683–693. https://doi.org/10.1080/02533839.2021.1940293
42. Srinivasu R, Rao A S, Reddy G M, and Rao K S, Friction stir surfacing of cast A356 aluminum silicon alloy with boron carbide and molybdenum disulphide powders. *Def Technol* **11** (2015) 140–146.
43. Mehdi H, and Mishra R, Microstructure and mechanical characterization of tungsten inert gas-welded joint of AA6061 and AA7075 by friction stir processing. *Proc Inst Mech Eng Part L J Mater Des Appl* **235** (11), (2021) 2531–2546. https://doi.org/10.1177/14644207211007882
44. Yadav D, and Bauri R, Processing, microstructure and mechanical properties of nickel particles embedded aluminium matrix composite. *Mater Sci Eng A* **528** (2011) 1326–1333.
45. Salah A N, Mabuwa S, Mehdi H, Msomi V, Kaddami M, and Mohapatra P, Effect of multipass FSP on Si-rich TIG welded joint of dissimilar aluminum alloys AA8011-H14 and AA5083-H321: EBSD and microstructural evolutions. *Silicon* (2022). https://doi.org/10.1007/s12633-022-01717-4
46. Chen X G, Silva M D, Gougeon P, and Georges L S, Microstructure and mechanical properties of friction stir welded AA6063–B4C metal matrix composites. *Mater Sci Eng A* **518** (12), (2009) 174–184.
47. Palanivel R, Dinaharan I, Laubscher R F, and Paulo Davim J, Influence of boron nitride nanoparticles on microstructure and wear behavior of AA6082/TiB_2 hybrid aluminum composites synthesized by friction stir processing. *Mater Des* **106** (2016) 195–204.
48. Srivastava M, Rathee S, Siddiquee A N, et al., Investigation on the effects of silicon carbide and cooling medium during multipass FSP of Al-Mg/ SiC surface composites. *Silicon* **11** (2019) 2149–2157.
49. Azizieh M, Kokabi A H, and Abachi P, Effect of rotational speed and probe profile on microstructure and hardness of AZ31/Al2O3 nanocomposites fabricated by friction stir processing. *Mater Des* **32** (4), (2011) 2034–2041.
50. Chang C I, Lee C J, and Huang J C, Relationship between grain size and Zener-Holloman parameter during friction stir processing in AZ31 Mg alloys. *Scripta Mater* **51** (2004) 509–514.
51. Manroo S, Khan N, and Ahmad B, Study on surface modification and fabrication of surface composites of magnesium alloys by friction stir processing: a review. *J Eng Appl Sci* **69** (2022) 25. https://doi.org/10.1186/s44147-022-00073-9
52. Lloyd D J, Particle reinforced aluminum and magnesium matrix composites. *Int Mater Rev* **39** (1), (1994) 1–23.
53. Salehi M, Saadatmand M, and Aghazadeh Mohandesi J, Optimization of process parameters for producing AA6061/SiC nanocomposites by friction stir processing. *Trans Nonferrous Met Soc China* **22** (5), (2012) 1055–1063.
54. Pandey C, Saini N, Mahapatra M M, and Kumar P, Study of the fracture surface morphology of impact and tensile tested cast and forged (C&F) Grade 91 steel at room temperature for different heat treatment regimes. *Eng Fail Anal* **71** (2017) 131–147.
55. Mehdi H, and Mishra R S, Study of the influence of friction stir processing on tungsten inert gas welding of different aluminum alloy. *SN Appl Sci* **1** (7), (2019) 712.
56. Xu W, Liu J, Luan G, and Dong C, Temperature evolution, microstructure and mechanical properties of friction stir welded thick 2219-O aluminum alloy joints. *Mater Des* **30** (2009) 1886–1893.
57. Ebrahimzadeh V, Paidar M, Safarkhanian M A, and Ojo O O, Orbital friction stir lap welding of AA5456-H321/AA5456-O

aluminum alloys under varied parameters. *Int J Adv Manuf Technol* **96** (2018) 1237–1254.
58. Jain S, and Mishra R S, Effect of Al2O3 nanoparticles on microstructure and mechanical properties of friction stir-welded dissimilar aluminum alloys AA7075-T6 and AA6061-T6. *Proc Inst Mech Eng Part E J Process Mech Eng* (2021). https://doi.org/10.1177/09544089211065534
59. Puri A B, and Bhattacharyya B, Modeling and analysis of white layer depth in a wire-cut EDM process through response surface methodology. *Int J Adv Manuf Technol* **25** (2005) 301–307.
60. Patel V V, Badheka V, and Kumar A, Influence of friction stir processed parameters on superplasticity of Al-Zn-Mg-Cu alloy. *Mater Manuf Proc* **31** (2016) 1573–1582.
61. Ghangas G, and Singhal S, Modelling and optimization of process parameters for friction stir welding of armor alloy using RSM and GRA-PCA approach. *Mater Res Express* **6** (2018) 026553
62. Shanavas S, and Dhas J E R, Parametric optimization of friction stir welding parameters of marine grade aluminium alloy using response surface methodology. *Trans Nonferrous Met Soc China* **27** (2017) 2334–2344.

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

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

NEURAL REGENERATION RESEARCH
www.nrronline.org

Review

New insights into the therapeutic approaches for the treatment of tauopathies

Himanshi Singh[1, 2, 3], Asmita Das[2], Mohammad Moshahid Khan[3, 4, 5, *], Tayebeh Pourmotabbed[1, *]

https://doi.org/10.4103/1673-5374.385288

Date of submission: May 2, 2023

Date of decision: July 5, 2023

Date of acceptance: August 10, 2023

Date of web publication: September 22, 2023

Abstract

Tauopathies are a group of neurological disorders, including Alzheimer's disease and frontotemporal dementia, which involve progressive neurodegeneration, cognitive deficits, and aberrant tau protein accumulation. The development of tauopathies cannot currently be stopped or slowed down by treatment measures. Given the significant contribution of tau burden in primary tauopathies and the strong association between pathogenic tau accumulation and cognitive deficits, there has been a lot of interest in creating therapies that can alleviate tau pathology and render neuroprotective effects. Recently, small molecules, immunotherapies, and gene therapy have been used to reduce the pathological tau burden and prevent neurodegeneration in animal models of tauopathies. However, the major pitfall of the current therapeutic approach is the difficulty of drugs and gene-targeting modalities to cross the blood-brain barrier and their unintended side effects. In this review, the current therapeutic strategies used for tauopathies including the use of oligonucleotide-based gene therapy approaches that have shown a promising result for the treatment of tauopathies and Alzheimer's disease in preclinical animal models, have been discussed.

Key Words: dementia; gene therapies; immunotherapy; neurodegeneration; oligonucleotides; tau; tauopathies; therapies

From the Contents

Introduction	1020
Search Strategy	1021
Periodic Prevalence of Tau Protein	1021
Tau Protein Expression and Regulation in Different Organs	1022
The Role of Tau Protein in Neurodegeneration	1022
Therapeutic Approaches for Targeting the Aberrant Tau Protein	1022
Conclusion and Future Perspectives	1024

Introduction

Tauopathy is a neurodegenerative disorder caused by aberrant tau protein aggregates in cells. A multipurpose, distinctive, and naturally disordered protein, tau protein is found in the brain and central nervous system. The tau protein forms oligomers or filaments and interacts with other molecules involved in various signaling pathways. The microtubule-associated protein tau (MAPT) gene contains 16 exons where alternative splicing of exons 2, 3, and 10 transcripts leads to the formation of six different molecular isoforms (**Figure 1**) ranging from 352–441 amino acids with a molecular weight of 45 kDa–65 kDa. These isoforms largely vary from each other by the number of repeats; three repeats (3R) or four repeats (4R), in their microtubule-binding domains (MTBR). Insertion or deletion of a 31 or 32 amino acid sequence coded by exon 10 results in 3R (R1, R3, R4) or 4R (R1, R2, R3, R4) tau protein isoforms (Wang and Mandelkow, 2016). The isoforms 3R and 4R form abnormal aggregates, thus tauopathies are categorized into three groups: 3R, 4R, and 3R/4R tauopathies (comprising 3R tau & 4R tau in equal proportions). Recent studies have demonstrated that paired helical filament (PHF)-6 motifs (275VQIINK280 or 306VQIVYK311) present within R2 and R3 MTBR are required for the pathological aggregation of tau protein and make them seed competent (Mirbaha et al., 2018). Exons 2 and 3 encode 29 or 58 amino acid sequences that, when inserted or deleted, result in the production of the tau protein isoforms 0N, 1N, and 2N, respectively. Interestingly, exon 3 is never expressed alone; it is always expressed in conjunction with exon 2. On the other hand, exon 2 seems to have no such anomaly. Moreover, the 2N tau protein isoforms are under expressed in the human brain (Wang and Mandelkow, 2016).

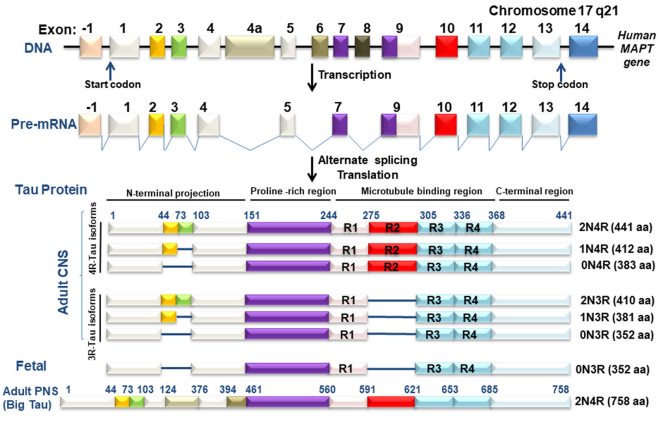


Figure 1 | Schematic representation of the wild-type MAPT gene and its variants.
MAPT gene exons are numbered from-1 to 14. In the adult human brain, six isoforms are produced as a result of alternate splicing of exon 2 (yellow), exon 3 (green), and exon 10 (red) in the CNS. Each isoform is distinguished by either three (3R) or four (4R) microtubule-binding domains and 0N, 1N, and 2N amino-terminal end insertions or deletions. Exons 4a (brown) and 6 (dark brown) have been represented in isoforms that express in the PNS, whereas exon 8 (black) is not transcribed in the human brain. Fetal brain tau lacks exon 2 (yellow), exon 3 (green), and exon 10 (red). Created with Microsoft PowerPoint. aa: Amino acid; CNS: central nervous system; DNA: deoxyribonucleic acid; MAPT: microtubule-associated protein tau; mRNA: messenger RNA; PNS: peripheral nervous system.

The mature neurons of the human brain predominantly express tau protein, whereas oligodendrocytes and astrocytes express tau protein to a much smaller extent (Chung et al., 2021). Under normal circumstances, tau protein stabilizes the microtubules, a crucial component of the neuronal cytoskeleton, and regulates intracellular trafficking (**Figure 2**). Disruption in tau protein regulation and/or function have been linked to neuropathological conditions i.e., tauopathies including Alzheimer's disease (AD), frontotemporal dementia (FTD), Pick disease, corticobasal syndrome (CBS), progressive supranuclear palsy (PSP), chronic traumatic encephalopathy (CTE), and argyrophilic grain disease (Chung et al., 2021; **Figure 2**). It is evident that the neurons of

[1]Department of Microbiology, Immunology and Biochemistry, University of Tennessee Health Science Center, Memphis, TN, USA; [2]Department of Biotechnology, Delhi Technological University, Shahbad Daulatpur, Delhi, India; [3]Department of Neurology, College of Medicine, University of Tennessee Health Science Center, Memphis, TN, USA; [4]Neuroscience Institute, University of Tennessee Health Science Center, Memphis, TN, USA; [5]Center for Muscle, Metabolism and Neuropathology, Division of Regenerative and Rehabilitation Sciences and Department of Physical Therapy, College of Health Professions, University of Tennessee Health Science Center, Memphis, TN, USA

*Correspondence to: Mohammad Moshahid Khan, PhD, mkhan26@uthsc.edu; Tayebeh Pourmotabbed, PhD, tpourmot@uthsc.edu.
https://orcid.org/0000-0003-4679-294X (Mohammad Moshahid Khan); https://orcid.org/0000-0003-3468-5880 (Tayebeh Pourmotabbed)

Funding: *This work was supported by National Institute of Health grant number R03AG075597 (to MMK and TP) and Department of Defense Award Number HT9425-23-1-0043 (to MMK).*
How to cite this article: *Singh H, Das A, Khan MM, Pourmotabbed T (2024) New insights into the therapeutic approaches for the treatment of tauopathies. Neural Regen Res 19(5):1020-1026.*

Downloaded from http://journals.lww.com/nrronline by BhDMf5ePHKav1zEoum1tQfN4a+kJLhEZgbsIHo4XMi0hCywCX1AWnYQp/IlQrHD3i3D0OdRyi7TvSFl4Cf3VC4/OAVpDDa8KKGKV0Ymy+78= on 11/03/2023

NEURAL REGENERATION RESEARCH
www.nrronline.org

Downloaded from http://journals.lww.com/nrronline by BhDMf5ePHKav1zEoum1tQfN4a+kJLhEZgbsIHo4XMi0hCywCX1AWnYQp/IlQrHD3i3D0OdRyi7TvSFl4Cf3VC4/OAVpDDa8KKGKV0Ymy+78= on 11/03/2023

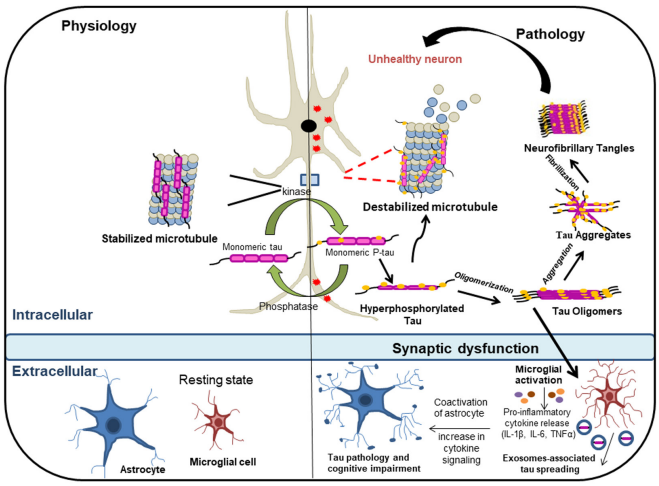


Figure 2 | **Schematic representation of the progression of tau pathology.**
In a healthy state, tau protein controls microtubule stabilization. Hyperphosphorylation of tau protein causes a reduction of microtubule affinity in tauopathies. Eventually, pathological soluble tau oligomers are formed and further aggregate to form insoluble neurofibrillary tangles. Tau oligomers are released into the extracellular space, which helps in the propagation of pathological tau protein to nearby neurons. Pro-inflammatory mediators like IL-1β, IL-6, and TNF-α are produced by microglia in response to inflammatory stimuli, which increases the activity of kinases implicated in tau protein phosphorylation and further exacerbates the disease. Created with Microsoft PowerPoint. IL-1β: Interleukin-1 beta; IL-6: interleukin-6; TNF-α: tumor necrosis factor alpha.

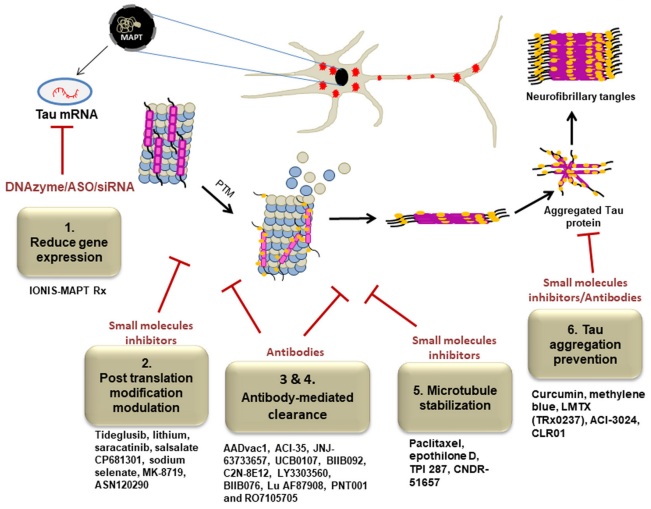


Figure 3 | **Schematic representation of potential therapeutic strategies for the treatment of tauopathies.**
Tau gene expression is decreased by targeting anti-sense oligonucleotides (DNAzyme, ASOs, siRNA) against the mRNA from the MAPT gene (targeted strategy 1). Kinase-mediated hyperphosphorylation, O-GlcNAc removal inhibition, and tau acetylation inhibition are blocked to sustain healthy PTM pathways (targeted strategy 2). Antibodies are utilized to remove or sequester pathologic tau species, inhibiting the spread of the condition from cell to cell (targeted strategies 3 & 4). Loss of function of tau is addressed by restoring microtubule stabilization (targeted strategy 5). The formation of pathological tau protein aggregates is prevented through antibodies mediated clearance and small molecules based inhibitors (targeted strategy 6). Each targeted strategy included examples of potential drug candidates for the treatment of tauopathies. Created with Microsoft PowerPoint. ASOs: Antisense oligonucleotides; DNAzymes: single-stranded catalytic DNA molecules; MAPT: microtubule-associated protein tau; PTM: post-translational modification; O-GlcNAc: O-linked-N-acetylglucosamine acetylation; siRNA: small interfering RNA.

patients with neurological disorders possess aggregates of various aberrant proteins, of which hyperphosphorylated tau is a key component. Glycogen synthase kinase-3 (GSK3) has been identified as the primary kinase involved in the hyperphosphorylation of tau protein (Sayas and Ávila, 2021). The GSK3-mediated phosphorylation of tau protein is attributed to the presence of amyloid beta (Aβ), which may inhibit insulin receptors to eventually promote GSK3 activation. The complexes that form as a result of the self-assembling of tau protein further complicate the disease condition through the degeneration of the glial and neuronal cells (You et al., 2022).

The presence of hyperphosphorylated tau protein and its aberrant aggregates in neuronal cells induce apoptosis and cell death (Sayas and Ávila, 2021). There is a significant amount of evidence suggesting that the hyperphosphorylated tau protein could bind to normal microtubule-associated proteins (MAPs) such as MAP1B and MAP2, resulting in synaptic degeneration (Alonso et al., 1997). In AD, the intracellular tangles caused by pathological tau protein inversely affect the abundance of healthy neurons. These pathological conditions attributed directly or indirectly to the existence of abnormal tau protein aggregates uncover the urgency for the discovery of efficient therapeutics for treating tauopathies. Although there are no established and approved pharmacological treatments available to date, several small molecules have been developed to attenuate the causes that lead to tauopathy, such as hyperphosphorylation of tau protein. Immunotherapies and catalytic nucleic acid-based gene therapies are the new treatment strategies used against tauopathies (**Figure 3**). In this review, we will cover key features and findings of tauopathies, existing anti-tau RNA/protein therapies, their shortcomings, and therapeutic strategies that have been emerging as potential treatments in clinical settings.

Search Strategy

The literature search was performed for the period from 1994 to 2023 with the major emphasis on the last 5 years of publication using the PubMed database. Some information was also taken from the websites page of the World Health Organization and AlzForum (https://www.alzforum.org/). The search keywords were as follows: "Tauopathies" and "Neurodegenerative disorders" OR "COVID-19" OR "Dementia" OR "Small Molecules Inhibitors for tauopathies" OR "Vaccines for tauopathies" OR "Gene Therapy for tauopathies" OR "Oligonucleotides" OR Clinical Trials OR Animal Models OR Drug Discovery for tauopathies.

Periodic Prevalence of Tau Protein

Tauopathies are progressive and heterogeneous neurodegenerative disorders defined by the buildup of aberrant tau protein deposits in the brain. In line with the World Health Organization (WHO), more than 55 million individuals globally are estimated to be affected by dementia, and each year approximately 10 million new cases of dementia are emerging. The number is expected to rise as much as 138 million by 2050. FTD is considered one of the commonly occurred presenile dementia which contributes to 2.6% of all types of dementia (Leroy et al., 2021). The reported age-standardization incidence was 2.90 per 100,000 individuals per year, which is increasing with time and goes up between the age groups of 75–79 (Turcano et al., 2020; Leroy et al., 2021). In spite of having a similar progression rate and mental state examination, FTD patients are underdiagnosed due to a longer referral delay compared to AD patients. Reports demonstrated the shortest survival rate for FTD patients with primarily motor symptoms whereas the survival rate of those patients with behavioral variant FTD, and language variant FTD did not differ significantly (Leroy et al., 2021). For a better understanding of the perspective of FTD disease, a European observational research study is currently being conducted (Borroni et al., 2022). A handful of population-based studies have been conducted to understand the PSP and CBS, which are uncommon groups of tauopathies. It has been found that the PSP and CBS incidence are about 3–8 per 100,000 individuals per year which increase with the age of above 75–79 (Stang et al., 2020). The average survival rate of individuals diagnosed with PSP or CBS is about 6 years (Stang et al., 2020). Argyrophilic grain disease is a 4R, late-onset tauopathy associated with advanced age (> 80 years old). It affects both genders equally. Studies indicated that lower socioeconomic status and appetite disorders are also the contributing factor other than the older age (Wurm et al., 2020). CTE is another notable dementia that has received little attention due to a scarcity of cases. Several studies have linked CTE to sports, presenting evidence that the pathology of CTE is commonly observed in athletes, primarily men (Bieniek et al., 2020).

AD is the most prevalent type of dementia, contributing 60–70 percent of total cases worldwide. Global data show that low-income nations have a greater prevalence of AD (Leroy et al., 2021). The incidence of dementia is increasing at an accelerated rate among the aging population by all means, which increases the medical cost of AD, including the unpaid and unseen efforts of care providers (El-Hayek et al., 2019). Since its discovery, research has progressed to understand the pathological processes of AD over the past several decades, although the development of therapeutic tools to delay the progress of neurodegeneration is still unmet. The majority of the research attention in the search for AD therapies for managing disease has been focused on two pathological aspects of the disease: neurofibrillary tangles (NFTs), which are composed of hyperphosphorylated tau protein, and Aβ protein that aggregates into plaques. Unfortunately, efforts to either remove Aβ peptide or reduce its production in order to mitigate the disease course remained largely unsuccessful and showed negative results in several clinical trials over the decades. The failure of drugs targeting Aβ pathology in the brain has fueled increased interest in tau protein-targeting strategies. Aberrant tau protein is emerging as a promising therapeutic candidate for the discovery of effective treatments for AD and related tauopathies. Additionally, during the unprecedented COVID-19 era, the risk of cognitive decline has additionally been noted in both AD and non-AD patients with mild to severe COVID-19. The elevated expression of pathological tau protein as well as upregulation of deleterious pathways causing hyperphosphorylation of tau protein in COVID-19 patients (Frontera et al., 2022; Reiken et al., 2022) has alarmed the ongoing burden of disease by upticking future patients (Gordon et al., 2022). However, therapeutic intervention is needed to underpin the new wave of cognitive decline or dementia.

Tau Protein Expression and Regulation in Different Organs

The expression of the tau protein isoforms is tightly controlled. The 45 kDa 0N3R tau protein isoform is expressed only in the fetal human brain. While the mature human brain expresses 3R and 4R tau protein isoforms with the same proportion (1:1). The level of tau protein phosphorylation is highest during neuronal development, and the degree of tau protein phosphorylation contributes to cytoskeletal plasticity in developing neurons in the brain (Betters et al., 2023). In spite of having a highly phosphorylated state, the fetal tau protein isoform does not form PHF (Hefti et al., 2019). Tau protein is expressed not only in the adult central nervous system (CNS) but also in the muscle, liver, lung, kidney, skin, and other tissues. The adult peripheral nervous system specifically the sciatic nerve, optic nerves, spinal cord, and the cell lines derived from neuronal cells express a 110 kDa tau protein variant referred to as "Big tau" (Dugger et al., 2016). Big tau protein contains exon 4a and is developmentally regulated and expressed more in PNS neurons as they mature during adulthood (Fischer and Baas, 2020). The specific function of "Big tau" protein is still unknown; however, indirect evidence suggests that it may contribute to axonal stability and aid in microtubule stabilization (Fischer and Baas, 2020). The differential expression of tau protein in various tissues and the nervous system opens the door to further research into their relationship to the onset of AD and tauopathies.

The Role of Tau Protein in Neurodegeneration

Tau protein oligomers have become an impetus for tauopathies, especially in their abnormally hyperphosphorylated aggregate form. In these forms, tau protein loses its propensity for microtubules and begins to self-assemble to induce the development of PHF, straight filaments, or twisted ribbons, which together evolve as NFTs (Oakley et al., 2020; **Figure 2**). Along with its connection to synaptic dysfunction and neuronal death, NFTs density is highly linked with cognitive decline in AD (Hsieh et al., 2019). Several recent studies speculated that hyperphosphorylation of tau protein, particularly on serine and threonine residues, leads to their aggregation (Zhou et al., 2018). In addition, tau protein goes through additional post-translational modifications (PTMs), including sumoylation, acetylation, ubiquitination, cleavage, and glycation (Alquezar et al., 2021). Site-specific tau protein PTMs can alter tau protein solubility or disrupt the connection between tau protein and microtubules, which can affect tau protein pathophysiology. In comparison to the non-aggregated tau protein, the aggregated ones are more difficult to get dephosphorylated by the protein phosphatase 2A (PP2A) and result in tau pathology (Miao et al., 2019). There is evidence that a pathogenic tau protein similar to prion disease can directly infect nearby and synaptically linked cells (Duyckaerts et al., 2019). Tau protein is an intracellular protein, which upon axonal degeneration, neuronal death, or through direct translocation from cytoplasm to plasma membrane will release to extracellular space (Brunello et al., 2020). Tau protein translocation may also take place via presynaptic vesicle secretion, exosomes, and ectosomes or in its naked form (Polanco and Götz, 2022). Once released, it interacts with low-density lipoprotein receptor-related protein-1, heparan sulfate proteoglycans, or muscarinic receptors and will be internalized by other neurons via dynamin-mediated endocytosis and tunneling nanotubes connecting the cytoplasmic content of adjacent cells (Zhang et al., 2021). Once seeded, it stimulates the aggregation of natively folded tau protein inside the naive cells, causing cellular toxicity and the spread of disease, resembling a prion-like propagation hypothesis (Duyckaerts et al., 2019; Brunello et al., 2020). In a mouse model of tauopathy, the suppression of exosome production and the reduction in microglia activation stop the spread of abnormal tau protein (Asai et al., 2015). These results show that microglia play a crucial role in the spread of tauopathies through phagocytosis and the production of exosomes that carry tau protein. This suggests that microglia may be a new target in the progression of tauopathy therapies.

Therapeutic Approaches for Targeting the Aberrant Tau Protein

Small molecule-based therapies

Tau protein undergoes several PTMs which disruptions lead to their pathological conditions, including tauopathy and AD. Increased activity of kinases and/or lack of activity of phosphatases can lead to the hyperphosphorylation of tau protein, one of the major causes of tauopathies. In the cytosol, hyperphosphorylated tau protein forms oligomers, or is self-assembled into PHFs or straight filaments, disrupting microtubule stability. Therefore, the majority of the initial treatments addressing tauopathies were based on small molecules inhibiting kinases, tau protein aggregations, PTMs, and/or stabilization of microtubules. However, these therapies have been turned down due to their toxicity, poor efficacy, or inability to cross the blood-brain barrier (BBB) (Soeda and Takashima, 2020; Wang et al., 2020). Considering the role of protein kinases in tau protein hyperphosphorylation, several kinase inhibitors targeting glycogen synthase kinase-3 beta (GSK-3β), cyclin-dependent kinase 5, extracellular signal-regulated kinase 2, and tyrosine kinase Fyn, are currently under investigation (Yadikar et al., 2020). In clinical trials, the GSK-3β inhibitor tideglusib has been shown to be well tolerated in AD patients, and lithium treatment was found to be beneficial in improving the cognitive performance of patients with mild cognitive impairment and AD (Snitow et al., 2021). Although the tideglusib has entered clinical trials, phase 2 didn't yield any positive therapeutic outcomes, whereas for lithium, clinical trials are ongoing, and patients are being recruited for phase 3 studies. Saracatinib, a Fyn kinase inhibitor, did not have a significant effect on slowing down disease progression in AD patients (Nygaard, 2018). CP681301, a cyclin-dependent kinase 5 inhibitor, increased tau protein phosphorylation in transgenic mice overexpressing p25 and also diminished the activity of GSK-3β by indirect phosphorylation at S9 (Soeda and Takashima, 2020). Nonetheless, human clinical trials using kinase inhibitors have not produced encouraging outcomes in reducing the emergence of cognitive deficits.

In addition to the inhibition of kinases, many small-molecule therapies are deployed to improve the activities of phosphatases in order to combat the disease. Inactivation of various phosphatases is the most commonly observed pathological condition in AD and other tauopathies. The activity of PP2A is low in the AD brain (Wei et al., 2020). PP2A is responsible for around 70 percent of the overall tau protein phosphatase activity in the brain, suggesting that decreased PP2A activity in AD is responsible for the accumulation of hyperphosphorylated tau protein in the brain. Studies demonstrated that activation of PP2A by sodium selenate lowered tau protein phosphorylation, and reversed memory deficits in animal models of tauopathies (Ahmed et al., 2020). However, due to selenium buildup in the CNS, sodium selenate therapy failed in phase 2a clinical studies (Cardoso et al., 2019). Other compounds such as memantine, apolipoprotein E mimetic (COG112), and resveratrol are known to attenuate the phosphorylation of tau protein by interfering with the activity of PP2A (Taleski and Sontag. 2018). Moreover, tau protein is primarily acetylated in the fibril-forming core, which encourages stacking of the beta strand and decreases the solubility of tau protein. Salsalate, a nonsteroidal anti-inflammatory drug reduced the amount of tau protein acetylation at K174, preventing hippocampus atrophy and memory loss in PS19 transgenic (PS19Tg) mice, a model of tauopathy (Min et al., 2015). Salsalate, however, had no therapeutic impact in phase 1 open-label clinical trial with patients who had been diagnosed with Richardson's syndrome (VandeVrede et al., 2020). The outcome has not been made available to the public. O-linked-N-acetylglucosamine (O-GlcNAc) acylation is a PTM in which O-GlcNAc transferase adds O-GlcNAc to serine and threonine residues of nucleus and cytoplasmic proteins. O-GlcNAcase is responsible for the removal of O-GlcNAc. It has been demonstrated that glycosylation of tau protein by O-GlcNAc prevents tauopathies, thus, inhibiting O-GlcNAcase activity in principle should prevent AD/tauopathies. As a result, reducing tau protein deglycosylation may be achieved by inhibiting the O-GlcNAcase enzyme. MK-8719, an O-GlcNAcase inhibitor, increased tau protein glycosylation levels in rTg4510 mouse models, reduced aberrant tau protein pathology, and slowed down brain atrophy (Wang et al., 2020). MK-8719 and ASN120290 were given orphan drug designations and have been used for the treatment of PSP; however, no clinical trials have been reported. A significant amount of research is also being carried out to target ubiquitination, sumoylation, and other PTMs (Alquezar et al., 2021).

The autophagy-lysosome pathway and the ubiquitin proteasomal system are the two main mechanistic pathways for the degradation of abnormal proteins to maintain cellular homeostasis, and aberrant tau protein accumulation in the brain impairs their ability to function. So, it stands to reason to stimulate these pathways in order to deal with tau protein accumulation. Protein breakdown by the lysosome or proteasome is accelerated by ubiquitination. A reduced level of phosphorylated tau protein and higher tau protein clearance were seen in the P301L mouse model after, a de-ubiquinating enzyme, USP13, was knocked down (Liu et al., 2019). Rapamycin, an mTOR kinase inhibitor is tested in phase 1 and phase 2 clinical trials for patients with mild cognitive impairment and early-onset AD (Silva et al., 2020). However, other mTOR kinase inhibitors, namely AZD2014, OSI-027, and AZD8055, showed greater efficacy compared to rapamycin in neuronal cell models from patient's induced pluripotent stem cells (Silva et al., 2020).

Tau protein-aggregates and oligomers (Oakley et al., 2020) are one of the main components in AD and other tauopathies. Misfolded and oligomerized tau protein can gradually overload the nerve cells, resulting in neuronal cell death. Curcumin, a major component of turmeric has been shown to inhibit tau protein aggregation (Bijari et al., 2018). Nevertheless, due to its poor bioavailability and low stability in the body, curcumin failed to demonstrate significant clinical efficacy (Chen et al., 2018). However, clinical trial results of its analogs with improved bioavailability are still awaited (Okuda et al., 2017; Cascio et al., 2019). The anti-inflammatory compound, resveratrol has also been shown to reduce the level of hyperphosphorylated tau protein in mice (Yu et al., 2018). Though the molecule possesses additional benefits like antioxidant and anti-inflammatory properties, its low bioavailability and rapid metabolism (Chimento et al., 2019) are its major shortcomings. Interestingly, nickel administration, either in its metallic form or in the form of its conjugate with morpholine, inhibited tau protein aggregation *in vitro* (Gorantla et al., 2020). However, nickel deposition in the liver, kidney, and spleen is a matter of concern. An *in vitro* experiment revealed that folic acid prevents tau protein aggregation and reduces its phosphorylation by stabilizing its native conformation and regulating PP2A methylation (Ghasemzadeh and Riazi, 2020), respectively. Furthermore, methylene blue has been shown to prevent tau protein aggregation and improve memory impairment (deficit) in a mouse model of tauopathy (Soeda et al., 2019). However, a derivative of methylene blue, LMTX (TRx0237), failed to show any substantial improvement in phase 3 clinical trials (Gauthier et al., 2016). ACI-3024 is a small-molecule tau protein aggregation inhibitor. In cell-based experiments, ACI-3024 was shown to reduce intracellular pathogenic tau protein in a dose-dependent manner, as well as prevent microglial activation and neuronal death caused by tau PHF. Consistent with cell-based study, ACI-3024 treatment was claimed

Downloaded from http://journals.lww.com/nrronline by BhDMf5ePHKav1zEoum1tQfN4a+kJLhEZgbsIHo4XMi0hCywCX1AWnYQp/IlQrHD3i3D0OdRyi7TvSFl4Cf3VC4/OAVpDDa8KKGKV0Ymy+78= on 11/03/2023

Downloaded from http://journals.lww.com/nrronline by BhDMf5ePHKav1zEoum1tQfN4a+kJLhEZgbsIHo4XMi0hCywCX1AWnYQp/IlQrHD3i3D0OdRyi7TvSFl4Cf3VC4/OAVpDDa8KKGKV0Ymy+78= on 11/03/2023

to lower aggregated and insoluble tau protein in rTg4510 mice. In these mice, treatment apparently decreased cerebrospinal fluid total tau protein and microglia activation in inverse relation to plasma drug levels. A phase 1 clinical study was carried out on healthy individuals, but there is no registry registration or published research to be discovered (Imbimbo et al., 2022). CLR01, a member of the broad-spectrum protein aggregation inhibitors called molecular tweezers was found to lessen tau protein aggregation, hyperphosphorylation, and oligomerization in P301S-tau mice (Di et al., 2021).

Reduced binding affinity caused by hyperphosphorylation of tau protein in tauopathies destabilizes the microtubules, which in turn causes axonal deficiency and tau pathology. Paclitaxel was the first microtubule stabilizer to be researched, and the focus of that study was on its anti-tumor capabilities. Its further study was put on hold due to the low quality of crossover BBB. Several microtubule-stabilizing compounds, such as epothilone D and TPI 287 have been tested in animal models of tauopathy (Miller and Das, 2020) and showed promising results in clinical trials. However, the clinical trials of epithilone D and TPI 287 were discontinued due to the adverse side effects and lack of tolerance in patients with AD (Tsai et al., 2020), respectively. Triazolo-pyrimidines CNDR-51657, another potential anti-cancer agent, when given orally to 9-month-old female PS19Tg mice, attenuated tau pathology, axonal dystrophy, and microtubule abnormalities (Zhang et al., 2018). No negative consequences were noted. Triazolo-pyrimidines interact with tubulin and their strong oral bioavailability and brain penetration, making them strong candidates for microtubule stabilizers (Oukoloff et al., 2021).

Acetylcholinesterase inhibitors like donepezil, rivastigmine, galantamine, and N-methyl D-aspartate receptor-like memantine have been shown to improve cognitive decline in AD patients. Few studies also reported the beneficial effect of acetylcholinesterase inhibitors in preclinical models of tauopathies (Yoshiyama et al., 2010; Riedel et al., 2020). For instance, a study by Yoshiyama and coworkers (Yoshiyama et al., 2010) reported that donepezil treatment ameliorated neuroinflammation, tau protein pathology, and neurodegeneration in PS19Tg mice. Acetylcholinesterase inhibitors alleviate the levels of synaptic acetylcholine, improve the interaction within cholinergic neurons, and delay the progression of diseases. In spite of having a positive impact on AD patients, these molecules are often associated with significant side effects (Sharma, 2019). Neurotrophic factor supplementation, including brain-derived neurotrophic factor (Giuffrida et al., 2018), ciliary neurotrophic factor (Wei et al., 2021), and nerve growth factor (Cuello et al., 2019), have been found to ameliorate deficits in neurogenesis, synaptic plasticity, and cognition. These approaches, however, have been slowed by concerns about the specificity, limited BBB permeability, and potential side effects of the available drugs in animal studies (Thoenen and Sendtner, 2002). Although histone deacetylase inhibitors have shown neuroprotective roles (Lee et al., 2018), successful treatments in the animal model are rarely translated into clinical trials.

Several small molecules inhibitors are developed to target the PTMs of tau protein, but very few have entered clinical trials, notably, sodium selenate, BPN14770, rapamycin, LMTX, curcumin, lithium, tideglusib, and saracatinib. These small molecules prevent phosphorylation of tau protein isoforms at tyrosine, serine, and threonine residues, acetylation at lysine 174, and aggregation by inhibiting the disulfide bond formation in oligomers. However, low stability, poor bioavailability, CNS toxicity, off-target effect, or lack of human efficacy constitutes the bottlenecks of small-molecule mode of therapeutic approaches. Multi-target drug candidates or combination therapy can be viable alternatives to combat tauopathies but can be hampered by the risk of drug-drug interactions and poor BBB permeability. Yet, it cannot be averted while developing treatments for challenging disorders, especially like AD. Novel formulation or drug delivery methods need to be adopted for the small-molecule therapeutics of tauopathies that lack brain exposure despite having efficacy. Growing research in nanodrug delivery methods can bring remarkable improvement in bioavailability and brain exposure of drug candidates against tauopathies.

Tau protein immunotherapies

Immunization is a desirable treatment strategy since it can boost the host immune system in response to disease or induce the generation of autoantibodies through active immunization. The same effect can also be produced by the administration of laboratory-made antibodies through passive immunization. Full-length human tau protein, shorter N-terminal region of tau protein, and phosphorylated tau protein, are available as immunogens. The effectiveness of an anti-tau protein vaccine called AV1980R/A with an adjuvant called AdvaxCpG has been evaluated in wild-type and transgenic mice. Interestingly, elevated levels of IgG in the brains of animals were observed upon immunization and showed significant improvement in memory function. Additionally, AV1980R/A showed a decrease in the level of hyperphosphorylated tau protein at Ser 396 with no effect on other phosphorylated tau protein forms in the vaccinated mouse brain (Joly-Amado et al., 2020). Several tau protein vaccines and antibodies are being tested in mouse models of tauopathy (Albert et al., 2019). AADvac1, a synthetic peptide that targets the MTBR of tau protein, was the first anti-tau protein vaccine to enter clinical trials. AADVac1 was proven to be risk-free in a phase 2 clinical study on patients with moderate AD dementia, but there was no clear evidence that it had any effect on cognitive tests (Novak et al., 2021). ACI-35, a liposomal vaccine, constituting tau protein fragments with several phosphorylated serine residues decreased soluble tau protein phosphorylation at S396 and induced antibody formation in P301L mice (Theunis et al., 2013). The phase 1b/2a clinical trial has shown that ACI-35 treatment is both safe and immunogenic in patients with mild to moderate AD dementia (No authors listed, 2022). Considering the pathogenicity of tau protein and the seeding of tau protein aggregation in naive cells, passive immunotherapies are emerging to target both extra and intracellular tau protein. JNJ-63733657 and UCB0107 antibodies binding to the N-terminus or C-terminus region and mid-domain region of phosphorylated tau protein inhibited cellular uptake of tau protein and efficiently prevented the seeding and propagation of the aberrant protein in naive or healthy cells (Ji and Sigurdsson, 2021). While JNJ-63733657 targets MTBR of tau protein, UCB0107 targets central region (amino acids 235-246) of tau protein. Studies demonstrated that targeting intracellular tau protein is more effective in reducing tau protein pathology in rTg4510 mouse model of tauopathy. The intrabodies, CP13i and PHF1i targeting intracellular tau protein, reduced tau protein pathology, whereas single-chain variable fragments, which target extracellular tau, had no therapeutic effect in the rTg4510 mice (Goodwin et al., 2021). However, the high molecular weight of antibodies is hindering their penetration into the cells (Soeda and Takashima, 2020). To address this limitation, anti-tau protein monoclonal antibodies that target intracellular tau protein have been developed. As of now, a significant number of antibodies have been developed and are being investigated at various stages of clinical trials. Monoclonal antibodies such as BIIB092, C2N-8E12, UCB0107, LY3303560, BIIB076, JNJ-63733657, Lu AF87908, PNT001, and RO7105705 are found to be well tolerated and safe as per phase 1 or 2 study data, and for a few, the trials are ongoing, and results are awaited (Soeda and Takashima, 2020). However, in two phase 2 clinical trials, the antibodies that bind to the N-terminal domain of tau protein (C2N-8E12, tilavonemab, and BIIB092; gosuranemab) were found to be ineffective for PSP patients (Höglinger et al., 2021; Jabbari and Duff, 2021). The binding occurs as expected, as evidenced by a decrease in cerebrospinal fluid-free tau protein and an increase in tau protein in the lysosomes of newly developing perivascular vesicular astrocytes (Kim et al., 2021). Factors affecting the outcome of clinical trials are challenges associated with the earlier PSP diagnosis, particularly in preclinical stages, ineffective epitopes, and extracellular targeting [204]. A phase 2 clinical trial is being conducted to examine C2N-8E12 in AD patients.

The progression of AD is dependent on tau protein phosphorylation at S396/S404 and S422 residues; thus, antibodies such as Lu AF87908 and ACI-35 that can bind to the phosphorylation sites in a sequence-specific manner might be effective in preventing disease progression. However, RG7345, which targets phosphorylated tau protein in tauopathies, was discontinued from clinical development due to the unfavorable pharmacokinetics of the antibody (Sandusky-Beltran and Sigurdsson, 2020). Various tau protein immunotherapies with a wide array of advantages are currently under development. Although most of them succeed to some extent in addressing the advancement of AD and related tauopathies, tau immunotherapies do have a few roadblocks. Expensive production, a short half-life, systemic administration, and the accumulation and formation of anti-antibodies causing undesirable immunological side effects together contributed to the failure of immunotherapies as effective therapeutic tools and hence a parallel need for alternative therapies.

Immune-related adverse anomalies are one of the considerable side effects of tau immunotherapies. To circumvent this issue, prior establishment of a safety profile needs to be done to achieve prophylactic immunotherapy for tauopathy. While developing antibodies for inducing passive immunization, specificity of antibodies and deliverability to the brain should be considered as highly preferable features. Efficient new techniques that can deliver the designed antibodies into the target brain region in sufficient quantities should be developed which can render immunotherapy a successful therapeutic tool to treat tauopathies.

Oligonucleotide based therapies

Theoretically, gene therapy delivers a transgene that either substitutes a defective gene or generally supports cells to combat the disease environment in order to treat a disease. But the selection of the vector, delivery system, dose optimization, host immune response, and vector interaction makes it a complex process. With the recent advancement in technology, oligonucleotide-based therapies like antisense oligonucleotide (ASO) technology, RNA interference (RNAi), and single-stranded catalytic DNA molecules (DNAzyme) are emerging as gene silencing approaches. These methods make use of short, single-stranded (15–20 nucleotide) oligonucleotides that bind to the corresponding RNA sequences based on Watson-Crick base pairing, leading to the eradication of the target RNA. As a major breakthrough, these have been used to address neurodegenerative diseases, and researchers are now targeting tau mRNA or RNA that interferes with the expression of tau mRNA (RNAi) to reduce the production of tau protein to stop their aggregation (**Figure 4**; DeVos et al. 2013; Xu et al. 2014). Although adeno-associated virus containing RNAi and ASOs therapies effectively reduce mRNA and protein expression and mitigate disease progression, they require direct injection into the brain and have been shown to have off-target effects. Given safety concerns with repeated injections, the risks may limit their use during AD and other tauopathies, which would then limit their therapeutic value.

ASOs

ASOs are short single-stranded DNA sequence ranging from 18–30 nucleotides in length (Scoles et al., 2019). They bind to the complementary mRNA target in a Watson-Crick- based hybridization manner. Upon binding, they modify the expression of target mRNA in an RNase H dependent and independent

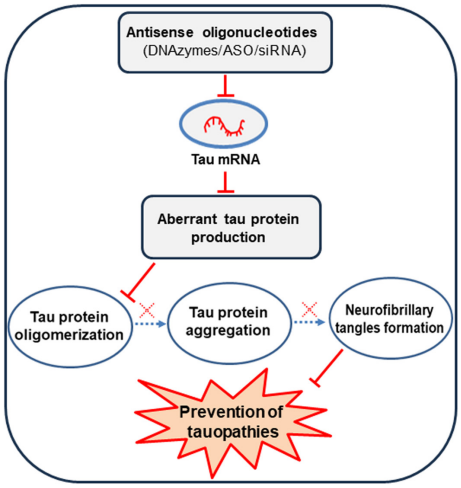


Figure 4 | Overview of oligonucleotide-based tau therapies for the treatment of tauopathies.
The accumulation of aberrant tau protein may be minimized by inhibiting the tau gene expression using anti-sense oligonucleotides (DNAzyme, ASOs, siRNA) against the mRNA from the MAPT gene. Created with Microsoft PowerPoint. ASOs: Antisense oligonucleotides; DNAzymes: single-stranded catalytic DNA molecules; MAPT: microtubule-associated protein tau; siRNA: Small interfering RNA.

manner. The RNase H enzyme detects DNA-RNA complexes and eventually cleaves and degrades the target mRNA from the complexes. However, the activation of RNase H upon its hybridization with ASO is highly dependent on the sequence specificity of ASO towards target RNA. On the other hand, the RNase H independent ASOs act as steric blockers and modulate the processing events of mRNA such as splicing, 5' cap addition, or 3' end polyadenylation (Roberts et al., 2020). The ASOs are subject to nuclease degradation in biological systems in unmodified form. Several modifications have been made to enhance the biological effectiveness of target therapy. Alterations in the backbone of ASOs using phosphorothioate have proven to improve the susceptibility to degradation by nuclease while retaining the cleavage activity (Jadhav et al., 2019). Further modification of ASOs through the incorporation of a sugar moiety at the 2' position not only enhanced their potency and nuclease resistance but also remarkably improved their toxicity profile, which makes ASOs feasible for therapeutic purposes (Anderson et al., 2021). In SH-SY5Y cells, ASO targeting the exon 4 of the *MAPT* gene reduced tau RNA expression and protein levels by 96% and 74% respectively (Chakravarthy et al., 2020). ASO has also been shown to decrease tau mRNA and protein in PS19Tg mice with a reversal in tau protein pathology (DeVos et al., 2017). The ASO, IONIS-MAPT Rx, is currently being tested against tauopathy and showed competitive binding to pre-mRNA of MAPT gene through intron 9. In the phase 1 clinical trials IONIS-MAPT Rx delivered intrathecally in patients with mild AD appeared to be safe (NCT03186989), and it did reduce both total tau protein and phospho tau-181 levels in the cerebrospinal fluid by more than 50 percent from baseline (Mummery et al., 2023). This finding endorses the ASO translatability to humans and suggested that tau-lowering therapy is well-tolerated and safe. However, further studies are required to evaluate the therapeutic effect of ASOs in AD and other tauopathies.

Small interfering RNA
Small interfering RNAs (siRNAs) are short non-coding double-stranded RNA (dsRNA) sequences ranging from 19–21 nucleotides in length, with a 2 nucleotide (TT or UU) overhang on the 3' end of either strand (Alshaer et al., 2021). They show high specificity towards target RNA and degrade the RNA in RNase H dependent manner. They emerge as a drug-targeting tool by modulating the expression of targeted genes through transcriptional gene silencing. When it enters the cell, the dsRNA is recognized by either the Dicer-2 or Argonaute-2 enzymes to form the RNA-inducing silencing complex and result in the degradation of the passenger strand. The remaining siRNA then binds to the complementary RNA and stimulates its degradation (Alshaer et al., 2021). Hoglinger and coworkers have shown that siRNA against *MAPT* (AccellTM SMART pool), in the absence of carrier molecule, was able to inhibit the expression of human tau protein in primary cortical neurons of PS19 transgenic mice expressing P301S mutant form of human tau protein. Moreover, neither neurotoxicity, neuroinflammation, nor apoptosis was observed during the presence of MAPT siRNA in the mouse hippocampus (Xu et al., 2014). However, siRNA shortcomings are its limited BBB permeability, cellular uptake, and susceptibility to nuclease-mediated degradation (Dana et al., 2017). Several studies have shown that direct delivery of adeno-associated virus containing RNAi molecules to the CNS can effectively reduce tau protein production and slow or prevent AD progression in mouse models of AD (Ising et al., 2017). Given safety concerns with repeated CNS injections, the risks may limit their use in the course of tauopathy, which would then limit their therapeutic value.

DNAzymes
Breaker and Joyce in 1994, using *in-vitro* selection methods, discovered a population of single-stranded catalytic DNA (DNAzyme) that could cleave target RNA in a cations-dependent manner (Breaker and Joyce, 1994). Since then, DNAzyme has emerged as a molecular tool in gene therapy, and many investigators globally have explored the pertinent growing scope of DNA catalytic activity, including its applications (Wang et al., 2021). DNAzyme, also known as a single-stranded catalytic DNA molecule, is an RNA-cleaving DNA enzyme. The structure of a DNAzyme comprises three key components, two variable flanking domains or binding arms that are bound to a central catalytic domain with RNA endonuclease activity composed of a specific sequence of 15 or 13 deoxynucleotides. Through Watson-Crick base hybridization, the binding arms help bind the DNAzyme to a specific target mRNA. DNAzyme is a metal-dependent enzyme. It uses a metal ion such as Pb^{2+}, Mg^{2+}, Zn^{2+}, or Cu^{2+} to polarize a water molecule, catalyzing the hydrolysis of the phosphodiester bond in an RNA molecule. The mRNA cleavage products are further degraded by intracellular RNases, whereas the DNAzyme is then available for a next cycle of catalytic action. With this high turnover cycle, a relatively low concentration of DNAzyme is sufficient to considerably reduce levels of specific disease-causing proteins, thus constituting an effective gene therapy approach. Presently, 10–23 and 8–17 designs of DNAzymes are known to catalyze the efficient cleavage of RNA into degraded products. DNAzymes can be engineered to function in an allele-specific manner to selectively cleave an RNA molecule having a point mutation. Thus, the cleavage of diseased or abnormal mRNAs diminishes the level of disease-causing proteins without interfering with the expression of the wild-type protein (Ma and Liu, 2020). The backbone modification or the addition of a sugar moiety to DNAzyme molecule has been shown to substantially improve its stability and potency in serum (Wang et al., 2021). Structural fine tuning of DNAzyme has improved its half-life, resistance against endogenous nucleases, and most importantly, affinity towards the target complementary sequences (Wang et al., 2021). The distribution, clearance, and safety of DNAzyme have recently been described. (Hallett et al., 2013). In this study, Hallett et al. (2013) have shown that DNAzymes are stable, can cross the BBB, and can be delivered repeatedly by systemic injections, thus avoiding the need for direct brain injections. The therapeutic utility of DNAzyme is further supported by clinical trials of DNAzyme targeting c-jun (Grassi and Grassi, 2013), EBV LMP1 protein (Cao et al., 2014), and GATA-3 (Cho et al., 2013) that have shown that it is safe and well-tolerated.

The molecular events contributing to neuronal death or degeneration are not yet clear and are under investigation. Evidence suggests that there is a close relationship between AD and the level of β-secretase (also known as BACE1 or β-site APP cleaving enzyme 1). BACE1 is an aspartyl protease that contributes to the development of Aβ plaque, the pathophysiological hallmark of AD. DNAzyme generated against BACE mRNA has been shown to reduce BACE gene expression by 50–70% in the HEK293 cells (Nawrot, 2004). In SH-SY5Y cells, DNAzyme (RNV563) targeting MAPT exon 13 reduced MAPT mRNA expression by more than 50% (Chakravarthy et al., 2020). The finding from these preclinical studies encourages more comprehensive and rigorous studies to examine the distribution, safety, and therapeutic benefit of DNAzyme, which is essential for clinical translation and the expansion of therapeutic applications for tauopathies and several other neurodegenerative diseases.

Oligonucleotide-based therapies have made astonishing advances over the last few years with the approval of ASO and with promising developments of several potential therapeutic agents for neurodegenerative diseases. However, off-target effect and effective delivery of oligonucleotides to its site of action remains a major issue. Efficient oligonucleotides with improved nuclease resistance, susceptibility to degradation, and toxicity profile need to be developed to make the treatment method more feasible and safer. Several strategies including structural modification of the oligonucleotide, utilization of various lipid or polymeric nanocarriers, tagging of oligonucleotides to target receptors, and use of small molecules can be adopted to enhance oligonucleotide effectiveness.

Conclusion and Future Perspectives

Tauopathies including AD are a group of neurodegenerative diseases that are characterized by an abnormal accumulation of hyperphosphorylated tau protein in the brain cells and progressive cognitive decline. Despite gaining considerable knowledge about the pathological mechanisms of tauopathies over the last several decades, we know little about how to stop or slow down the ongoing neurodegenerative process. In the search for effective disease-modifying intervention for tauopathies, several biological targets have been and continue to be determined. Considering the strong association between pathogenic tau protein accumulation and cognitive deficits along with genetic evidence that tau mutations cause progressive neurodegeneration, there has been a lot of interest in developing anti-tau RNA/protein therapies as a means to treat AD and related tauopathies. The majority of the tau-based therapies are targeting either the gain of toxic function of tau protein by aiming PTMs and tau protein aggregation, or the loss of function of tau protein by aiming microtubule instability. Each approach has its advantages and shortcomings. A common limitation that can be seen in these approaches is either poor BBB permeability and/or off-target effect. Thus, new therapeutic strategies to downregulate the production of pathological tau aggregates, and deliver the therapeutic agent to CNS, with tolerable toxicity will signify promising interventions in treating tauopathies. Given that tau pathology may differ across tauopathies, a realistic therapeutic approach may be to regulate the tau transcripts instead of tau protein, which would rule out the possibility of tau protein toxicity. The current advancements in gene therapy, genome

Downloaded from http://journals.lww.com/nrronline by BhDMf5ePHKav1zEoum1tQfN4a+kJLhEZgbsIHo4XMi0hCywCX1AWnYQp/IlQrHD3i3D0OdRyi7TvSFl4Cf3VC4/OAVpDDa8KKGKV0Ymy+78= on 11/03/2023

Downloaded from http://journals.lww.com/nrronline by BhDMf5ePHKav1zEoum1tQfN4a+kJLhEZgbsIHo4XMi0hCywCX1AWnYQp/IlQrHD3i3D0OdRyi7TvSFl4Cf3VC4/OAVpDDa8KKGKV0Ymy+78= on 11/03/2023

editing techniques, and oligonucleotide-based therapies are opening the doors to a better understanding of the cellular mechanisms of tauopathy and the development of new therapies for these complex disorders. To this end, ASO, RNAi, and DNAzymes represent potential therapeutic approaches that target tau mRNA to regulate its accumulation and toxicity during pathological conditions **(Figure 4**). While ASOs and DNAzymes have been investigated for several neural and extraneural diseases, there has been considerable interest in testing the therapeutic benefits of ASOs and DNAzyme in tauopathies. Considering the positive outcome of tau-targeting ASO MAPTRx in AD, more preclinical studies are required to determine the therapeutic benefits of these tau-lowering gene therapies for other primary tauopathies. Moreover, attempts to improve the BBB penetrability of ASOs or DNAzyme are required to enhance the effectiveness of these therapies which may possibly persuade the approach to be a much more competent option to treat neurodegenerative disorders in the future. Although gene therapies including ASOs and DNAzyme are safe and effective in reducing mRNA expression of target genes, these are still in the initial stage of development and further comprehensive studies will be needed for evaluation of these approaches in preclinical models of tauopathies for clinical translation.

Author contributions: *HS searched the database, collected the information required to incorporate into the review article, drew the figures and wrote the first draft of manuscript. AD reviewed and edited the draft. TP and MMK provided the conceptualization of the topic, edited the final draft, and critically reviewed the article. All authors read and approved the final manuscript.*
Conflicts of interest: *The authors declare no conflicts of interest.*
Data availability statement: *Not applicable.*
Open access statement: *This is an open access journal, and articles are distributed under the terms of the Creative Commons AttributionNonCommercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.*

References

Ahmed T, Van der Jeugd A, Caillierez R, Buée L, Blum D, D'hooge R, Balschun D (2020) Chronic sodium selenate treatment restores deficits in cognition and synaptic plasticity in a murine model of tauopathy. Front Mol Neurosci 13:570223.

Albert M, Mairet-Coello G, Danis C, Lieger S, Caillierez R, Carrier S, Skrobala E, Landrieu I, Michel A, Schmitt M (2019) Prevention of tau seeding and propagation by immunotherapy with a central tau epitope antibody. Brain 142:1736-1750.

Alonso AdC, Grundke-Iqbal I, Barra HS, Iqbal K (1997) Abnormal phosphorylation of tau and the mechanism of Alzheimer neurofibrillary degeneration: sequestration of microtubule-associated proteins 1 and 2 and the disassembly of microtubules by the abnormal tau. Proc Natl Acad Sci U S A 94:298-303.

Alquezar C, Arya S, Kao AW (2021) Tau post-translational modifications: dynamic transformers of tau function, degradation, and aggregation. Front Neurol 11:595532.

Alshaer W, Zureigat H, Al Karaki A, Al-Kadash A, Gharaibeh L, Ma'mon MH, Aljabali AA, Awidi A (2021) siRNA: mechanism of action, challenges, and therapeutic approaches. Eur J Pharmacol 905:174178.

Anderson BA, Freestone GC, Low A, De-Hoyos CL, Iii WJD, Østergaard ME, Migawa MT, Fazio M, Wan WB, Berdeja A (2021) Towards next generation antisense oligonucleotides: mesylphosphoramidate modification improves therapeutic index and duration of effect of gapmer antisense oligonucleotides. Nucleic Acids Res 49:9026-9041.

Asai H, Ikezu S, Tsunoda S, Medalla M, Luebke J, Haydar T, Wolozin B, Butovsky O, Kügler S, Ikezu T (2015) Depletion of microglia and inhibition of exosome synthesis halt tau propagation. Nat Neurosci 18:1584-1593.

Betters RK, Luhmann E, Gottschalk AC, Xu Z, Shin MR, Ptak CP, Fiock KL, Radoshevich LC, Hefti MM (2023) Characterization of the tau interactome in human brain reveals isoform-dependent interaction with 14-3-3 family proteins. eNeuro 10. doi: 10.1523/ENEURO.0503-22.2023.

Bieniek KF, Blessing MM, Heckman MG, Diehl NN, Serie AM, Paolini MA, Boeve BF, Savica R, Reichard RR, Dickson DW (2020) Association between contact sports participation and chronic traumatic encephalopathy: a retrospective cohort study. Brain Pathol 30:63-74.

Bijari N, Balalaie S, Akbari V, Golmohammadi F, Moradi S, Adibi H, Khodarahmi R (2018) Effective suppression of the modified PHF6 peptide/1N4R Tau amyloid aggregation by intact curcumin, not its degradation products: another evidence for the pigment as preventive/therapeutic "functional food". Int J Biol Macromol 120:1009-1022.

Borroni B, Graff C, Hardiman O, Ludolph AC, Moreno F, Otto M, Piccininni M, Remes AM, Rowe JB, Seelaar H (2022) FRONTotemporal dementia Incidence European Research Study—FRONTIERS: rationale and design. Alzheimers Dement 18:498-506.

Breaker RR, Joyce GF (1994) A DNA enzyme that cleaves RNA. Chem Biol 1:223-229.

Brunello CA, Merezhko M, Uronen R-L, Huttunen HJ (2020) Mechanisms of secretion and spreading of pathological tau protein. Cell Mol Life Sci 77:1721-1744.

Cao Y, Yang L, Jiang W, Wang X, Liao W, Tan G, Liao Y, Qiu Y, Feng D, Tang F (2014) Therapeutic evaluation of Epstein-Barr virus-encoded latent membrane protein-1 targeted DNAzyme for treating of nasopharyngeal carcinomas. Mol Ther 22:371-377.

Cardoso BR, Roberts BR, Malpas CB, Vivash L, Genc S, Saling MM, Desmond P, Steward C, Hicks RJ, Callahan J (2019) Supranutritional sodium selenate supplementation delivers selenium to the central nervous system: results from a randomized controlled pilot trial in Alzheimer's disease. Neurotherapeutics 16:192-202.

Cascio FL, Puangmalai N, Ellsworth A, Bucchieri F, Pace A, Piccionello AP, Kayed R (2019) Toxic tau oligomers modulated by novel curcumin derivatives. Sci Rep 9:1-16.

Chakravarthy M, Chen S, Wang T, Veedu RN (2020) Development of novel chemically-modified nucleic acid molecules for efficient inhibition of human MAPT gene expression. Genes (Basel) 11:667.

Chen M, Du ZY, Zheng X, Li DL, Zhou RP, Zhang K (2018) Use of curcumin in diagnosis, prevention, and treatment of Alzheimer's disease. Neural Regen Res 13:742.

Chimento A, De Amicis F, Sirianni R, Sinicropi MS, Puoci F, Casaburi I, Saturnino C, Pezzi V (2019) Progress to improve oral bioavailability and beneficial effects of resveratrol. Int J Mol Sci 20:1381.

Cho EA, Moloney FJ, Cai H, Au-Yeung A, China C, Scolyer RA, Yosufi B, Raftery MJ, Deng JZ, Morton SW (2013) Safety and tolerability of an intratumorally injected DNAzyme, Dz13, in patients with nodular basal-cell carcinoma: a phase 1 first-in-human trial (DISCOVER). Lancet 381:1835-1843.

Chung DC, Roemer S, Petrucelli L, Dickson DW (2021) Cellular and pathological heterogeneity of primary tauopathies. Mol Neurodegener 16:57.

Cuello AC, Pentz R, Hall H (2019) The brain NGF metabolic pathway in health and in Alzheimer's pathology. Front Neurosci 13:62.

Dana H, Chalbatani GM, Mahmoodzadeh H, Karimloo R, Rezaiean O, Moradzadeh A, Mehmandoost N, Moazzen F, Mazraeh A, Marmari V (2017) Molecular mechanisms and biological functions of siRNA. Int J Biomed Sci 13:48.

DeVos SL, Goncharoff DK, Chen G, Kebodeaux CS, Yamada K, Stewart FR, Schuler DR, Maloney SE, Wozniak DF, Rigo F (2013) Antisense reduction of tau in adult mice protects against seizures. J Neurosci 33:12887-12897.

DeVos SL, Miller RL, Schoch KM, Holmes BB, Kebodeaux CS, Wegener AJ, Chen G, Shen T, Tran H, Nichols B (2017) Tau reduction prevents neuronal loss and reverses pathological tau deposition and seeding in mice with tauopathy. Sci Transl Med 9:eaag0481.

Di J, Siddique I, Li Z, Malki G, Hornung S, Dutta S, Hurst I, Ishaaya E, Wang A, Tu S, Boghos A, Ericsson I, Klärner FG, Schrader T, Bitan G (2021) The molecular tweezer CLR01 improves behavioral deficits and reduces tau pathology in P301S-tau transgenic mice. Alzheimers Res Ther 13:6. Erratum in: Alzheimers Res Ther 13:88.

Dugger BN, Whiteside CM, Maarouf CL, Walker DG, Beach TG, Sue LI, Garcia A, Dunckley T, Meechoovet B, Reiman EM (2016) The presence of select tau species in human peripheral tissues and their relation to Alzheimer's disease. J Alzheimers Dis 51:345-356.

Duyckaerts C, Clavaguera F, Potier MC (2019) The prion-like propagation hypothesis in Alzheimer's and Parkinson's disease. Curr Opin Neurol 32:266-271.

El-Hayek YH, Wiley RE, Khoury CP, Daya RP, Ballard C, Evans AR, Karran M, Molinuevo JL, Norton M, Atri A (2019) Tip of the iceberg: assessing the global socioeconomic costs of Alzheimer's disease and related dementias and strategic implications for stakeholders. J Alzheimers Dis 70:323-341.

Fischer I, Baas PW (2020) Resurrecting the mysteries of big tau. Trends Neurosci 43:493-504.

Frontera JA, Boutajangout A, Masurkar AV, Betensky RA, Ge Y, Vedvyas A, Debure L, Moreira A, Lewis A, Huang J (2022) Comparison of serum neurodegenerative biomarkers among hospitalized COVID-19 patients versus non-COVID subjects with normal cognition, mild cognitive impairment, or Alzheimer's dementia. Alzheimers Dement 18:899-910.

Gauthier S, Feldman HH, Schneider LS, Wilcock GK, Frisoni GB, Hardlund JH, Moebius HJ, Bentham P, Kook KA, Wischik DJ (2016) Efficacy and safety of tau-aggregation inhibitor therapy in patients with mild or moderate Alzheimer's disease: a randomised, controlled, double-blind, parallel-arm, phase 3 trial. Lancet 388:2873-2884.

Ghasemzadeh S, Riazi GH (2020) Inhibition of Tau amyloid fibril formation by folic acid: in-vitro and theoretical studies. Int J Biol Macromol 154:1505-1516.

Giuffrida ML, Copani A, Rizzarelli E (2018) A promising connection between BDNF and Alzheimer's disease. Aging (Albany NY) 10:1791.

Goodwin MS, Sinyavskaya O, Burg F, O'Neal V, Ceballos-Diaz C, Cruz PE, Lewis J, Giasson BI, Davies P, Golde TE (2021) Anti-tau scFvs targeted to the cytoplasm or secretory pathway variably modify pathology and neurodegenerative phenotypes. Mol Ther 29:859-872.

Gorantla NV, Das R, Balaraman E, Chinnathambi S (2020) Transition metal nickel prevents Tau aggregation in Alzheimer's disease. Int J Biol Macromol 156:1359-1365.

Gordon MN, Heneka MT, Le Page LM, Limberger C, Morgan D, Tenner AJ, Terrando N, Willette AA, Willette SA (2022) Impact of COVID-19 on the onset and progression of Alzheimer's disease and related dementias: a roadmap for future research. Alzheimers Dement 18:1038-1046.

Grassi G, Grassi M (2013) First-in-human trial of Dz13 for nodular basal-cell carcinoma. Lancet 381:1797-1798.

Hallett MA, Dalal P, Sweatman TW, Pourmotabbed T (2013) The distribution, clearance, and safety of an anti-MMP-9 DNAzyme in normal and MMTV-PyMT transgenic mice. Nucleic Acid Ther 23:379-388.

Hefti MM, Kim S, Bell AJ, Betters RK, Fiock KL, Iida MA, Smalley ME, Farrell K, Fowkes ME, Crary JF (2019) Tau phosphorylation and aggregation in the developing human brain. J Neuropathol Exp Neurol 78:930-938.

Höglinger GU, Litvan I, Mendonca N, Wang D, Zheng H, Rendenbach-Mueller B, Lon H-K, Jin Z, Fisseha N, Budur K (2021) Safety and efficacy of tilavonemab in progressive supranuclear palsy: a phase 2, randomised, placebo-controlled trial. Lancet Neurol 20:182-192.

Hsieh YC, Guo C, Yalamanchili HK, Abreha M, Al-Ouran R, Li Y, Dammer EB, Lah JJ, Levey AI, Bennett DA (2019) Tau-mediated disruption of the spliceosome triggers cryptic RNA splicing and neurodegeneration in Alzheimer's disease. Cell Rep 29:301-316.

Imbimbo BP, Ippati S, Watling M, Balducci C (2022) A critical appraisal of tau-targeting therapies for primary and secondary tauopathies. Alzheimers Dement 18:1008-1037.

Downloaded from http://journals.lww.com/nrronline by BhDMf5ePHKav1zEoum1tQfN4a+kJLhEZgbsIHo4XMi0hCywCX1AWnYQp/IlQrHD3i3D0OdRyi7TvSFl4Cf3VC4/OAVpDDa8KKGKV0Ymy+78= on 11/03/2023

Ising C, Gallardo G, Leyns CE, Wong CH, Jiang H, Stewart F, Koscal LJ, Roh J, Robinson GO, Remolina Serrano J (2017) AAV-mediated expression of anti-tau scFvs decreases tau accumulation in a mouse model of tauopathy. J Exp Med 214:1227-1238.
Jabbari E, Duff KE (2021) Tau-targeting antibody therapies: too late, wrong epitope or wrong target? Nat Med 27:1341-1342.
Jadhav S, Avila J, Schöll M, Kovacs GG, Kövari E, Skrabana R, Evans LD, Kontsekova E, Malawska B, de Silva R (2019) A walk through tau therapeutic strategies. Acta Neuropathol Commun 7:22.
Ji C, Sigurdsson EM (2021) Current status of clinical trials on tau immunotherapies. Drugs 81:1135-1152.
Joly-Amado A, Davtyan H, Serraneau K, Jules P, Zitnyar A, Pressman E, Zagorski K, Antonyan T, Hovakimyan A, Paek H (2020) Active immunization with tau epitope in a mouse model of tauopathy induced strong antibody response together with improvement in short memory and pSer396-tau pathology. Neurobiol Dis 134:104636.
Kim B, Mikytuck B, Suh E, Gibbons GS, Van Deerlin VM, Vaishnavi SN, Spindler MA, Massimo L, Grossman M, Trojanowski JQ (2021) Tau immunotherapy is associated with glial responses in FTLD-tau. Acta Neuropathol 142:243-257.
Lee HY, Fan SJ, Huang FI, Chao HY, Hsu KC, Lin TE, Yeh TK, Lai MJ, Li YH, Huang HL (2018) 5-Aroylindoles act as selective histone deacetylase 6 inhibitors ameliorating Alzheimer's disease phenotypes. J Med Chem 61:7087-7102.
Leroy M, Bertoux M, Skrobala E, Mode E, Adnet-Bonte C, Le Ber I, Bombois S, Cassagnaud P, Chen Y, Deramecourt V, Lebert F, Mackowiak MA, Sillaire AR, Wathelet M, Pasquier F, Lebouvier T, Abied R, Adnet C, Barois A, Baude S, et al. (2021) Characteristics and progression of patients with frontotemporal dementia in a regional memory clinic network. Alzheimers Res Ther 13:19.
Liu X, Hebron ML, Mulki S, Wang C, Lekah E, Ferrante D, Shi W, Kurd-Misto B, Moussa C (2019) Ubiquitin specific protease 13 regulates tau accumulation and clearance in models of Alzheimer's disease. J Alzheimers Dis 72:425-441.
Ma L, Liu J (2020) Catalytic nucleic acids: biochemistry, chemical biology, biosensors, and nanotechnology. iScience 23:100815.
Miao J, Shi R, Li L, Chen F, Zhou Y, Tung YC, Hu W, Gong C-X, Iqbal K, Liu F (2019) Pathological tau from Alzheimer's brain induces site-specific hyperphosphorylation and SDS-and reducing agent-resistant aggregation of tau in vivo. Front Aging Neurosci 11:34.
Miller JH, Das V (2020) Potential for treatment of neurodegenerative diseases with natural products or synthetic compounds that stabilize microtubules. Curr Pharm Des 26:4362-4372.
Min SW, Chen X, Tracy TE, Li Y, Zhou Y, Wang C, Shirakawa K, Minami SS, Defensor E, Mok SA (2015) Critical role of acetylation in tau-mediated neurodegeneration and cognitive deficits. Nat Med 21:1154-1162.
Mirbaha H, Chen D, Morazova OA, Ruff KM, Sharma AM, Liu X, Goodarzi M, Pappu RV, Colby DW, Mirzaei H (2018) Inert and seed-competent tau monomers suggest structural origins of aggregation. Elife 7:e36584.
Mummery CJ, Borjesson-Hanson A, Blackburn DJ, Vijverberg EGB, De Deyn PP, Ducharme S, Jonsson M, Schneider A, Rinne JO, Ludolph AC, Bodenschatz R, Kordasiewicz H, Swayze EE, Fitzsimmons B, Mignon L, Moore KM, Yun C, Baumann T, Li D, Norris DA, et al. (2023) Tau-targeting antisense oligonucleotide MAPT(Rx) in mild Alzheimer's disease: a phase 1b, randomized, placebo-controlled trial. Nat Med 29:1437-1447.
Nawrot B (2004) Targeting BACE with small inhibitory nucleic acids-a future for Alzheimer's disease therapy? Acta Biochim Pol 51:431-444.
No authors listed (2022) Two new stabs at vaccinating people against pathologic tau. Available at: https://www.alzforum.org/news/conference-coverage/two-new-stabs-vaccinating-people-against-pathologic-tau. Accessed July 13, 2023.
Novak P, Kovacech B, Katina S, Schmidt R, Scheltens P, Kontsekova E, Ropele S, Fialova L, Kramberger M, Paulenka-Ivanovova N (2021) ADAMANT: a placebo-controlled randomized phase 2 study of AADvac1, an active immunotherapy against pathological tau in Alzheimer's disease. Nat Aging 1:521-534.
Nygaard HB (2018) Targeting Fyn kinase in Alzheimer's disease. Biol Psychiatry 83:369-376.
Oakley SS, Maina MB, Marshall KE, Al-Hilaly YK, Harrington CR, Wischik CM, Serpell LC (2020) Tau filament self-assembly and structure: tau as a therapeutic target. Front Neurol 11:590754.
Okuda M, Fujita Y, Hijikuro I, Wada M, Uemura T, Kobayashi Y, Waku T, Tanaka N, Nishimoto T, Izumi Y (2017) PE859, a novel curcumin derivative, inhibits amyloid-β and tau aggregation, and ameliorates cognitive dysfunction in senescence-accelerated mouse prone 8. J Alzheimers Dis 59:313-328.
Oukoloff K, Nzou G, Varricchio C, Lucero B, Alle T, Kovalevich J, Monti L, Cornec A-S, Yao Y, James MJ (2021) Evaluation of the structure–activity relationship of microtubule-targeting 1, 2, 4-triazolo [1, 5-a] pyrimidines identifies new candidates for neurodegenerative tauopathies. J Med Chem 64:1073-1102.
Polanco JC, Götz J (2022) Exosomal and vesicle-free tau seeds—propagation and convergence in endolysosomal permeabilization. FEBS J 289:6891-6907.
Reiken S, Sittenfeld L, Dridi H, Liu Y, Liu X, Marks AR (2022) Alzheimer's-like signaling in brains of COVID-19 patients. Alzheimers Dement 18:955-965.
Riedel G, Klein J, Niewiadomska G, Kondak C, Schwab K, Lauer D, Magbagbeolu M, Steczkowska M, Zadrozny M, Wydrych M, Cranston A, Melis V, Santos RX, Theuring F, Harrington CR, Wischik CM (2020) Mechanisms of anticholinesterase interference with tau aggregation inhibitor activity in a tau-transgenic mouse model. Curr Alzheimer Res 17:285-296.
Roberts TC, Langer R, Wood MJ (2020) Advances in oligonucleotide drug delivery. Nat Rev Drug Discov 19:673-694.
Sandusky-Beltran L, Sigurdsson E (2020) Tau immunotherapies: lessons learned, current status and future considerations. Neuropharmacology 175:108104.
Sayas CL, Ávila J (2021) GSK-3 and Tau: a key duet in Alzheimer's disease. Cells 10:721.
Scoles DR, Minikel EV, Pulst SM (2019) Antisense oligonucleotides: a primer. Neurol Genet 5:e323.
Sharma K (2019) Cholinesterase inhibitors as Alzheimer's therapeutics. Mol Med Rep 20:1479-1487.
Silva M, Nandi GA, Tentarelli S, Gurrell IK, Jamier T, Lucente D, Dickerson BC, Brown DG, Brandon NJ, Haggarty SJ (2020) Prolonged tau clearance and stress vulnerability rescue by pharmacological activation of autophagy in tauopathy neurons. Nat Commun 11:1-18.
Snitow ME, Bhansali RS, Klein PS (2021) Lithium and therapeutic targeting of GSK-3. Cells 10:255.
Soeda Y, Saito M, Maeda S, Ishida K, Nakamura A, Kojima S, Takashima A (2019) Methylene blue inhibits formation of tau fibrils but not of granular tau oligomers: a plausible key to understanding failure of a clinical trial for Alzheimer's disease. J Alzheimers Dis 68:1677-1686.
Soeda Y, Takashima A (2020) New insights into drug discovery targeting tau protein. Front Mol Neurosci. 13:590896.
Stang CD, Turcano P, Mielke MM, Josephs KA, Bower JH, Ahlskog JE, Boeve BF, Martin PR, Upadhyaya SG, Savica R (2020) Incidence and trends of progressive supranuclear palsy and corticobasal syndrome: a population-based study. J Parkinsons Dis 10:179-184.
Taleski G, Sontag E (2018) Protein phosphatase 2A and tau: an orchestrated 'Pas de Deux'. FEBS Lett 592:1079-1095.
Theunis C, Crespo-Biel N, Gafner V, Pihlgren M, López-Deber MP, Reis P, Hickman DT, Adolfsson O, Chuard N, Ndao DM (2013) Efficacy and safety of a liposome-based vaccine against protein Tau, assessed in tau. P301L mice that model tauopathy. PLoS One 8:e72301.
Thoenen H, Sendtner M (2002) Neurotrophins: from enthusiastic expectations through sobering experiences to rational therapeutic approaches. Nat Neurosci 5:1046-1050.
Tsai RM, Miller Z, Koestler M, Rojas JC, Ljubenkov PA, Rosen HJ, Rabinovici GD, Fagan AM, Cobigo Y, Brown JA (2020) Reactions to multiple ascending doses of the microtubule stabilizer TPI-287 in patients with Alzheimer disease, progressive supranuclear palsy, and corticobasal syndrome: a randomized clinical trial. JAMA Neurol 77:215-224.
Turcano P, Stang CD, Mielke MM, Martin PR, Upadhyaya SG, Josephs KA, Boeve BF, Knopman DS, Petersen RC, Savica R (2020) Incidence of frontotemporal disorders in Olmsted County: a population-based study. Alzheimers Dement 16:482-490.
VandeVrede L, Dale ML, Fields S, Frank M, Hare E, Heuer HW, Keith K, Koestler M, Ljubenkov PA, McDermott D (2020) Open-label phase 1 futility studies of salsalate and young plasma in progressive supranuclear palsy. Mov Disord Clin Pract 7:440-447.
Wang L, Kumar R, Pavlov PF, Winblad B (2020) Small molecule therapeutics for tauopathy in Alzheimer's disease: walking on the path of most resistance. Eur J Med Chem 209:112915.
Wang X, Li W, Marcus J, Pearson M, Song L, Smith K, Terracina G, Lee J, Hong KLK, Lu SX (2020) MK-8719, a novel and selective O-GlcNAcase inhibitor that reduces the formation of pathological tau and ameliorates neurodegeneration in a mouse model of tauopathy. J Pharmacol Exp Ther 374:252-263.
Wang Y, Mandelkow E (2016) Tau in physiology and pathology. Nat Rev Neurosci 17:22-35.
Wang Y, Nguyen K, Spitale RC, Chaput JC (2021) A biologically stable DNAzyme that efficiently silences gene expression in cells. Nat Chem 13:319-326.
Wei H, Zhang HL, Wang XC, Xie JZ, An DD, Wan L, Wang JZ, Zeng Y, Shu XJ, Westermarck J, Lu YM, Ohlmeyer M, Liu R (2020) Direct activation of protein phosphatase 2A (PP2A) by tricyclic sulfonamides ameliorates Alzheimer's disease pathogenesis in cell and animal models. Neurotherapeutics 17:1087-1103.
Wei W, Liu Y, Dai CL, Baazaoui N, Tung YC, Liu F, Iqbal K (2021) Neurotrophic treatment initiated during early postnatal development prevents the Alzheimer-like behavior and synaptic dysfunction. J Alzheimers Dis 82:631-646.
Wurm R, Klotz S, Rahimi J, Katzenschlager R, Lindeck-Pozza E, Regelsberger G, Danics K, Kapas I, Bíró Z, Stögmann E (2020) Argyrophilic grain disease in individuals younger than 75 years: clinical variability in an under-recognized limbic tauopathy. Eur J Neurol 27:1856-1866.
Xu H, Rösler TW, Carlsson T, de Andrade A, Fiala O, Hollerhage M, Oertel WH, Goedert M, Aigner A, Höglinger GU (2014) Tau silencing by siRNA in the P301S mouse model of tauopathy. Curr Gene Ther 14:343-351.
Yadikar H, Torres I, Aiello G, Kurup M, Yang Z, Lin F, Kobeissy F, Yost R, Wang KK (2020) Screening of tau protein kinase inhibitors in a tauopathy-relevant cell-based model of tau hyperphosphorylation and oligomerization. PLoS One 15:e0224952.
Yoshiyama Y, Kojima A, Ishikawa C, Arai K (2010) Anti-inflammatory action of donepezil ameliorates tau pathology, synaptic loss, and neurodegeneration in a tauopathy mouse model. J Alzheimers Dis 22:295-306.
You G, Yao J, Liu Q, Li N (2022) The strategies for treating "Alzheimer's disease": insulin signaling may be a feasible target. Curr Issues Mol Biol 44:6172-6188.
Yu KC, Kwan P, Cheung SK, Ho A, Baum L (2018) Effects of resveratrol and morin on insoluble tau in tau transgenic mice. Transl Neurosci 9:54-60.
Zhang B, Yao Y, Cornec AS, Oukoloff K, James MJ, Koivula P, Trojanowski JQ, Smith AB, Lee VMY, Ballatore C (2018) A brain-penetrant triazolopyrimidine enhances microtubule-stability, reduces axonal dysfunction and decreases tau pathology in a mouse tauopathy model. Mol Neurodegener 13:59.
Zhang H, Cao Y, Ma L, Wei Y, Li H (2021) Possible mechanisms of tau spread and toxicity in Alzheimer's disease. Front Cell Dev Biol 9:707268.
Zhou Y, Shi J, Chu D, Hu W, Guan Z, Gong CX, Iqbal K, Liu F (2018) Relevance of phosphorylation and truncation of tau to the etiopathogenesis of Alzheimer's disease. Front Aging Neurosci 10:27.

C-Editors: Liu WJ, Zhao M, Qiu Y; T-Editor: Jia Y

Journal of the Brazilian Society of Mechanical Sciences and Engineering (2023) 45:583
https://doi.org/10.1007/s40430-023-04504-3

TECHNICAL PAPER

Novel electrode for thermal additive centrifugal force-assisted abrasive flow machining

Anant Bhardwaj[1] · **Krovvidi Srinivas**[1] · **Rajiv Chaudhary**[1]

Received: 14 September 2022 / Accepted: 25 September 2023
© The Author(s), under exclusive licence to The Brazilian Society of Mechanical Sciences and Engineering 2023

Abstract
Thermal additive centrifugal abrasive flow machining (TACAFM) process is a recent hybrid of the abrasive flow machining process, which combines the advantage of a centrifugal force-assisted abrasive flow machining process along with the material removal capability of the spark electro-discharge machining. The present work involves the parametric investigations of the TACAFM process using a novel spline shape electrode with a curved blade on percentage improvement in the surface finish ($\%\Delta R_a$) and material removal. The investigations include the modelling and simulations of flow parameters on Ansys Fluent and analysing experimental results using the artificial neuron network approach. The optimal values of the parameters are also computed using the artificial neuron network-coupled genetic algorithm (ANN-GA) approach. The simulated results satisfy the experimental and predicted effects of the ANN-GA. The SEM images justify the validity of the TACAFM process of the proposed electrode.

Keywords TACAFM · Abrasive flow machining · Artificial neuron network · Surface roughness · Material removal · Genetic algorithm

1 Introduction

The success of abrasive flow machining (AFM) in finishing is often quoted, but its limitation of low material removal (MR) is a significant issue of concern. Considering this fact, many researchers have come forward with new hybridisation to increase the MR and the application of the conventional AFM process. AFM is an unconventional nano-finishing process that uses a non-Newtonian media (containing polymer, gel and abrasives) extruded to and flowing through the internal surface of the workpiece. The media is extruded by the pressure difference maintained between the upper and lower ends of the media cylinder. The abrasives present in the media pass through the restrictive passage and abrade the surface of the workpiece pertaining to the desired MR at the requisite finish, which found its applicability in automobiles, biomedical, space, and specific industrial applications where precise finishing and polishing are required, like the finishing of microchannels after the EDM process, fuel injectors and ink-jet printer nozzle and pumps [1]. To improve the efficiency of the AFM, It is divided into one-way, two-way and orbital AFM based on their configuration [2]. To improve the efficiency and overcome the limitation of low material removal of abrasive flow machining, various hybrids of AFM are continuously developed based on the energy source in constant development. These are centrifugal force-assisted AFM [3, 4], magnetic force-assisted AFM process [5], helical AFM [6], electrochemical-aided AFM [7], hybrid of electrochemical and centrifugal force-assisted AFM Process [8], ultrasonic force-assisted AFM [9], drill-bit-guided AFM process [10], rotational AFM [11], hybrid of magnetic and centrifugal force-assisted AFM process [12]. The new hybrid in this regard is the TACAFM process, which clubs the advantage of the centrifugal force-assisted AFM process and the electro-discharge machining process [13–16]. The advantage of the AFM process is that the finishing and the material removal are done up to the nano-level. Besides the (TACAFM), the conventional form of AFM has many applications.

Conventional AFM has the advantage of finishing over the other process. It uses a non-Newtonian viscous media

Technical Editor: Adriano Fagali de Souza.

✉ Anant Bhardwaj
anant5892@gmail.com

[1] Department of Mechanical Engineering, Delhi Technological University, Delhi, India

Published online: 14 October 2023

whose viscosity plays a crucial role in the process. Ali et al. [17] used carbon nano-tube particles in the abrasive media and drastically improved the MR and the $\%\Delta R_a$. The key parameters were the researchers using extrusion pressure, media flow rate, and abrasive. The researchers also used Ansys fluent and analysed the flow parameters of the conventional AFM and concluded that there is about a 10-per cent improvement in surface finish compared to the traditional AFM Process. Gautam and Jinesh [18] use the Taguchi method and grey relation analysis to optimise the process parameters for the machining of SS304 material. The parameters taken by the researchers were abrasive grain size, extrusion pressure and the number of cycles. The researchers concluded that the optimum parameters are 300 µm, 55 bar extrusion pressure, and 80 cycles.

The researchers focused on the machining and finishing processes of AFM through the media's development, which is a mixture of polymer, abrasive and gel. The development of media is achieved by varying and improving the property of the polymer, gel or using different type of abrasives, which improves the machining characteristic of the AFM process.

Kar et al. [19] developed the newly formed media based on the butyl rubber-based media and studied its effect on the machining parameters. The researchers studied the morphological and rheological characteristics of the process. They concluded that the developed media, a mixture of butyl rubber, silicon carbide and naphthenic oil, performs better than the natural rubber. Uhlmann et al. [20] highlighted the application of AFM in the automobile industry and focused on machining and improving the holes. The researchers used the conventional AFM process for machining intersecting holes, fuel rails and feed lines at a high aspect ratio and found satisfactory results and concluded the capability of the traditional process of AFM in machining the intrinsic part of the automotive industry with affirmation. Venketesh et al. [21] applied the conventional AFM process in machining the bevel gear for industrial application. The researchers used media flow rate, abrasive mesh size, processing time and extrusion pressure as the main parameters of the investigations. The researchers concluded that after the finishing from AFM, the EN-8 bevel gears showed more than fifty per cent improvement in surface finish and extrusion pressure contributed 57 per cent to the process. Hashmi et al. [22] used AFM to machine 3D-printed FDM (fused deposition material) based material. The researcher used parameters like abrasive mesh size, abrasive concentration, finishing time and layer thickness. The researchers synthesised the media using waste vegetable edible oil as a liquid synthesiser, Sic particles, natural polymer and glycerine. They concluded that there is an improvement of 57 per cent in surface finish compared to the traditional AFM process.

To improve the efficiency of the AFM and its hybrids process, the researchers used mathematical modelling and simulations for the AFM process. Bhardwaj et al. [23–26] used the Ansys fluent in analysing the effect of flow parameters on the material removal and the surface roughness. The researchers used the pressure and temperature distribution to study the machining effect of the process. Zhang et al. [27] prepared a water-based media containing silicon for polishing using the AFM process and highlighted its successful application. Shadab et al. [28, 29] developed silicon-based media with alumina to finish titanium-based alloy using magnetic abrasive finishing and magnetic force-assisted AFM. Dixit et al. [30] developed the xanthan-based media for the conventional AFM process and jotted the media contribution to improve the process. Yadav et al. [31] studied the mechanism of material removal by the viscoelastic media and highlighted the application of AFM in the modern finishing process. Fu et al. [32] suggested an improved material removal process through AFM by suggesting efficient fixturing methods.

The AFM has covered a significant manufacturing area, from micro-nano-machining to the additive and selective laser melting process. Goyal et al. [33] and Singh et al. [34] accounted for the recent advancement in AFM, highlighting the major application and advantages of the process over other non-traditional processes. Mofidi et al. [35] developed the mathematical model to study the magnetic force-assisted AFM process and highlighted the advantage of clubbing the magnetic assistance with the conventional AFM process. Manjunath et al. [36] used AFM for the surface finishing of the aluminium propeller. Rajender et al. [37] developed a novel fixture to improve the conventional AFM process efficiency. Han et al. [38] and [39] used the AFM process for the machining of microchannels produced after the selective laser melting (SLM) process. Kuma et al. [40] highlighted AFM's application for finishing additively manufactured metal components. Basha et al. [41] developed a novel media for machining pure copper made by atomic diffusion technique through an additive manufacturing process and highlighted the successful implementation of the process. Bhardwaj et al. [42, 43] highlighted the effect of electrode geometry for enhancing the productivity of TACAFM process.

The proposed literature review highlights the broad reach of the AFM process, but the major challenge of low MR and considerable machining time limits its capabilities. For this purpose, the novel TACAFM process was proposed. The electrode used in the process was spline shape with straight blade. In the present work, the researcher used spline shape electrode with curved blade in order to increase the centrifugal action of the media on the wall of the workpiece. Due to this, the dynamic number of abrasive particle increases near the wall of the workpiece, enhancing the material removal.

The basis of the curved shape of the electrode is taken from the curved blade of the centrifugal fan whose efficiency increases as compared to centrifugal fan with the straight fan blade. Furthermore, ANN modelling is done to map the input parameters to output parameters, and multi-objective optimization was done to find optimal input parameters.

1.1 Thermal additive centrifugal force-assisted abrasive flow machining

The TACAFM process is developed by Ali et al. [13–16]. The method combines the advantage of the centrifugal force-assisted AFM process and the EDM process. The process consists of a centrifugal force generating a negatively charged electrode, which pulsed DC supply electrode charges. The material removal is done on the anodic workpiece, which should be conductive. As the current is supplied to the electrode, it emits electrons, which tend toward the positively charged workpiece. In between, these electrons collide with the media particle, which is non-conductive. As a result of these collisions and also due to the media's repulsive action, the spark generated between the workpiece and the electrode surface, which makes the process capability of more material removal and continuous extrusion of media with abrasive, provides a smoother and improved surface finish. Figure 1 shows the mechanism of spark formation, and Fig. 1(i) shows the top view of the electrode and the workpiece with the gap between the electrode and the workpiece. Figure 1(ii) shows the enlarged view of the gap between the electrode and the workpiece. When the pulsed DC is supplied to the workpiece and electrode, the electrons move towards the workpiece from the electrode surface, where they strike with media particles and ionise them. As the concentration of electrons increases, the ionisation increases, resulting in the formation of the plasma layer. This plasma layer forms the required spark and removes the material by melting and erosion.

1.2 Novel electrode for TACAFM process

The geometry of the centrifugal force-generating electrode plays a vital role in the centrifugal force-assisted AFM process (CFAAFM) and the TACAFM. Thus, selecting an optimum shape for the centrifugal force-generating electrode is essential, which can maximise MR, and gives the optimum value of the $\%\Delta R_a$. Considering this, a novel centrifugal force-generating (CFG) electrode is proposed in this work, a spline shape with a curved blade. The initial shape of the CFG electrode is a spline shape with a straight blade. The straight blade is because the gap between the edge and the workpiece is adequate for generating the spark. Also, the gap between the spline blades allows the melted material from the workpiece to mix with the media, which passes through the gap (Fig. 2).

Amjadimanesh et al. [44] simulated and conducted various experiments to determine the optimum shape of the fan blade and concluded that the circular shape blade is the most efficient blade type. The reason the curved blade emerged from the centrifugal fan. The shape of the centrifugal fan blade is curved forward and backward, producing more centrifugal force than the fan with a straight blade [45]. The curved blade's rotation hits the media containing the abrasives, polymer, and gel. These abrasives, in turn, strike the surface of the workpiece, which results in the abrasion process. The efficiency of this abrasion process improves by making the blade curved rather than straight. Thus, the

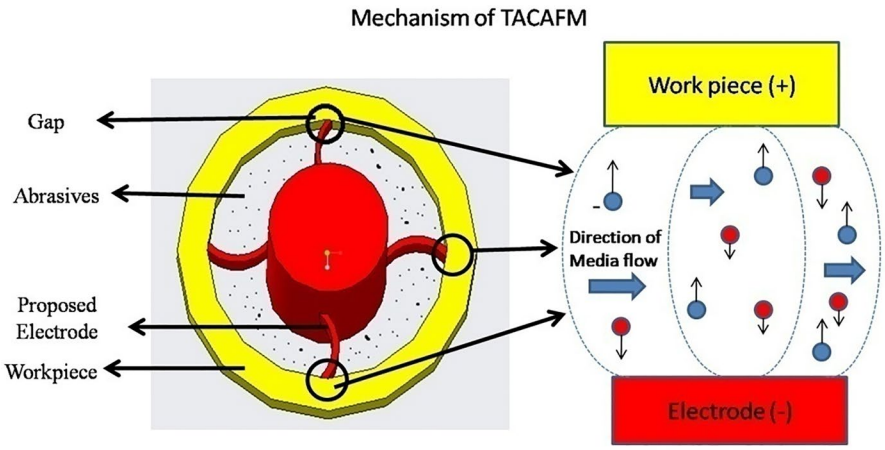


Fig. 1 Mechanism of spark formation

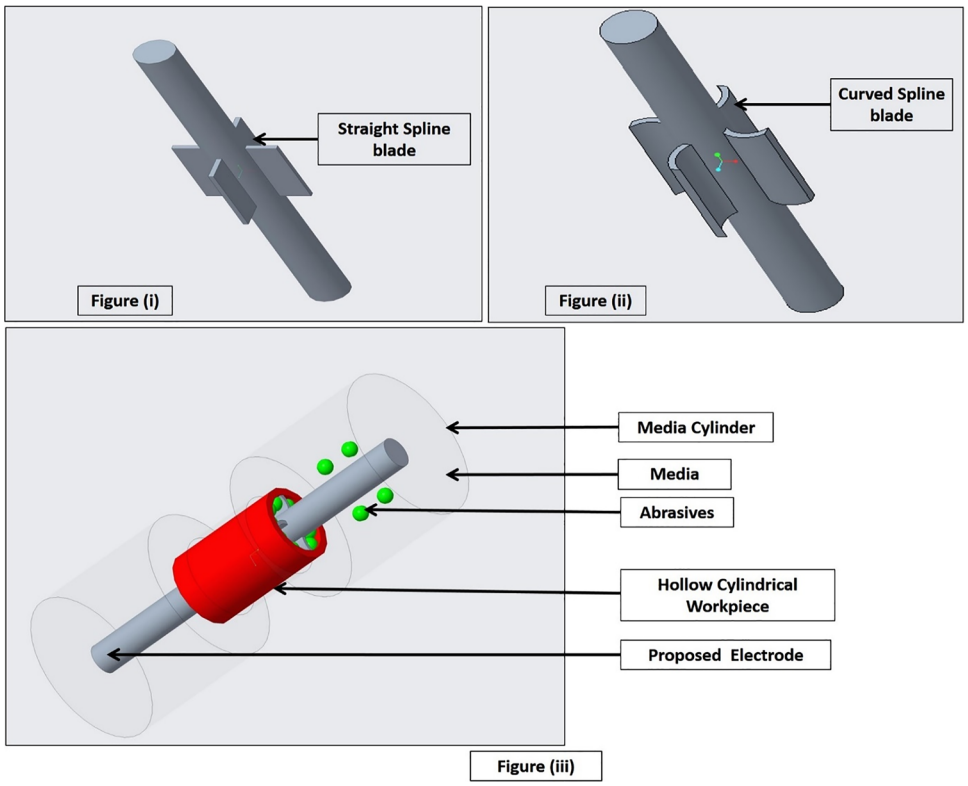


Fig. 2 Novel electrode, (i) spline shape electrode used in set-up, (ii) proposed spline shape electrode with curved blade, (iii) arrangement of electrode and the workpiece

validity of the proposed electrode with the curved blade is proved.

2 Materials and methods

To test the proposed CFG electrode, the researchers modelled it, simulated it on ANSYS 22.0 and validated the experimental results. The methodology is discussed in the following section. The modelling aims to find the feasibility of the proposed electrode with the TACAFM process and to analyse the flow parameters' effect on the machining process.

2.1 Modelling and simulation

The initial model is made on CREO 3.0 and then exported to ANSYS Fluent, where it is simulated under the required boundary conditions. The modelling of the initial set-up was done on ANSYS Fluent, where the media used is viscous, containing a mixture of abrasive polymer and gel, and the abrasive used is alumina. A similar property of media was used as given to the experimental media for validation. The Boolean operations were used to create a proper flow of media through the workpiece, and the proper name selection was made for the post-CFD computation of results. The material of the workpiece used in the simulation was brass, as brass is used as a prominent component in modern manufacturing, automobile, biomedical and aerospace industries. Also, the alloy shows well-defined mechanical and thermal properties [46]. The dimension of the workpiece is taken according to Kohut's recommendations [47], which is 10-mm outer diameter, 8-mm inner diameter and 10-mm height. The material of the CFG electrode is copper due to its good thermal and electric conductivity. The dimensions of the CFG electrode are designed based on the fact that the minimum area of flow is given to the media during the extrusion process so that the dynamic numbers of media particles increase near the workpiece wall, resulting in the enhancement of the machining process [48]. Also, a minimum gap is maintained between the workpiece and electrode to generate the spark. The next step was meshing, for which an adequately refined mesh was prepared. The size of the elements was 0.5 mm, and the mesh was well prepared at the corners with proper skewness.

The total number of elements was 83,150, with the maximum tetragon. The mesh file was exported to the experimental set-up, where the viscous pressured-based model with energy equation was considered with double-precision accuracy for the media flow of the simulation set-up. The boundary condition of the simulation is the rotation of the CFG electrode with 100 RPM. The inlet is a pressure inlet, and the outlet is a pressure outlet. The inlet pressure is kept at 40 MPa, and the pressure outlet is 20 MPa (Fig. 3).

Again to generate a spark and produce the EDM effect on the workpiece, a minimum gap between the spline and

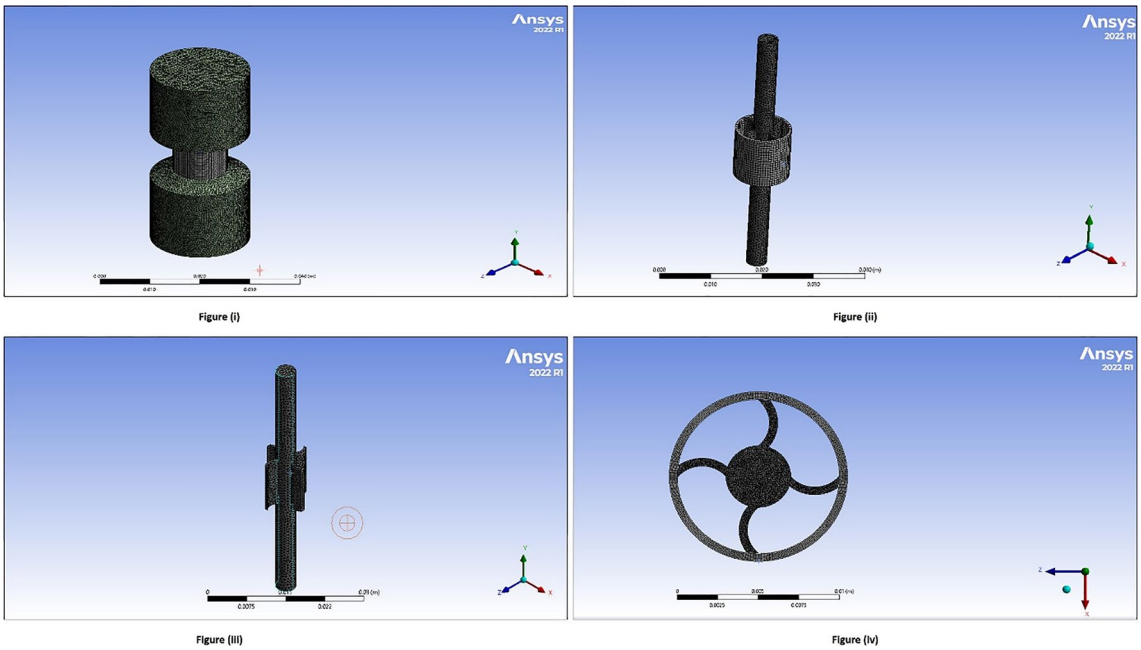


Fig. 3 Mesh: (i) mesh file of the set-up, (ii) mesh file of the proposed electrode and workpiece, (iii) mesh file of the proposed electrode front view, (iv) mesh file of the proposed electrode and workpiece top view

workpiece is less than 0.01 mm. The rotating electrode is given a temperature of 2300 °C. The minimum gap between the electrode and the workpiece with the electrolyte environment (provided by media) generates a spark, which results in the melting of the workpiece. The following figure shows the meshing of the proposed CFG electrode for the TACAFM process. The material of the proposed electrode is copper, as it provides EDM effects and is a good conductor of electricity. Figure 4 shows the mesh files of the proposed electrode. Figure 3(i) shows the mesh file of the proposed electrode containing the workpiece, electrode, and abrasives. Here, the wall is made up of a non-friction wall with the property of the viscous fluid, whose properties are already listed in Table 1. Table 2 shows the boundary conditions. Figure 3(ii) shows the mesh file of the electrode and the workpiece. Figure 3(iii) shows the mesh file of the proposed electrode in the front view, and Fig. 3(iv) shows the top view of the proposed electrodes. Figure 4 shows the boundary simulation set-up with the boundary conditions.

After the successful meshing, the model was prepared for the calculations. The advection scheme was kept at high resolution, and solutions were solved for 100 iterations. The length scale control was conservative, with time scale control set to auto time scale. The time scale factor is kept at 1, and the convergence is up to 10^{-06}. The result obtained was grid-tested and was independent of the mesh size.

The laminar modal is used in the analysis as the media used in actual TACAFM process is a mixture of polymer, gal and abrasive, which is highly viscous in nature, and as it passes through the constrained path, the media velocity is also very low, which makes the model into the laminar domain. Figure 4 shows the boundary condition and conversion curve. Figure 4(i) shows the boundary conditions of the fluid, which includes solid stationary wall. Figure 4(ii) shows the boundary condition applied to the modal. The detail boundary condition is listed in Table 2. Figure 4(iii) shows the conversion curve with the convergence rate on the X-axis and the number of iterations on the Y-axis. The conversion process is smooth with the energy line below the X, Y and Z velocity and the continuity equation line.

2.2 Experimental set-up

To validate the simulation results, experiments were performed to analyse the effect of supply current, the rotational speed of the CFG electrode and the number of cycles on the MR and $\%\Delta R_a$. For this purpose, the experimental set-up was fabricated to club the advantage of spark EDM with a centrifugal force-assisted AFM process. The set-up consists of a conventional AFM set-up, a three-phase induction motor with a power drive, an EDM controller and the EDM power supply.

EDM controller is a device used to maintain the proper supply of current and maintain an adequate potential gap between the workpiece and the electrode for spark generation by sensing the gap between the electrode and workpiece with the help of an inbuilt MOSFET. The set-up (shown in Fig. 5) consists of a 25-Ampere EDM controller and a

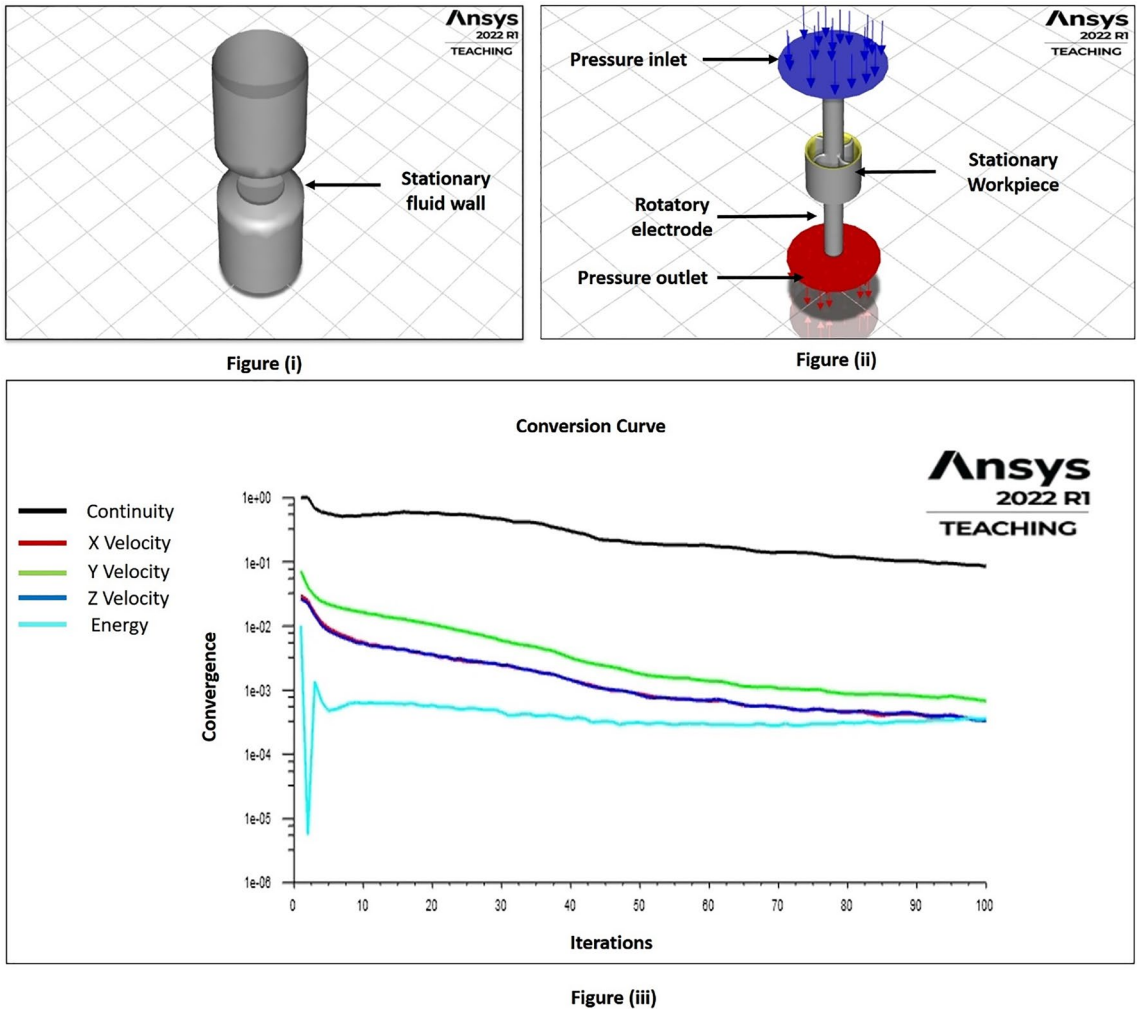


Fig. 4 Simulation set-up for ANSYS Fluent: (i) boundary condition of fluid wall, (ii) boundary conditions of electrode and workpiece, (iii) conversion curve

Table 1 Material properties for simulations

Material	Media	Brass	Alumina abrasive
Type	Fluid	Solid	Solid
Density (kg/m^3)	1219	8300	3950
Viscosity (kg/m-s)	0.789	109	12
Thermal conductivity (w/m–k)	0.22	401	451

Table 2 Boundary conditions

Sr. no.	Components	Boundary type	Parameters
1	Inlet	Pressure Inlet	40 MPa normal to surface
2	Fluid (media)	Wall	Stationary wall
3	CFG electrode	Rotating wall with temperature input	100 rpm rotation 2300 °C
4	Workpiece	Wall	Stationary wall
4	Outlet	Pressure outlet	20 MPa

pulse DC power supply with 3KVA power, with 415 V and 50 Hz frequency (Fig. 5(i)). To facilitate the rotation of the CFG electrode, the 3-phase induction motor (800 rpm and 1 HP) is used with the frequency control power drive mechanism (0.75 KW/1 HP, 240 V/Phase) shown in Fig. 5(ii). Figure 5(iii) shows the arrangement of the nylon fixture containing the workpiece and CFG electrode. Also, the set-up consists of the terminals connected to the electrode and the workpiece and the power supply for the DC supply. The fixture's design is based on the fixture used by Ali et al. [13–15]. Figure 5(iv) shows the TACAFM set-up with all the requisite accessories used in the investigation. Figure 5(v) shows the details of the conventional AFM set-up with the 3-phase arrangement.

Fig. 5 Experimental set-up: (i) EDM controller and power supply, (ii) 3-phase induction motor with power drive, (iii) fixture containing the workpiece and CFG electrode, (iv) TACAFM set-up, (v) conventional AFM set-up with labelled parts, (vi) side view showing power transmission mechanism for rotation of CFG electrode

The conventional AFM works on the maximum pressure of 20 MPa with a media flow rate of 290 cm^3/min. The media used contain polymer, gel (polymer to gel ratio 1:1) and abrasive (abrasive to media 50% by weight, three hundred mesh size). The polymer is prepared with the help of hydrocarbon oil, and the gel is prepared using Lewis acid and aluminium stearate. The abrasive used in the investigation is alumina with a 300 mesh size. Figure 5(vi) shows the experimental set-up's side view, highlighting the intermediate gear connecting the fixture gear with the induction motor gear for the power supply to support the proposed electrode's rotation mechanism.

2.3 Experimentation

The researchers performed experiments with the help of spline shape electrodes whose blades were curved. The idea behind the curved blades was to direct the media toward the wall of the hollow cylindrical workpiece where the machining of the surface takes place effectively. To experiment,

the researchers used the TACAFM experimental set-up. The initial parameters were electrode rotation speed, the current supply to cater for the EDM effect, and the number of oscillations. Table 3 shows the parameters and their different levels. The values of speed are kept low for the experimental purpose. This is because, at a higher rotational speed, the centrifugal force generated on the workpiece would be high, which may lead to the breakdown of the workpiece. Also, higher rotational speed leads to unwanted vibrations that may affect the experimentation environment [14].

Table 3 shows three different parameters of the TACAFM process, which include the speed of the rotation of the electrode in which the current is supplied, the current supplied, and the number of oscillations. The number of oscillations means a movement of media from the upper position to the lower and back to the upper due to the pressure difference applied. In this investigation, the pressure difference lay was 20 MPa. The initial experiment was performed on the TACAFM set-up, considering listed parameters at various levels to compute the percentage improvement in the surface finish, defined as the ratio of the difference between the initial and final surface finish and the initial surface finish. The material removal is calculated by the difference between the initial weight before and after machining.

Taylor Hobson Talysurf computed the surface roughness of the workpiece. Taylor Hobson Talysurf works on the DC power and can give R_a and R_z values with a maximum specimen length of 60 mm. The researchers also computed material removal, which is the difference between the weight of the workpiece before the machining and the weight of the workpiece after the machining process. The precision balance took the weight of the workpiece. The precision balance used is of Sartorius company and works on the load cell principle. The precision balance used for the investigation has a capacity of 200 gm, and the resolution of the precision balance is 0.10 mg.

The workpiece used for the computation is a hollow cylinder whose internal diameter is 8 mm, outside diameter is 10 mm, and length is 10 mm [47]. The material of the workpiece used is brass. The media used for the experiment consisted of a mixture of abrasive, polymer and gel. The property of the media kept similar to that used in the simulation process. Figure 6 shows the experimental requirement for set-up. Figure 6(i) shows the brass workpiece. Figure 6(ii) shows the nylon fixture used in the investigation. The fixture is made of nylon, which is a bad conductor of electricity; as a result, it does not affect the set-up when the potential difference is created between the anodic workpiece and the cathodic electrode for the spark generation. Also, a suitable arrangement is made in the fixture to avoid a short circuit when the current is provided.

Figure 6(iii) shows the set-up of the Taylor Hobson Talysurf before the machining process. The workpiece is supported on the V block, and initial roughness is taken to calculate the $\%\Delta R_a$ before and after finishing. Figure 6(iv) shows the set-up of the Taylor Hobson Talysurf after finishing. Here, the workpiece is broken longitudinally into two halves and the surface roughness is measured accordingly. The V block holds the workpiece for the Taylor Hobson Talysurf. Figure 6(v) shows the top view of the proposed electrode. The profile is adjusted according to the internal diameter of the workpiece, and a tolerance of 0.05 mm is maintained so that the spark can be obtained between the curved flange and the internal surface of the workpiece. Figure 6(vi) shows the front view of the proposed electrode. Figure 5(vii) shows the original profile of the proposed electrode obtained from the wire-cut EDM process.

Table 3 shows the values of the percentage improvement in surface roughness and the material removal against the specified parameters obtained after the experiment on the TACAFM set-up. The experimental data obtained from the set-up are analysed by artificial neural network (ANN). In this work, the backpropagation method for the ANN is used, and the given data are divided into training, testing and validation model. MATLAB software was used for applying the ANN. The architecture contains three layers: the input layer, the hidden layer and the output layer. The first layer has the 3 neurons corresponding to the input parameters (rotational speed, current supply and number of oscillations), and the second layer is the hidden layer. The software has inbuilt algorithms that adjust the hidden layer's neurons corresponding to the training data. The third layer contains two neurons corresponding to the output parameters (MR and $\%\Delta R_a$). Figure 7 shows the architecture for the ANN network used in the present investigation. Initially, the data are imported into the software as input and output matrices. The second step involves inputting the training's functions, the Levenberg–Marquardt (trainlm). The network training performance is based on the mean-squared error, and the calculation is based on MEX.

After the successful training of the data, the following graphs were obtained and verified by the regression. Figure 7(i) shows the line diagram containing the input, hidden and output layers. For this investigation, three input layers are the rotational speed of the electrode, the supply current and the number of oscillations of the media between the internal surfaces of the workpiece. The output

Table 3 Process parameters and levels

Parameters	Units	Level 1	Level 2	Level 3	Level 4
Speed of electrode	rpm	10	20	40	60
Supply current	Ampere	4	8	12	16
Number of oscillations	cycles	2	4	6	8

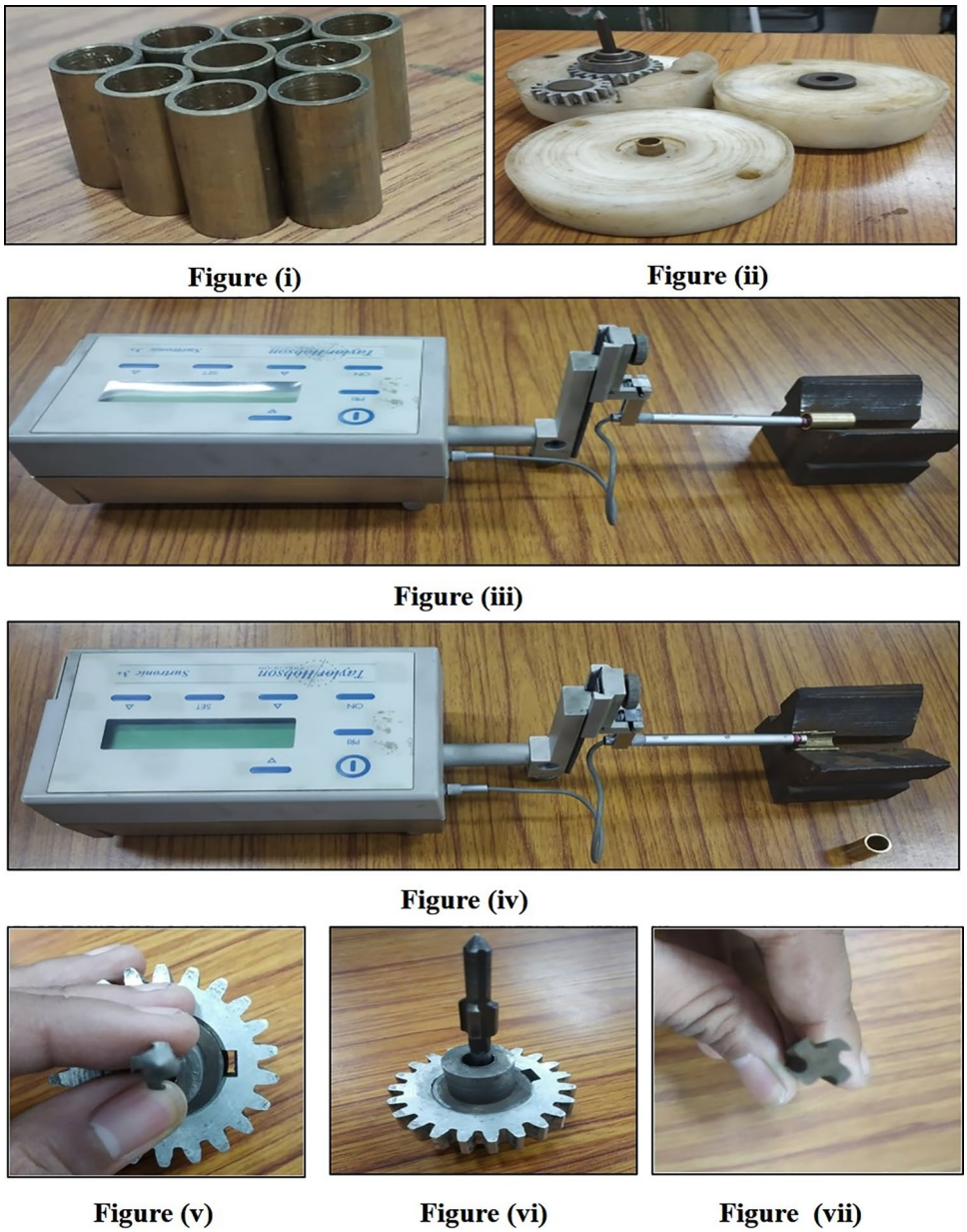


Fig. 6 Experimental requirement: (i) hollow brass workpiece, (ii) nylon fixture, (iii) surface roughness measurement before machining, (iv) surface roughness measurement after machining, (v) top view of proposed electrode, (vi) front view of proposed electrode, (vii) cross section of proposed electrode

layers are $\%\Delta R_a$ and MR. Figure 6(ii) shows the training window of the ANN application on MATLAB software. It gives real-time information about the neural network structure containing input, output and hidden layers. The training windows provide information about the algorithms (data division, training, performance and calculations) used in the investigations and also about the progress and checking, which includes information on epoch, performance, gradient and validation check and about mu, which are essential parameters for the development of ANN networks. Also, it gives information regarding the real-time plots containing the information on performance training status and the regression analysis with the curve fit.

3 Result and discussion

Based on the modelling of ANSYS 22.0, the experiment and the analysis of the ANN, the results were divided into subsequent sections in which detailed arguments are discussed. The simulation part consists of the distribution of the pressure and temperature over the brass workpiece,

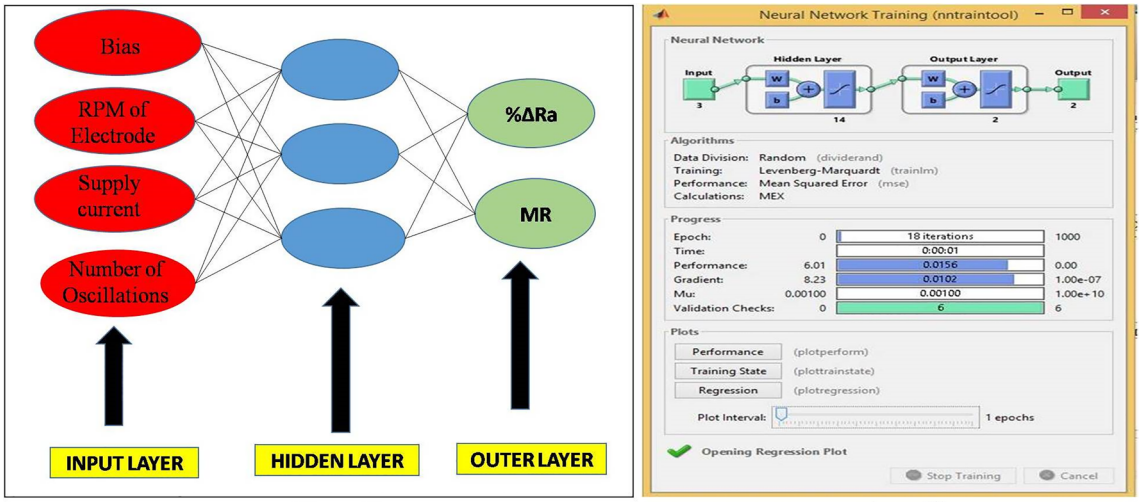


Fig. 7 ANN modelling: (i) ANN architecture of the model and (ii) training window of ANN application on MATLAB software

and the experimental results were analysed to validate the simulation results.

3.1 Simulation results

After successfully implementing the modelling on ANSYS Fluent, the plots were obtained and the feasibility of the proposed electrode with the TACAFM process was analysed. Figure 8(i) shows the pressure distribution over the workpiece when the straight spline electrode is considered. The maximum pressure value is 25.65 MPa, while Fig. 8(ii) shows the pressure variation of the workpiece under the same conditions by the proposed electrode. The maximum pressure is present at the upper part of the workpiece, which is 37 MPa, much higher than the straight spline electrode. It is due to the combination of the spark effect and the rotation of the CFG electrode.

Being the curved blade instead of the straight spline blade, more centrifugal force is generated, which increases the pressure on the workpiece and thus results in more material removal. Therefore, the machining, in this case, would result from the axial extrusion, CFG electrode, and melting due to the spark produced through the EDM effect. Also, the curved spline has more value of maximum pressure distribution compared to the straight spline rod used in the initial investigation of the TACAFM process by Ali et al. [13–16]. It is also evident that the proposed electrode has a maximum pressure value between the extrusion pressure of 20–40 MPa. The high-pressure value signifies the high value of cutting forces, resulting in MR enhancement [24–26]. Figure 8(iii) shows the temperature distribution of the workpiece. It is evident from the workpiece that the maximum temperature is present where the curved flange is present near the workpiece. The presence of maximum temperature on the flange justifies the spark present due to the EDM effect, which can be validated through the initial investigation of the TACAFM process [12]. The intensity of the spark in the actual process varies with the intensity of the current, but to show the EDM effect, the EDM temperature of the tool from the literature is given to the electrode for the simulation purpose, which is about 2300 °C [13]. Figure 8(iv) shows the streamline representing the passage of media through the hollow workpiece and the proposed electrode placed coaxially with the workpiece. The streamline is equally spaced and flows through the constricted passage where the workpiece is placed. The important point for observation is that the streamline is more compactly spaced inside the workpiece, which represents the increase in the dynamic number of abrasive particles inside the workpiece, resulting in the enhancement of MR due to the increase in the abrasive and the work-surface interaction which is desirable [48]. Figure 8(v) shows the pressure distribution over the media. The pressure distribution on the media is mainly due to the extrusion pressure, which is higher at the inlet side and lowers at the outlet side. The critical point of consideration is that the pressure value at the workpiece's bottom side is lower than the outlet pressure. The pressure value at the lower side of the workpiece in Fig. 8(ii) is around 18 MPa, which is lower than the outlet pressure. The reason behind this is that when the media flows from the inlet to the workpiece, due to viscosity and internal friction with the wall and flow, the pressure decreases with the flow, but when media enters inside the workpiece, the flow area decreases. As a result, the pressure

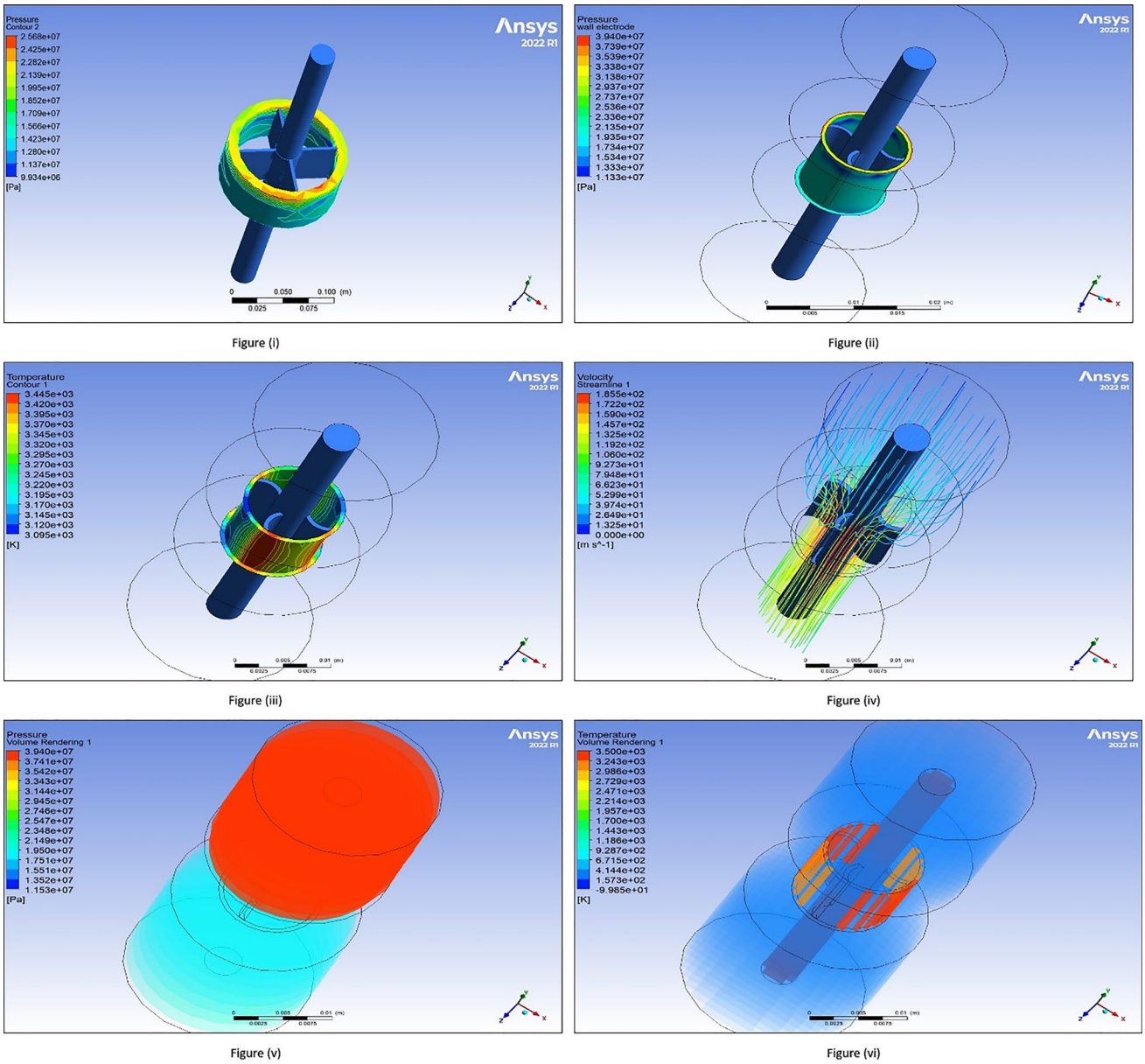


Fig. 8 Simulation results: (i) pressure distribution over workpiece by straight spline CFG electrode, (ii) pressure distribution over workpiece by proposed CFG electrode, (iii) temperature distribution over workpiece by proposed CFG electrode, (iv) velocity streamline during TACAFM process, (v) pressure distribution over the media during TACAFM process, (vi) temperature distribution over the media

increases inside the workpiece and machining occurs. When the media flows out from the workpiece, the flow area drastically increases, which results in a decrease in pressure at the local region. This phenomenon is known as the bell-mounting effect [24, 25, 49]. Due to this decrease in pressure, the dynamic number of abrasive particles increases near the bottom side of the workpiece and the abrasive and work–surface interaction increases resulting in the effective machining process. Figure 8(vi) shows the temperature distribution over the media during the TACAFM process. The figure shows that the workpiece is at a higher temperature, highlighting the machining process through the melting and erosion process.

3.2 Experimental results

After the successful training of the data, the following graphs were obtained and verified by the regression.

Figure 9(i) shows the performance curve after the experiment. The network was trained for 18 epochs. An epoch in

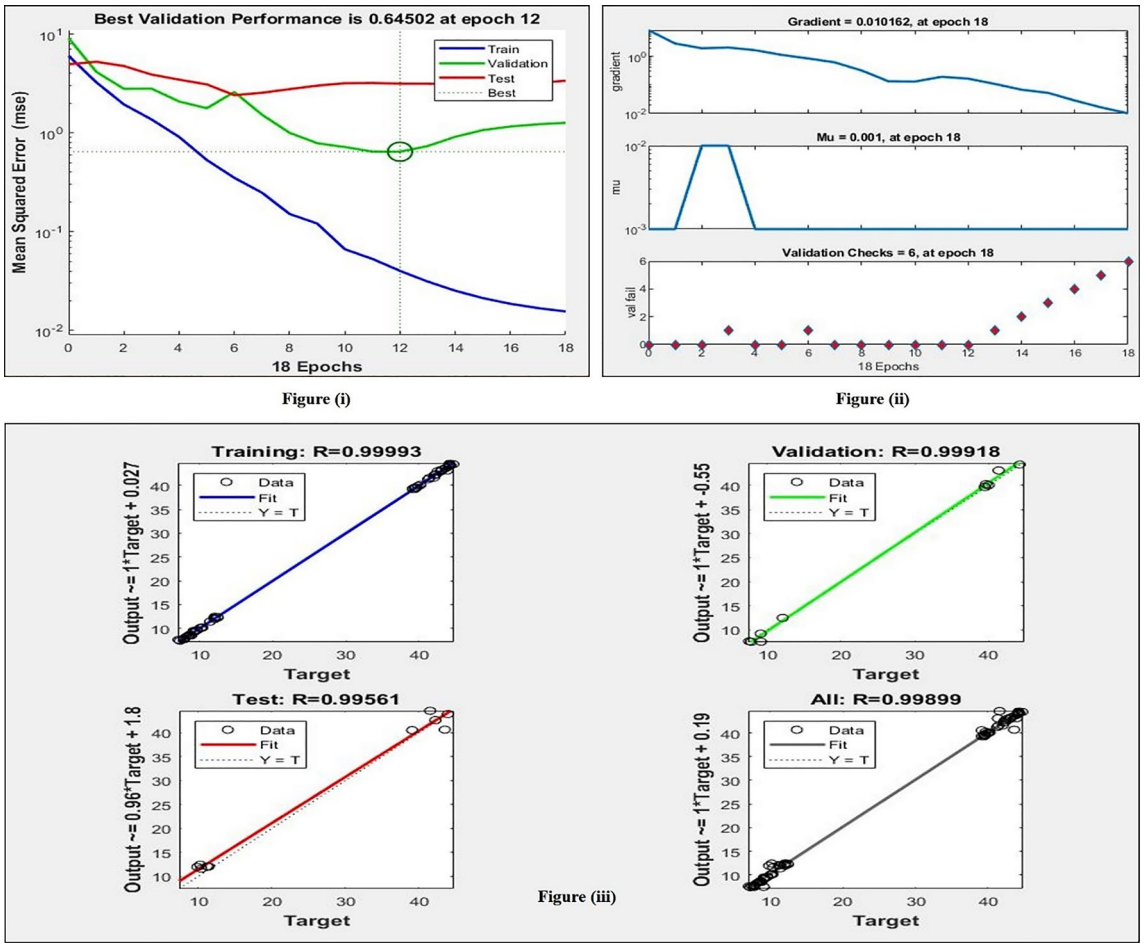


Fig. 9 Graphical representation of the analysis; (i) performance of the neural network, (ii) training status, (iii) regression graph of the data

the neural network is the neural network training with all the training data for 1 cycle. The data is initially divided into training, testing and validation. About 70 per cent of the data is under training, thirty per cent is under testing, and thirty per cent is under validation. The blue line shows the training curve. The red line shows the testing curve, and the green line shows the validation curve. It is clear from the curve that the training line is below the test and the validation line, which validates the result and gives a smooth conversion. The curve shows the dynamics of mean square error for all data sheets in a decreasing manner in the logarithmic scale. It also shows the validation and testing curve on the same scale. The important point of observation is that the best validation performance is observed at 12 epochs of 0.64502. It is because, from this point, the validation curve takes increasing trends. The other critical point that is taken into consideration is that the training curve is in descending order, which is the foremost requisite, which means that the mean square error is low for the training process.

Figure 8(ii) shows the neural network training status where the first graph shows the gradient vs epochs, the second shows the Mu vs epochs, and the third is validation fail vs epochs. The first graph shows the gradient value as 0.010162 at the end of the cycle 18 epoch. It indicates that the local minimum of the goal function required for training has been achieved after 18 epochs; it is considered a better value as it approaches zero. The gradient is a value of backpropagation on each iteration in a logarithmic scale. The graph of mu vs epochs shows the value of Mu as 0.001 at 18 epochs. Mu is the control parameter for the algorithm used to train the neural network and is used to control the weights of neurons involved in backpropagation during training. The validation fails indicate 6 validation fails within 18 epochs. Validation fails are iterations when validation MSE increases its value. Many failures mean overtraining, and MATLAB stops training after continuous failures in a row [48].

Figure 8(iii) shows the regression curve for the training, testing and validation of the data. The regression curve shown in the blue line has a regression coefficient value of 0.9993, which is very close to 1, and the square is very close to 1, which shows the efficient training process. The regression curve is indicated by a green curve

Table 4 Experimental values of MR and $\%\Delta R_a$

Sr. no.	RO	Rotational speed	Supply current	No. of oscillations	MR (mg)	$\%R_a$
1	1	10	4	2	7.82	39.98
2	2	10	4	4	9.14	39.46
3	3	10	4	6	8.56	39.42
4	4	10	4	8	7.92	39.61
5	5	10	8	2	8.84	41.42
6	6	10	8	4	9.16	42.03
7	7	20	8	6	10.34	43.05
8	8	20	8	8	7.87	42.96
9	9	20	12	2	9.56	43.61
10	10	20	12	4	10.14	43.74
11	11	20	12	6	12.43	42.11
12	12	20	12	8	11.43	42.29
13	13	40	16	2	9.88	41.56
14	14	40	16	4	10.42	43.98
15	15	40	16	6	9.26	44.28
16	16	40	16	8	10.12	44.16
17	17	40	4	2	12.46	39.86
18	18	40	4	4	11.23	39.11
19	19	60	4	6	9.65	39.42
20	20	60	4	8	8.12	39.14
21	21	60	8	2	9.11	39.62
22	22	60	8	4	8.60	40.19
23	23	60	8	6	9.12	41.23
24	24	60	8	8	8.42	42.32
25	25	10	12	2	10.32	44.16
26	26	10	12	4	12.42	44.68
27	27	10	12	6	10.26	42.28
28	28	10	12	8	11.42	43.56
29	29	10	16	2	12.08	44.17
30	30	10	16	4	12.08	44.11
31	31	20	16	6	11.99	44.15
32	32	20	16	8	12.05	44.19
33	33	20	4	2	7.56	41.32
34	34	20	4	4	7.11	43.89
35	35	20	4	6	7.86	42.56
36	36	20	4	8	7.31	43.14

whose regression coefficient is 0.99918 and whose square is also close to 1, which justifies the validation part. The regression curve is shown in red, and the regression curve is 0.99561, which is close to 1. The curve, shown in black, is the summation of the training, testing and validation, and the coefficient of co-relation is 0.99899, which is very close to 1. Thus, after the network training, the predicted output value for the MR and $\%\Delta R_a$ is obtained [50].

Once the predicted data are exported from the network, the percentage error is computed for both MR and $\%\Delta R_a$. Table 4 gives the value of the predicted output and the percentage error obtained after considering the difference in the data. It can be observed from the table that the maximum value of the percentage error, which is defined as the ratio of the difference between the predicted value and the experimental value multiplied by 100 to the obtained, is 3.24, which is at serial no. 34.

In contrast, the minimum value is zero, which is there in most cases. Considering the absolute percentage error for the percentage improvement in the surface finish, which is calculated as the ratio of the difference between the predicted value of the percentage improvement in surface finish and

the experimental value of the percentage improvement in surface finish to the experimental value of the percentage improvement in surface finish multiplied by 100, the maximum value went up to the 3.45 which is at 34 serial number. The minimum absolute percentage error in the percentage improvement in the surface finish came out to be zero for the most values, which satisfactorily proves the validation of the process and justifies the surface finish of the newly proposed electrode with the curved splines [51].

The average experimented value for MR is 9.77 mg, and the percentage improvement in the surface finish is 42.13 per cent. The average predicted value for MR is 9.73 mg and that for $\%\Delta R_a$ is 42.14 per cent. The cumulative average absolute per cent error for MR is 0.490 and that for $\%\Delta R_a$ is 1.05, which is acceptable for the TACAFM process and may be improved with further automation and experimental conditions [52]. From the data obtained from the experimental values, a comparison plot is obtained from the experimental data and the predicted data from the ANN functions. Figure 10 shows the comparison plot between the predicted and the practical value for the percentage improvement in surface finish. The accuracy of the network may be justified by the RMSE value for MR and $\%\Delta R_a$. The RMSE is the square root of the mean square error, which is calculated according to Eq. (1):

$$\text{RMSE value} = \sqrt{\frac{(\text{Predicted Value} - \text{Experimental Value})^2}{\text{experimental value}}} \tag{1}$$

It measures the average difference between values predicted by a model and the actual values. It estimates how well the model can predict the target value (accuracy) [53]. The average cumulative value of RMSE for MR is 0.034835, and that for the $\%\Delta R_a$ is 0.06501, indicating a better-trained model.

The experimental and predicted value shows similar trends at maximum locations, but variation is present. The reason for this is the less number of dataset. As the dataset increases, the epochs would increase and the regression value would be close to 1. The error would decrease when the regression value for the combined training, validation and testing approaches to 1 [54]. Figure 11 compares the experimental data of the material removal with the predicted data obtained from ANN. It is evident from the figure that the expected data mostly match the experimental data.

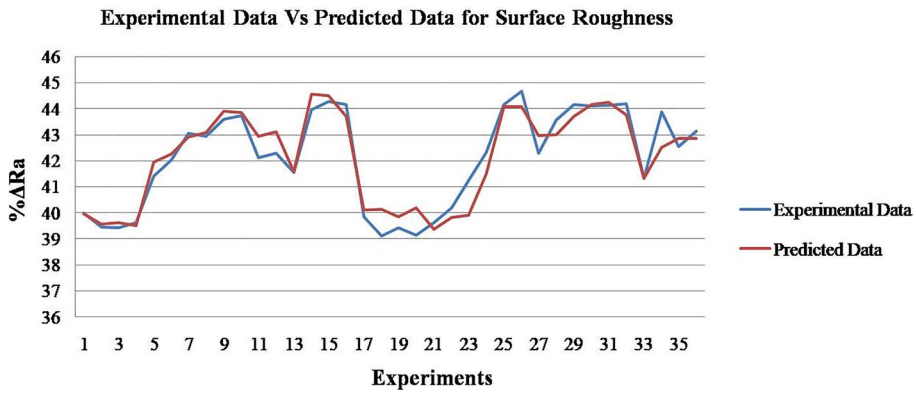


Fig. 10 Comparison of data and predicted value for percentage improvement in surface finish

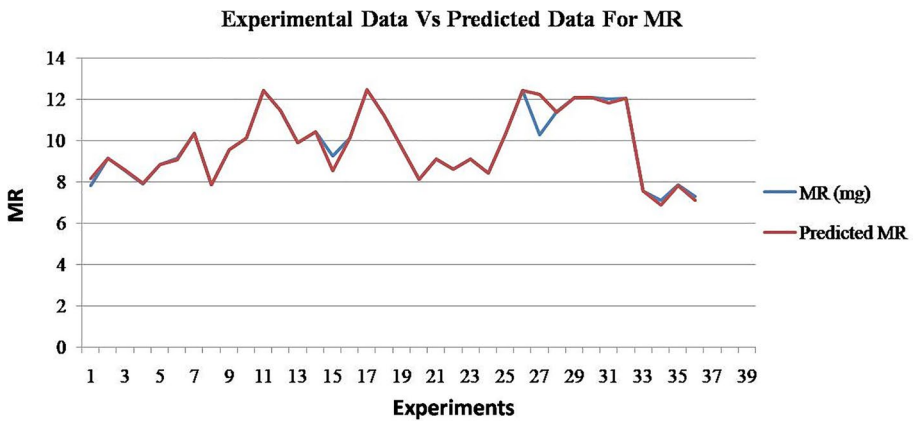


Fig. 11 Comparisons of data and predicted value for the material removal

Table 5 Predicted values percentage error and RMSE

Sr. no.	Running order	Rotational speed	Supply current	Number of oscillations	MR (mg)	$\%\Delta R_a$	Predicted MR	Predicted $\%\Delta R_a$	% error MR ABS	% error $\%\Delta R_a$ ABS	RMSE (for MR)	RMSE (for $\%\Delta R_a$)
01	01	10	04	2	7.82	39.98	7.86	39.96	0.51	0.05	0.121584	0.003163
02	02	10	04	4	9.14	39.46	9.14	39.56	0.00	0.25	0	0.015919
03	03	10	04	6	8.56	39.42	8.56	39.62	0.00	0.50	0	0.031855
04	04	10	04	8	7.92	39.61	7.92	39.51	0.00	0.25	0	0.015889
05	05	10	08	2	8.84	41.42	8.84	41.93	0.00	1.28	0	0.079244
06	06	10	08	4	9.16	42.03	9.08	42.24	0.87	0.53	0.026433	0.032392
07	07	20	08	6	10.34	43.05	10.34	42.91	0.00	0.35	0	0.021337
08	08	20	08	8	7.87	42.96	7.87	43.08	0.00	0.30	0	0.018308
09	09	20	12	2	9.56	43.61	9.56	43.89	0.00	0.70	0	0.0424
10	10	20	12	4	10.14	43.74	10.14	43.83	0.00	0.23	0	0.013608
11	11	20	12	6	12.43	42.11	12.43	42.92	0.00	2.03	0	0.124822
12	12	20	12	8	11.43	42.29	11.45	43.11	0.17	2.05	0.005916	0.126094
13	13	40	16	2	9.88	41.56	9.88	41.57	0.00	0.03	0	0.001551
14	14	40	16	4	10.42	43.98	10.42	44.54	0.00	1.40	0	0.084442
15	15	40	16	6	9.26	44.28	8.53	44.48	7.88	0.50	0.239893	0.030056
16	16	40	16	8	10.12	44.16	10.12	43.7	0.00	1.15	0	0.069222
17	17	40	04	2	12.46	39.86	12.46	40.11	0.00	0.63	0	0.039598
18	18	40	04	4	11.23	39.11	11.23	40.13	0.00	2.55	0	0.163101
19	19	60	04	6	9.65	39.42	9.65	39.84	0.00	1.05	0	0.066895
20	20	60	04	8	8.12	39.14	8.12	40.2	0.00	2.65	0	0.169432
21	21	60	08	2	9.11	39.62	9.11	39.37	0.00	0.63	0	0.039718
22	22	60	08	4	8.6	40.19	8.6	39.81	0.00	0.95	0	0.059941
23	23	60	08	6	9.12	41.23	9.12	39.9	0.00	3.33	0	0.207131
24	24	60	08	8	8.42	42.32	8.42	41.49	0.00	2.08	0	0.127587
25	25	10	12	2	10.32	44.16	10.32	44.05	0.00	0.28	0	0.016553
26	26	10	12	4	12.42	44.68	12.42	44.05	0.00	1.58	0	0.094251
27	27	10	12	6	10.26	42.28	10.28	42.95	0.00	1.68	0.618147	0.10304
28	28	10	12	8	11.42	43.56	11.37	42.97	0.44	1.48	0.014796	0.089394
29	29	10	16	2	12.08	44.17	12.08	43.7	0.00	1.18	0	0.070719
30	30	10	16	4	12.08	44.11	12.08	44.14	0.00	0.08	0	0.004517
31	31	20	16	6	11.99	44.15	11.82	44.24	1.42	0.23	0.049095	0.013545
32	32	20	16	8	12.05	44.19	12.05	43.76	0.00	1.08	0	0.064685
33	33	20	04	2	7.56	41.32	7.56	41.32	0.00	0.00	0	0
34	34	20	04	4	7.11	43.89	6.88	42.51	3.23	3.45	0.086257	0.208303

Table 5 (continued)

Sr. no.	Running order	Rotational speed	Supply current	Number of oscillations	MR (mg)	$\%\Delta R_a$	Predicted MR	Predicted $\%\Delta R_a$	% error MR ABS	% error $\%\Delta R_a$ ABS	RMSE (for MR)	RMSE (for $\%\Delta R_a$)
35	35	20	04	6	7.86	42.56	7.82	42.85	0.51	0.73	0.014268	0.044453
36	36	20	04	8	7.31	43.14	7.1	42.83	2.87	0.78	0.077671	0.047198
Total					352.06	1516.76	350	1517.07	17.90	38.02	1.254058	2.340363
Average					9.77	42.13	9.73	42.14	0.490	1.05	0.034835	0.06501

3.3 Optimal function

The next major challenge is getting the optimal or global solution for the input parameters. In our case, we need to get a solution such that the value of the percentage improvement in surface roughness gets minimised and the value of the material removal gets maximised. To achieve this, the researchers decided to consider a genetic algorithm account. The reason for considering the genetic algorithm is because of its evolutionary nature, and also, it is an adaptive heuristic search algorithm [55]. The genetic algorithm is based on genetic and natural selection, and the main objective of the algorithm is to get a high-quality solution for the optimisation problem. The present work has a multi-optical combination of the process parameters; hence, the ANN-GA model must get the optimal parameters [54]. An ANN-GA model is prepared with an initial population size of 50 in the first iteration with a 0.8 crossover rate and a mutation of 0.01. The flow chart for the combined ANN-GA approach is shown in Fig. 11.

Table 5 contains the optimum value of the parameters obtained after the ANN-GA model. The two main objectives were to minimise ($1/\%\Delta R_a$) and the second to maximise MR (Fig. 12 and Table 6).

Table 5 indicates that the maximum limit for the percentage improvement is more than 44 per cent, which shows the ability of the proposed electrode to achieve a high range of surface finish compared to a straight spline electrode in the same process. Also, the range of MR is about 8–13 mg, which is very effective compared to the conventional AFM process. Figure 13 shows the SEM images of the machined and the unmachined surfaces of the workpiece. Figure 14 shows the various plot of surface roughness obtained from the Taylor Hobson Talysurf after machining the brass workpiece. Hence, it justifies the effectiveness of the TACAFM process in $\%\Delta R_a$ improvement.

Figure 13(i) shows the unmachined internal surface of the brass workpiece. After breaking the workpiece longitudinally, the images focus on its curved inner surface. Initially, roughness and surface irregularity marks are observed on the workpiece. When the workpiece is machined on the TACAFM process, the combination of axial force due to the media extrusion [56], centrifugal force due to rotation of the CFG electrode [57] and the Coriolis force due to the combination of rotation and extrusion [14] takes place. These effective forces on the workpiece increase the resultant cutting force, resulting in MR enhancement. The spark generation enhances this machining process, improves surface quality, and caters to the considerable machining time [58]. Due to the generation of spark, molten metal and oxide layer will be present on the finished workpiece's surface, shown in Fig. 13(ii). The abrasive marks will appear as they will abrade the surface of the workpiece.

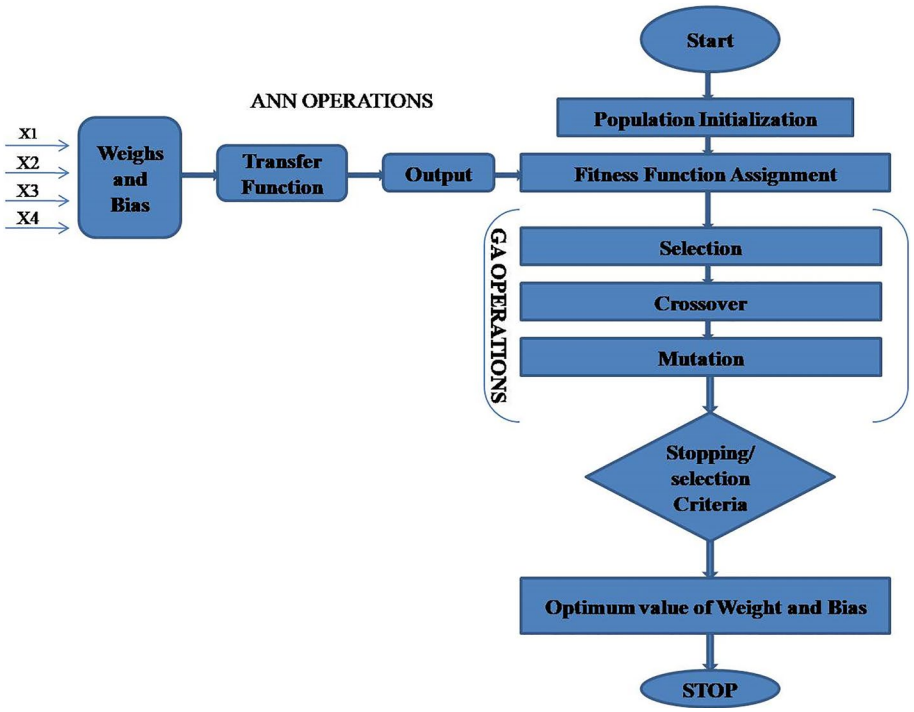


Fig. 12 Flow chart explaining ANN-GA process

Table 6 Set of optimal solutions corresponding to inputs and outputs

Sr. no.	The rotational speed of the electrode	Supply current	Number of oscillations	MR	$\%\Delta R_a$
1	29.51	14.13	3.43	8.87	44.74
2	44.40	11.02	6.74	13.50	43.08
3	24.95	14.08	4.01	10.48	44.65
4	24.52	14.04	4.86	11.79	44.34
5	26.33	14.05	3.67	9.75	44.71
6	26.01	14.05	3.51	9.56	44.72
7	25.94	14.02	5.45	12.21	43.98
8	24.95	13.81	5.01	11.90	44.15
9	24.88	13.59	6.48	12.63	43.29
10	43.08	11.06	6.75	13.24	43.09
11	26.03	14.06	3.98	10.25	44.66
12	26.52	13.94	3.53	9.45	44.73
13	24.70	14.05	4.61	11.41	44.46
14	25.62	14.07	4.80	11.48	44.37
15	25.44	14.06	3.76	10.03	44.70
16	41.19	11.92	6.67	12.79	43.13
17	26.73	14.04	3.67	9.67	44.72
18	26.00	14.05	4.51	10.97	44.49

Also, the patterns will be randomly dispersed due to the helical path of the abrasive over the workpiece [48], which is evident from Fig. 13(iii) as it shows the abrasive marks on the finished surface. The straight line appearing on the surface of the workpiece shows the machining marks generated due to the sparks. Figure 13(iv) shows the spark marks on the surface of the finished workpiece. The spark produced will be uniform around the workpiece as the workpiece is rotated. This uniformity of spark around the workpiece would result in the uniform material removal from the workpiece. Figure 14 shows the surface roughness plot of the finished workpiece obtained from Taylor Hobson Talysurf.

4 Conclusions

In this research work, an effort has been made to simulate and experiment on the TACAFM through a newly proposed CFG electrode. After critically analysing the results, the following conclusion can be drawn.

1. It is evident from the literature review and the experimental results that the material removal for the conventional AFM is low and the machining time is high; thus, hybridisation is needed to improve MR.
2. With the improvement in material removal, there is a requirement for the optimum surface finish value, which needs to cater. Thus, a suitable hybridisation process is required to be chosen. Thermal additive centrifugal AFM is an appropriate process in this regard.
3. The simulation results for the temperature variations show that the proposed electrode can carry the EDM effect. It is evident from the temperature profile result

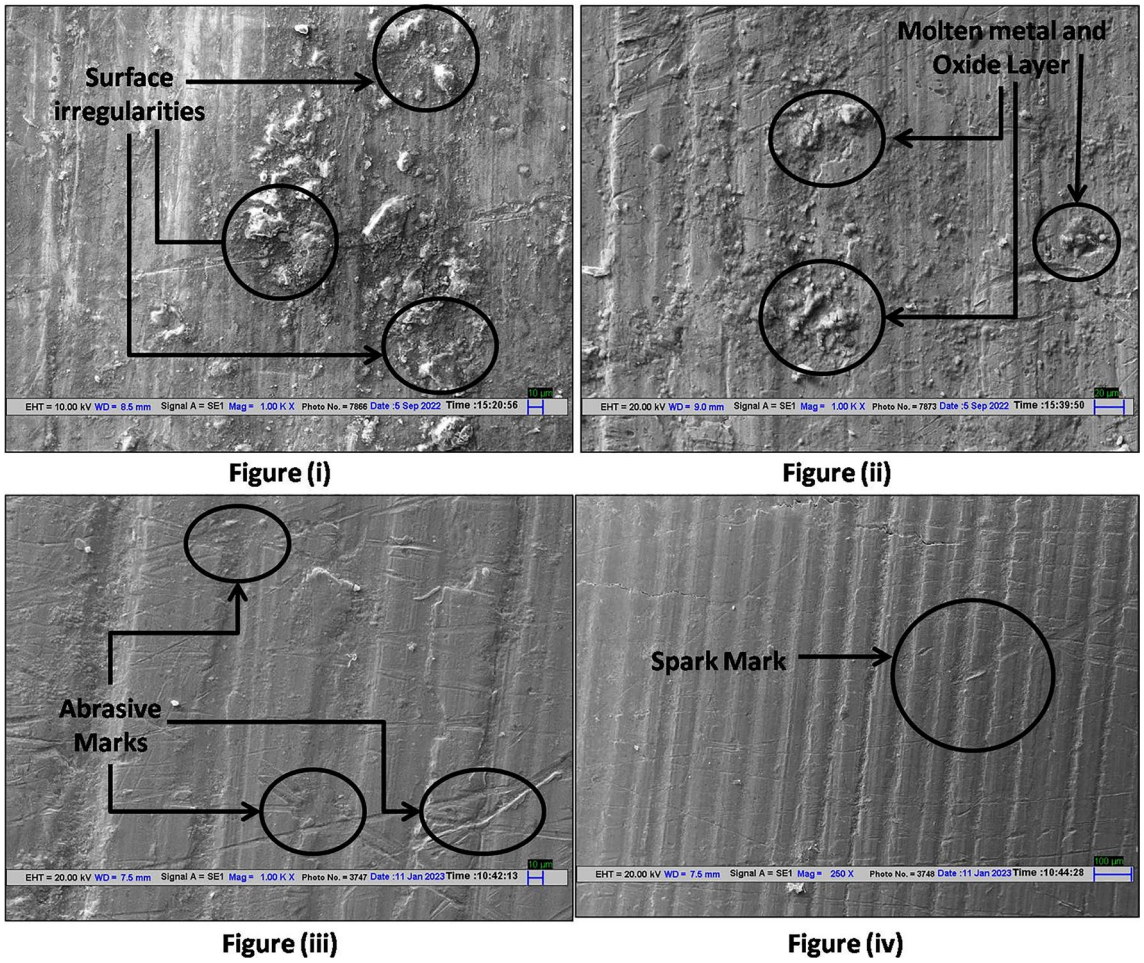


Fig. 13 SEM images of (i) unmachined surface, (ii) finished surface at 1000× magnification, (iii) finished surface at 500×, (iv) finished surface at 250×

that the spark is produced at the flanges, which justifies the occurrence of the TACAFM process.

4. The simulation results for the pressure variation clarify that under the same condition, the pressure on the workpiece is higher by the proposed electrode than that with the straight spline electrode. This indicates that the proposed electrodes give more MR than one with a straight spline electrode.
5. The streamlined curves highlight the increased number of abrasive particle particles near the inner surface of the workpiece, and the temperature distribution over the media justifies the machining process at the workpiece by spark generation.
6. The artificial neuron network is used to train, test and validate the data obtained after the experiment, and it was found that the predicted data follow the experimental data.
7. The maximum percentage error for the computation of MR is 3.87 per cent, while in the case of the percentage improvement in surface roughness, it is about 2.65 per cent.
8. The average experimental value for the MR and $\%\Delta R_a$ is 9.77 mg and 42.13%, respectively. The average predicted value for MR and $\%\Delta R_a$ is 9.73 mg and 42.14%, respectively, and the value of cumulative average absolute per cent error for MR and $\%\Delta R_a$ is 0.49 and 1.05%, respectively. The cumulative RMSE value for MR and $\%\Delta R_a$ is 0.034853 and 0.06501, respectively.
9. From the ANN-GA optimisation, optimal MR and $\%\Delta R_a$ can be predicted. The predicted results are that the $\%\Delta R_a$ is about 44 per cent, and the maximum MR is about 13 mg, which is significant compared to the conventional AFM process.

Fig. 14 R_a plots for machined surface after TACAFM process

10. The surface roughness plot obtained from the Taylor Hobson Talysurf justifies the surface roughness capability of the TACAFM process. Also, it gives an edge to the proposed electrode over the other for the TACAFM process.

Acknowledgements The authors thank the Delhi Technological University for providing a Research fellowship to cater for the financial requirement.

References

1. Parvesh A, Dhull S, Walia RS, Qasim M, Mohit T (2017) Hybrid AFM for nano finishing - a review. Mater Today Proc 4:7208–7218. https://doi.org/10.1016/j.matpr.2017.07.048
2. Anant B, Parvesh Ali, Walia RS, Qasim M, Pandey SM (2019) Development of hybrid forms of AFM process: a review. In: Shanker K, Shankar R, Sindhwani R (eds) Advances in industrial and production engineering. Springer Singapore, Singapore, pp 41–67
3. Walia RS, Shan HS, Kumar P (2008) Morphology and integrity of surfaces finished by centrifugal force assisted abrasive flow machining. Int J Adv Manuf Technol 39:1171–1179. https://doi.org/10.1007/s00170-007-1301-8
4. Walia RS, Shan HS, Kumar P (2006) AFMwith additional centrifugal force applied to the media. Mach Sci Technol 10(3):341–354. https://doi.org/10.1080/10910340600902157
5. Singh S, Shan HS (2002) Development of magneto AFM process. Int J Mach Tools Manuf 42:953–959. https://doi.org/10.1016/S0890-6955(02)00021-4
6. Brar BS, Walia RS, Singh VP, Sharma M (2013) A robust helical AFM(HLX-AFM) process. J Inst Eng India Ser C 94:21–29. https://doi.org/10.1007/s40032-012-0054-9
7. Brar BS, Walia RS, Singh VP (2015) Electrochemical-aided AFM(ECA2FM) process: a hybrid machining process. Int J Adv Manuf Technol 79:329–342. https://doi.org/10.1007/s00170-015-6806-y
8. Vaishya R, Walia RS, Kalra P (2015) Design and development of hybrid electrochemical and centrifugal force assisted abrasive flow machining. Mater Today Proc 2:3327–3341. https://doi.org/10.1016/j.matpr.2015.07.158
9. Venkatesh G, Sharma AK, Kumar P (2015) On ultrasonic assisted abrasive flow finishing of bevel gears. Int J Mach Tools Manuf 89:29–38. https://doi.org/10.1016/j.ijmachtools.2014.10.014

10. Sankar MR, Mondal S, Ramkumar J, Jain VK (2009) Experimental investigations and modelling of drill bit-guided abrasive flow finishing (DBG-AFF) process. Int J Adv Manuf Technol 42:678–688. https://doi.org/10.1007/s00170-008-1642-y
11. Ravi S, Jain VK, Ramkumar J (2010) Rotational abrasive flow finishing (R-AFF) process and its effects on finished surface topography. Int J Mach Tools Manuf 50:637–650. https://doi.org/10.1016/j.ijmachtools.2010.03.007
12. Singh R, Walia RS, Suri NM (2015) Parametric optimization of centrifugal-magnetic force assisted abrasive flow machining process using utility concept. Int J Res Eng Technol 4(8):2321–2730. https://doi.org/10.15623/ijret.2015.0408065
13. Ali P, Walia RS, Murtaza Q, Ranganath MS (2020) Modeling and analysis of developed thermal additive centrifugal abrasive flow machining process. Surf Topogr Metrol Prop 8(3):035013. https://doi.org/10.1088/2051-672X/abaffe
14. Parvesh A, Walia RS, Qasim M, Ranganath MS (2020) Material removal analysis of hybrid EDM-assisted centrifugal AFM process for performance enhancement. J Braz Soc Mech Sci Eng 42:302. https://doi.org/10.1007/s40430-020-02375-6
15. Parvesh A, Walia RS, Qasim M, Ranganath MS (2022) Characterisation of finished surface through thermal additive centrifugal AFM for better surface integrity. Indian J Eng Mater Sci 29(1):29–44
16. Parvesh A, Walia RS, Qasim M, Ranganath MS (2018) Enhancement of surface integrity using spark assisted AFM process. Int J Mech Prod Eng Res Dev (IJMPERD) 8(6):635–642
17. Parvesh A, Pandey SM, Ranganath MS, Walia RS, Qasim M (2020) Experimentation and modelling of CNT additive abrasive media for micro finishing. Measurement 151:107133. https://doi.org/10.1016/j.measurement.2019.107133
18. Krishna GG (2022) Optimisation of process parameters in AFM of SS 304L using Taguchi method and grey relation analysis. Mater Today Proc 58:335–338. https://doi.org/10.1016/j.matpr.2022.02.240
19. Kar KK, Ravikumar NL, Tailor PB, Ramkumar J, Sathiyamoorthy D (2009) Performance evaluation and rheological characterisation of newly developed butyl rubber based media for AFMprocess. J Mater Process Technol 209:2212–2221. https://doi.org/10.1016/j.jmatprotec.2008.05.012
20. Uhlmann E, Mihotovic V, Roßkamp S, Dethlefs A (2016) A pragmatic modeling approach in AFM for complex-shaped automotive components. Proced CIRP 46:51–54. https://doi.org/10.1016/j.procir.2016.03.172
21. Venkatesh G, Sharma AK (2014) Finishing of bevel gears using abrasive flow machining. Proced Eng 97:320–328. https://doi.org/10.1016/j.proeng.2014.12.255
22. Wahab Hashmi A, Singh Mali H, Meena A (2022) Experimental investigation on AFM(AFM) of FDM printed hollow truncated cone parts. Mater Today Proc 56:1369–1375. https://doi.org/10.1016/j.matpr.2021.11.428
23. Bhardwaj A, Srinivas K, Chaudhary R (2021) Modelling of wind turbine magnet for magnetic abrasive finishing and magnetic field assisted AFM process. JER https://doi.org/10.36909/jer.ICARI.15291
24. Bhardwaj A, Srinivas K, Chaudhary R (2021) Modelling and simulation of abrasive flow machining. Int J Adv Prod Ind Eng 6(2):01–09
25. Srinivas K, Bharadwaj A (2018) Pressure variation in abrasive flow machining: modelling and simulation. IJAPIE 3(4):15–19
26. Bhardwaj A, Computational and experimental analysis of parameters in centrifugal force assisted abrasive flow machining process. Diss. 2019, Delhi Technological University
27. Zhenzhong Z, Xianpeng Z, Xiao L, Peifa Z, Peng Y, Chuanzhen H, Hanlian L, Hongtao Z, Zou B (2019) Preparation and application of a novel water-based viscoelastic polishing fluid for abrasive flow machining. CIRP Bio Manuf Conf. https://doi.org/10.1016/j.procir.2020.05.153
28. Ahmad S, Singari RM, Mishra RS (2021) Development of Al_2O_3–SiO_2 based magnetic abrasive by sintering method and its performance on Ti-6Al-4V during magnetic abrasive finishing. Trans IMF 99(2):94–101. https://doi.org/10.1080/00202967.2021.1865644
29. Ahmad S, Gangwar S, Yadav PC, Singh DK (2017) Optimisation of process parameters affecting surface roughness in magnetic abrasive finishing process. Mater Manuf Process 32(15):1723–1729. https://doi.org/10.1080/10426914.2017.1279307
30. Nitin D, Varun S, Pradeep K (2021) Development and characterisation of Xanthan Gum based abrasive media and performance analysis using abrasive flow machining. J Manuf Process 67:101–115. https://doi.org/10.1016/j.jmapro.2021.04.053
31. Pawan Y, Jayswal SC (2019) Study of media in AFM and its hybrid processes. Mater Today Proc 18:3017–3026. https://doi.org/10.1016/j.matpr.2019.07.173
32. Youzhi Fu, Hang G, Qiusheng Y, Xuanping W (2019) A new predictive method of the finished surface profile in abrasive flow machining process. Precis Eng 60:497–505. https://doi.org/10.1016/j.precisioneng.2019.08.011
33. Aneesh G, Harvinder S, Rachin G, Rajdeep S, Swarn S (2022) Recent advancements in AFM and abrasive materials: a review. Mater Today Proc 56:3065–3072. https://doi.org/10.1016/j.matpr.2021.12.109
34. Swarn S, Harish K, Santosh K, Saurabh C (2022) A systematic review on recent advancements in Abrasive Flow Machining (AFM). Mater Today Proc 56:3108–3116. https://doi.org/10.1016/j.matpr.2021.12.273
35. Mohseni-Mofidi S, Pastewka L, Teschner M, Bierwisch C (2022) Magnetic-assisted soft AFM studied with smoothed particle hydrodynamics. Appl Math Model 101:38–54. https://doi.org/10.1016/j.apm.2021.07.015
36. Manjunath MA, Vinod P, Balashanmugam N, Sankar MR (2020) Abrasive flow finishing for surface roughness improvement of aluminium propeller: a case study. Mater Today Proc 26:1113–1118. https://doi.org/10.1016/j.matpr.2020.02.223
37. Baraiya R, Atul B, Jain V, Gupta D (2020) In-situ simultaneous surface finishing using AFMvia novel fixture. J Manuf Process 50:266–278. https://doi.org/10.1016/j.jmapro.2019.12.051
38. Han S, Salvatore F, Rech J, Bajolet J (2020) Abrasive flow machining (AFM) finishing of conformal cooling channels created by selective laser melting (SLM). Precis Eng 64:20–33. https://doi.org/10.1016/j.precisioneng.2020.03.006
39. Sangil H, Ferdinando S, Joel R, Julien B, Joel C (2020) Effect of AFM finish of selective laser melting (SLM) internal channels on fatigue performance. J Manuf Process. https://doi.org/10.1016/j.jmapro.2020.09.065
40. Kuma CW, Wu CH, Wan S, Kang CW (2020) Prediction and compensation of material removal for AFMof additively manufactured metal components. J Mater Process Tech. https://doi.org/10.1016/j.jmatprotec.2020.116704
41. Basha SM, Venkaiah N, Sankar MR (2023) Development and performance evaluation of Galactomannan polymer based abrasive medium to finish atomic diffusion additively manufactured pure copper using abrasive flow finishing. Addit Manuf 61:103290. https://doi.org/10.1016/j.addma.2022.103290
42. Bhardwaj A, Srinivas K, Chaudhary R (2023) Analysis of shapes of centrifugal force-generating rod in centrifugal force-assisted abrasive flow machining process. Mapan 38:459–479. https://doi.org/10.1007/s12647-023-00626-0
43. Bhardwaj A, Srinivas K, Chaudhary R (2023) Analysis of thermal additive centrifugal abrasive flow machining process with improved electrode geometry. Proc Inst Mech Eng Part E J Process Mech Eng. https://doi.org/10.1177/09544089231195171

44. Amjadimanesh A, Ajam H, Nezhad H (2014) Effect of blade type on a 3D FC centrifugal fan. J Serb Soc Comp Mech 8:23–35. https://doi.org/10.5937/jsscm1401023A
45. Wang Z, Zhu C (2017) Study of the effect of centrifugal force on rotor blade icing process. Int J Aerosp Eng 2017:1–9. https://doi.org/10.1155/2017/8695170
46. Gaurav PC, Vejendla PSB (2022) Deformation and annealing of brass. Mater Today Proc 64:1380–1383. https://doi.org/10.1016/j.matpr.2022.04.502
47. Kohut T (1988) Surface finishing with Abrasive flow machining. In: SME Technical Paper Proceedings of 4th international aluminium extraction technology seminar, Washington D.C, pp 35–42
48. Walia RS, Shan HS, Kumar PK (2008) Determining dynamically active abrasive particles in the media used in centrifugal force assisted AFMprocess. Int J Adv Manuf Technol 38:1157–1164. https://doi.org/10.1007/s00170-007-1184-8
49. Walia RS, Shan HS, Pradeep K (2006) Parametric optimization of centrifugal force-assisted AFM(CFAAFM) by the Taguchi method. J Mater Manuf Processes 21:4. https://doi.org/10.1080/10426910500411645
50. Chen Lv, Yang X, Junzhi Z, Xiaoxiang Na, Yutong Li, Teng L, Dongpu C, Fei-Yue W (2018) Levenberg-Marquardt backpropagation training of multilayer neural networks for state estimation of a safety critical cyber-physical system. IEEE Trans Ind Inform 14(8):3436–3446. https://doi.org/10.1109/TII.2017.2777460
51. Ahmad S, Singari RM, Mishra RS (2020) Modelling and optimisation of magnetic abrasive finishing process based on a non-orthogonal array with ANN-GA approach. Trans IMF 98(4):186–198. https://doi.org/10.1080/00202967.2020.1776966
52. Singh RK, Gangwar S, Singh DK, Ahmad S (2021) Multi-objective optimisation of magnetic abrasive finishing by desirability function with genetic algorithm. Int J Eng Sci Technol 13(1):1–9. https://doi.org/10.4314/ijest.v13i1.1S
53. Tayyab M, Ahmad S, Akhtar MJ, Sathikh PM, Singari RM (2022) Prediction of mechanical properties for acrylonitrile-butadiene-styrene parts manufactured by fused deposition modelling using artificial neural network and genetic algorithm. Int J Comput Integr Manuf. https://doi.org/10.1080/0951192X.2022.2104462
54. Hardik S, Bhadauria SS, Sharma VS (2021) Artificial neural network (ANN) based prediction of process parameters in additive manufacturing. IOP Conf Ser Mater Sci Eng 1136:012026. https://doi.org/10.1088/1757-899X/1136/1/012026
55. Ahmad S, Singari RM, Mishra RS (2021) Tri-objective constrained optimisation of pulsating DC sourced magnetic abrasive finishing process parameters using artificial neural network and genetic algorithm. Mater Manuf Processes 36(7):843–857. https://doi.org/10.1080/10426914.2020.1866196
56. Dirk B, Horst B, Martin S (2012) Investigation of one-way AFM and in-process measurement of axial forces. Proced CIRP 1:419–424. https://doi.org/10.1016/j.procir.2012.04.075
57. Walia RS, Shan HS, Kumar PK (2006) Finite element analysis of media used in the centrifugal force assisted AFMprocess. Proc Inst Mech Eng Part B J Eng Manuf 220:1775–1785. https://doi.org/10.1243/09544054JEM325
58. Babouri L, Belmokre K, Kabir A, Abdelouas A, Khettabi R, El Mendili Y (2018) Microstructure and crystallographic properties of Cu77Zn21 alloy under the effect of heat treatment. Mater High Temp. https://doi.org/10.1080/09603409.2018.1499243

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

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

Observational Constraints and Cosmological Implications of Scalar-Tensor $f(R,T)$ Gravity

Amine Bouali[1]★, Himanshu Chaudhary[2,3,4]†, Tiberiu Harko[5,6,7]‡, Francisco S. N. Lobo[8,9]§, Taoufik Ouali[1]¶, and Miguel A. S. Pinto[8,9]‖

[1]*Laboratory of Physics of Matter and Radiation, Mohammed I University, BP 717, Oujda, Morocco*
[2]*Department of Applied Mathematics, Delhi Technological University, Delhi-110042, India*
[3]*Pacif Institute of Cosmology and Selfology (PICS) Sagara, Sambalpur 768224, Odisha, India,*
[4]*Department of Mathematics, Shyamlal College, University of Delhi, Delhi-110032, India.*
[5]*Department of Physics, Babes-Bolyai University, Kogalniceanu Street, Cluj-Napoca 400084, Romania*
[6]*Department of Theoretical Physics, National Institute of Physics and Nuclear Engineering (IFIN-HH), Bucharest, 077125 Romania*
[7]*Astronomical Observatory, 19 Ciresilor Street, Cluj-Napoca 400487, Romania*
[8]*Instituto de Astrofísica e Ciências do Espaço, Faculdade de Ciências da Universidade de Lisboa, Edifício C8, Campo Grande, P-1749-016 Lisbon, Portugal*
[9]*Departamento de Física, Faculdade de Ciências da Universidade de Lisboa, Edifício C8, Campo Grande, P-1749-016 Lisbon, Portugal*

arXiv:2309.15497v1 [gr-qc] 27 Sep 2023

28 September 2023

ABSTRACT

Recently, the scalar-tensor representation of $f(R,T)$ gravity was used to explore gravitationally induced particle production/annihilation. Using the framework of irreversible thermodynamics of open systems in the presence of matter creation/annihilation, the physical and cosmological consequences of this setup were investigated in detail. In this paper, we test observationally the scalar-tensor representation of $f(R,T)$ gravity in the context of the aforementioned framework, using the Hubble and Pantheon+ measurements. The best fit parameters are obtained by solving numerically the modified Friedmann equations of two distinct cosmological models in scalar-tensor $f(R,T)$ gravity, corresponding to two different choices of the potential, and by performing a Markov Chain Monte Carlo analysis. The best parameters are used to compute the cosmographic parameters, i.e., the deceleration, the jerk and the snap parameters. Using the output resulting from the Markov Chain Monte Carlo analysis, the cosmological evolution of the creation pressure and of the matter creation rates are presented for both models. To figure out the statistical significance of the studied scalar-tensor $f(R,T)$ gravity, the Bayesian and the corrected Akaike information criteria are used. The latter indicates that the first considered model in scalar-tensor $f(R,T)$ gravity is statistically better than ΛCDM, i.e., it is more favored by observations. Besides, a continuous particle creation process is present in Model 1. On the other hand, for large redshifts, in Model 2 the particle creation rate may become negative, thus indicating the presence of particle annihilation processes. However, both models lead to an accelerating expansion of the Universe at late times, with a deceleration parameter equivalent to that of the ΛCDM model.

Key words: Observations – Dark Energy – Methods: Statistical – Methods: Numerical – Methods: Data Analysis – Miscellaneous

1 INTRODUCTION

General Relativity (GR) (Einstein 1915) can be unequivocally considered as one of the most important physical theories developed in recent times, and it is the standard theoretical foundation for the modern comprehension of gravitational phenomena. Essentially, GR is a geometric theory of gravity, in which the gravitational interaction is explained by assuming a Riemannian structure of the spacetime. The recent detection of the gravitational waves by the LIGO (Laser Interferometer Gravitational-Wave Observatory) and Virgo collaboration (Abbott et al. 2016), which started a new and important era for the astronomical, astrophysical, and cosmological probes of the Universe, confirmed once again the major predictive power of this theory. However, despite its remarkable successes and achievements, GR faces several problems on both small and large scales. On the former, we have the presence of singularities in black hole solutions, like, for example, in the Schwarzschild solution (Schwarzschild 1916). The appearance of singularities is unavoidable in GR (Hawking & Penrose 1970; Hawking & Ellis 2023). We must also take into account the fact that the theory is not renormalizable at high energies, and therefore cannot be quantized (Shomer 2007). Moreover, while all the standard model interactions are conformally invariant at the classical level, GR is not (Ghilencea 2021, 2022, 2023). Hence, from the perspective of GR, the gravitational interaction cannot be understood,

★ E-mail: a1.bouali@ump.ac.ma
† E-mail: himanshuch1729@gmail.com
‡ E-mail: t.harko@ucl.ac.uk
§ E-mail: fslobo@fc.ul.pt
¶ E-mail: t.ouali@ump.ac.ma
‖ E-mail: mapinto@fc.ul.pt

© 2023 The Authors

and treated in the same manner as the remaining fundamental interactions, which leads to fundamental problems in the quantization of the theory.

More recently, based on observations of Supernovae Type IA, it was found that the Universe is presently undergoing an accelerated expansion that is thought to be caused by an exotic fluid, dubbed as dark energy (Riess et al. 1998; Perlmutter et al. 1999). Einstein's cosmological constant Λ (Einstein 1986) is a particular type of this fluid, whose physical nature is still yet to be understood. In fact, the cosmological constant is a subject of great controversy, since if we assume that it is the vacuum energy, then the theoretical value predicted by quantum field theory is largely (of an order of 10^{120}) different from the observational value obtained from cosmological analysis, which, in short, constitutes the famous Cosmological Constant Problem (Weinberg 1989; Carroll 2001; Martin 2012). Furthermore, GR is unable to describe correctly the galactic rotation curves (Salucci et al. 2011; Persic et al. 1996; Borriello & Salucci 2001), without assuming the presence of a non-luminous matter, called as dark matter, to fill the mass gap that is supposedly missing. Dark matter is assumed to interact only gravitationally with normal baryonic matter. The variety of theoretical models for this kind of matter is very rich, with the main candidates being WIMPs (Weakly Interacting Massive Particles), sterile neutrinos, axions, and supersymmetric particles (for reviews of the many proposed particles that could be dark matter see (Overduin & Wesson 2004; Beylin et al. 2020; Lebedev 2021; Bian et al. 2021). Still, so far, no terrestrial experiment has directly detected dark matter particles, or any of their possible signatures (for a recent review on the state of the art regarding dark matter searches see Cebrián (2023).

Recently, the ΛCDM model must also face a number of other important challenges, resulting from the significant increase of the precision of the cosmological observations. One of the most challenging present day cosmological problems is the important difference between the expansion rate of the Universe as determined from the Cosmic Microwave Background (CMB) experiments and the low redshift (local) measurements. These important differences in the determinations of the present-day value of the Hubble constant led to what is generically called the Hubble tension. For a detailed discussion of the Hubble tension, and of its possible solutions see (Di Valentino et al. 2021) and references therein. The Hubble constant (H_0) as measured by the Planck satellite has the value of 66.93 ± 0.62 km/ s/ Mpc, while the SHOES collaboration obtained a value of 73.24 ± 1.74 km/ s/ Mpc (Di Valentino et al. 2021). If the Hubble tension is real, it strongly points toward the necessity of investigating new gravitational theories, and of the replacement the ΛCDM model with alternative cosmological models.

A possible path in an attempt to solve the aforementioned problems is to modify the Einstein-Hilbert action. A common example is the $f(R)$ modified theory of gravity (Sotiriou & Faraoni 2010; De Felice & Tsujikawa 2010), in which one substitutes the Ricci scalar, R, by an arbitrary and well-behaved function of the same quantity, $f(R)$. Such an extension allows for a broader and more diverse gravitational dynamics at low energies, which can, for example, explain the late-time cosmic acceleration without the necessity of considering the existence of a cosmological constant (Capozziello 2002). Indeed, admitting that the nature of the gravitational interaction changes on cosmological scales could pave the way to new breakthroughs in cosmology (Clifton et al. 2012; Avelino et al. 2016).

A particularly interesting class of modified theories of gravity are represented by those that contain nonminimal geometry-matter couplings, i.e., (direct) interaction terms between geometry (describing gravitational fields), and matter (non-gravitational fields) in the corresponding Lagrangian density (Koivisto 2006; Bertolami et al. 2007, 2008; Harko & Lobo 2010; Harko et al. 2013; Haghani et al. 2013; Odintsov & Sáez-Gómez 2013; Harko & Lobo 2014; Harko et al. 2014a,b, 2018; Harko & Lobo 2020; Harko 2014b; Harko et al. 2015, 2021). In particular, in these types of theories of gravity, the matter energy-momentum tensor is not conserved generally, and an extra-force is generated due to the presence of the geometry-matter coupling (Koivisto 2006; Bertolami et al. 2007). The most natural physical interpretation of the matter non-conservation is by assuming the presence of particle creation, which implies the irreversible matter/energy transfer from gravity (curvature) to matter fields. Hence, in models with geometry-matter couplings, the natural description of the thermodynamic phenomena occurs within the thermodynamics of open systems in the presence of irreversible processes. However, we would like to point out that presently the existence of the active matter creation processes in the Universe represents a hypothesis, which is not (yet) supported by any direct observational or experimental evidence.

A prominent example of a theory with a non-minimal curvature-matter coupling is the $f(R,T)$ gravity theory (Harko et al. 2011), where T is the trace of the energy-momentum tensor $T_{\mu\nu}$. As the name suggests, in this case, the Einstein-Hilbert action is generalized by the substitution of the Ricci scalar by a general $f(R,T)$ function. $f(R,T)$ gravity can be also seen as an extension of the $f(R)$ theory, in which now it is possible to have a direct exchange of energy and momentum between curvature and matter. The exploration of this theory in the usual geometrical representation was extensively done in the literature (see, for instance, (Houndjo 2012; Alvarenga et al. 2013; Shabani & Farhoudi 2014; Yousaf et al. 2016).

Additionally, the recently proposed scalar-tensor representation for this theory in (Rosa 2021) has triggered an intensive research effort (Rosa et al. 2021; Gonçalves et al. 2022b,a; Rosa & Kull 2022). The scalar-tensor representation not only helps in decreasing the complexity of the mathematical formalism by reducing the order of the metric in the field equations, but also allows to explore cosmological solutions by using a dynamical system approach. Moreover, the scalar-tensor $f(R,T)$ gravity has been recently considered in the context of irreversible matter creation, demonstrating capabilities of explaining the entropy production alongside a creation of cosmological matter, which can trigger the recent cosmological accelerated expansion of the Universe (Pinto et al. 2022). Although it is very important to formulate new modified theories of gravity, in order to reach a more complete description of the gravitational phenomena, it is equally important to test and constrain the existing theories as well, by using the most recent, and the best available observational probes, such as, for example, the Hubble measurements and Supernova type Ia (i.e Pantheon+) data.

Given that, in this paper, we have performed a detailed comparison with the aforementioned cosmological observations of two particular cosmological models in the scalar-tensor representation of $f(R,T)$ gravity, in which the gravitational interaction is described by two scalar fields φ and ψ, and a self-interaction potential $U(\varphi,\psi)$. The cosmological models adopted in the present study correspond to two different choices of the interaction potential of the theory, which is assumed to take the forms $U(\varphi,\psi) = \alpha\varphi + \beta\varphi^2 - (1/2\gamma)\psi^2$, and $U(\varphi,\psi) = \alpha\varphi^n\psi^{-m} + \beta\varphi$, respectively, with the first potential having a purely additive structure, while the second potential has an additive-multiplicative structure in the two basic fields φ and ψ. Some general features of the cosmologies corresponding to these choices have been already investigated in (Pinto et al. 2022). In (Pinto et al. 2022), the particle creation rates, the creation pressure, the temperature evolution, and the comoving entropy have been obtained by using system-

atic covariant thermodynamic formulation. The thermodynamics of open systems was considered together with the field equations of the scalar-tensor representation of the $f(R,T)$ theory, and thus an extension of the ΛCDM paradigm was obtained. One of the important features of this approach is that the matter creation rates, as well as the creation pressure, are included in the energy-momentum tensor of the cosmological fluid. Specific cosmological models were also constructed by using the scalar-tensor $f(R,T)$ gravity theory, and a comparison with the ΛCDM model, as well as the observational data for the Hubble function was performed by using the trial and error approach for fixing the initial conditions of the generalized Friedmann equations, and of the model parameters. Moreover, the thermodynamic properties of the cosmological models in the presence of particle creation were investigated in both the high and the low redshift regimes.

It is important to point out that the cosmological models depend on several arbitrary constants (parameters) α, β and γ, whose values cannot be predicted theoretically, but must be obtained from a comparison with the observational data.

As a first step in our analysis of the $f(R,T)$ cosmological models, we have solved numerically the Friedmann equations in the scalar-tensor representation of the $f(R,T)$ gravity for the two given potentials. Then, by means of a Markov Chain Monte Carlo (MCMC) analysis, we have extracted the best fit values of the parameters of the two models α, β, and γ, and we have also obtained their mean values. Based on these statistical results, the two $f(R,T)$ gravity models were compared, analyzed, and classified with respect to the standard ΛCDM model by using the Bayesian and the corrected Akaike Information Criterion.

As a second step of this work, we have analyzed in detail the footprints of these fitting results on the current acceleration of the Universe by studying the evolution of the cosmographic parameters. Indeed, from the observational data, the cosmographic parameters can describe the past history and the evolution of the scale factor through a series of its derivatives. Furthermore, as they are model-independent tools, the kinematic parameters are used to discriminate between competing cosmological models describing the current cosmic speed up without assuming any cosmological model (Capozziello et al. 2019, 2020). Cosmography has already been subject of many studies in modified gravity such as in $f(R)$ gravity (Teppa Pannia et al. 2019; Pizza 2015; Pires et al. 2010; Aviles et al. 2013; Bouhmadi-Lopez et al. 2010; Capozziello et al. 2008), in $f(T)$ teleparallel gravity (Capozziello et al. 2011; Piedipalumbo et al. 2015), in $f(R,T)$ gravity theory (Farias & Moraes 2021), and, recently, in the Barthel-Kropina dark energy model (Bouali et al. 2023). Another powerful tool considered in this paper and recently used to discriminate between alternative dynamical dark energy models from ΛCDM is the $Om(z)$ diagnostic (Sahni et al. 2008, 2003). We have also investigated the $Om(z)$ diagnostic for the considered $f(R,T)$ models, and compared its predictions with the ΛCDM standard cosmology.

This paper is organized in the following manner. The modified field equations of $f(R,T)$ gravity in both geometrical and scalar-tensor representations are presented in Section 2.1. The general cosmological assumptions and equations are considered in Section 2.2. We consider two different functional forms for the scalar interaction potential, which specify the two particular cosmological model we are working with. The generalized Friedmann equations for two distinct models are also explicitly written down. In Section 3, we perform several cosmological tests to constrain the free parameters of the considered particular cosmological models. Finally, we discuss and conclude the results obtained in our paper in Section 4.

2 FIELD EQUATIONS OF $F(R,T)$ GRAVITY, AND THEIR COSMOLOGICAL IMPLICATIONS

In the present Section, we briefly review the fundamentals of the $f(R,T)$ modified theory of gravity, and of its scalar-tensor representation. The (non-)conservation equations of the matter energy-momentum tensor in these representations are also presented. In order to consider the observational tests of the theory we will consider the cosmological implications of the scalar-tensor representation of the $f(R,T)$, by assuming an isotropic, homogeneous and spatial flat geometry. The generalized Friedmann equations, describing the large scale evolution are introduced first in a general form. Two specific cosmological models, corresponding to two different choices of the self-interaction potential $U(\varphi,\psi)$ are also presented. These models will be used in the next Section as toy models for the cosmological tests of the theory.

2.1 Metric and scalar tensor representation of $f(R,T)$ gravity

In the present Subsection, we write down the basic relations in the two equivalent representations of the $f(R,T)$ gravity theory. Since we are working under the metric formalism, henceforth we denote the geometrical representation as the metric representation.

2.1.1 Metric representation

The action for the $f(R,T)$ gravity theory is postulated to be (Harko et al. 2011)

$$S=\frac{1}{2\kappa^2}\int_{\Omega}\sqrt{-g}\,f(R,T)d^4x+\int_{\Omega}\sqrt{-g}\,\mathcal{L}_m d^4x, \tag{1}$$

where we have denoted $\kappa^2=8\pi G/c^4$, with G the gravitational constant, and c is the speed of light in vacuum. Moreover, g is the determinant of the metric tensor $g_{\mu\nu}$, while Ω denotes the 4-dimensional spacetime Lorentzian manifold on which the gravitational field acts, and on which it is possible to define a set of coordinates $\{x^{\mu}\}$. The gravitational Lagrangian density is given by $f(R,T)$, where f is an arbitrary function of the Ricci scalar $R=g^{\mu\nu}R_{\mu\nu}$, with $R_{\mu\nu}$ denoting the Ricci tensor, and T is the trace $T=g^{\mu\nu}T_{\mu\nu}$ of the matter energy-momentum tensor $T_{\mu\nu}$. The matter energy-momentum tensor is defined in terms of the variation of the matter Lagrangian density $\mathcal{L}_m$ as

$$T_{\mu\nu}=-\frac{2}{\sqrt{-g}}\frac{\delta\left(\sqrt{-g}\mathcal{L}_m\right)}{\delta g^{\mu\nu}}. \tag{2}$$

Henceforth, we adopt a system of geometrized units in such a way that $G=c=1$, and thus $\kappa^2=8\pi$.

By applying the action variational principle, $\delta S=0$, the variation of Eq. (1) with respect to the metric tensor $g_{\mu\nu}$ yields the modified gravitational field equations see (Harko et al. 2011) for a detailed derivation)

$$f_R R_{\mu\nu}-\frac{1}{2}g_{\mu\nu}f(R,T) + \left(g_{\mu\nu}\Box-\nabla_{\mu}\nabla_{\nu}\right)f_R = \kappa^2 T_{\mu\nu}-f_T(T_{\mu\nu}+\Theta_{\mu\nu}), \tag{3}$$

where the subscripts R and T denote the partial derivatives of f with respect to the respective quantities, respectively, ∇_{μ} is the covariant derivative and $\Box\equiv\nabla^{\sigma}\nabla_{\sigma}$ denotes the D'Alembert operator in its standard form. Furthermore, the auxiliary tensor $\Theta_{\mu\nu}$, introduced in Eq. (3), is defined as

$$\Theta_{\mu\nu}\equiv g^{\rho\sigma}\frac{\delta T_{\rho\sigma}}{\delta g^{\mu\nu}}. \tag{4}$$

The conservation equation for $f(R,T)$ gravity can be obtained by taking the divergence of Eq. (3) and using the identity $(\Box\nabla_\nu - \nabla_\nu\Box) f_R = R_{\mu\nu}\nabla^\mu f_R$ (Koivisto 2006). By doing so, we obtain

$$(\kappa^2 - f_T)\nabla^\mu T_{\mu\nu} = \left(T_{\mu\nu} + \Theta_{\mu\nu}\right)\nabla^\mu f_T + f_T\nabla^\mu\Theta_{\mu\nu} + f_R\nabla^\mu R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}\nabla^\mu f. \quad (5)$$

By looking at the form of Eq. (5), it is evident that the matter energy-momentum tensor is not necessarily conserved. In fact, one should note that, in general, modified theories of gravity that may possess non-minimal geometry-matter couplings (Koivisto 2006; Bertolami et al. 2007, 2008; Harko & Lobo 2010; Harko et al. 2013; Haghani et al. 2013; Odintsov & Sáez-Gómez 2013; Harko & Lobo 2014; Harko et al. 2014a,b, 2018; Harko & Lobo 2020; Harko 2014b; Harko et al. 2015; Haghani et al. 2018; Harko et al. 2021) imply the possible non-conservation of the matter energy-momentum tensor, with $\nabla_\mu T^{\mu\nu} \neq 0$. This result is standardly interpreted as a transfer of energy from the geometry sector to the matter sector (Harko 2014b; Harko et al. 2015; Haghani et al. 2018; Harko et al. 2021).

Next, we consider the equivalent scalar-tensor representation of $f(R,T)$ gravity, which will be utilized henceforth.

2.1.2 Scalar-Tensor Representation

One of the interesting and important properties of modified theories of gravity, featuring extra scalar degrees of freedom in comparison to GR, is that one can deduce a dynamically equivalent scalar-tensor theory. As such, it is possible to construct an equivalent scalar-tensor action to Eq. (1), in which two extra scalar fields are added as fundamental fields, each one related to a degree of freedom of the metric representation of $f(R,T)$ gravity (Rosa 2021). In the particular case of this theory, the degrees of freedom are R and T.

To obtain the scalar-tensor representation, one introduces the first two auxiliary fields, η and ζ, respectively. Then one rewrites Eq. (1) in the following form

$$S = \frac{1}{2\kappa^2}\int_\Omega \sqrt{-g}\left[f(\eta,\zeta) + (R-\eta)f_\eta + (T-\zeta)f_\zeta\right]d^4x + \int_\Omega \sqrt{-g}\mathcal{L}_m d^4x, \quad (6)$$

where the subscripts η and ζ indicate the operation of partial derivation with respect to the considered variables. After taking the variation of the action (6) with respect to the fields η and ζ, we find that the equations of motion in these variables are

$$(R-\eta)f_{\eta\eta} + (T-\zeta)f_{\eta\zeta} = 0, \quad (7)$$

$$(R-\eta)f_{\zeta\eta} + (T-\zeta)f_{\zeta\zeta} = 0. \quad (8)$$

Now, we rewrite the system of Eqs. (7) and (8) in a matrix form, $\mathcal{M}\mathbf{u} = 0$ as

$$\begin{pmatrix} f_{\eta\eta} & f_{\eta\zeta} \\ f_{\zeta\eta} & f_{\zeta\zeta}\end{pmatrix}\begin{pmatrix} R-\eta \\ T-\zeta\end{pmatrix} = 0. \quad (9)$$

A unique solution of the above algebraic equation is obtained if and only if $\det\mathcal{M} \neq 0$, a condition which gives the relation $f_{\eta\eta}f_{\zeta\zeta} \neq f_{\eta\zeta}^2$, which must be satisfied by the function $f(\eta,\zeta)$. Hence, it follows that Eq. (9) has the unique solution given by $R=\eta$ and $T=\zeta$. After inserting these results back into Eq. (6), one easily verifies that this equation takes the form of the initial action (1). Thus, we have proved the equivalence between the two representations, and we have shown that the scalar-tensor representation is well-defined and mathematically consistent.

As a last step, we introduce two scalar fields, φ and ψ, as well as an interaction potential between them, $U(\varphi,\psi)$, according to the definitions

$$\varphi \equiv \frac{\partial f}{\partial R}, \qquad \psi \equiv \frac{\partial f}{\partial T}, \quad (10)$$

and

$$U(\varphi,\psi) \equiv -f(\eta,\zeta) + \eta\varphi + \zeta\psi, \quad (11)$$

respectively. Now we rewrite Eq. (6), and display the action of the equivalent scalar-tensor representation of $f(R,T)$ gravity as

$$S = \frac{1}{2\kappa^2}\int_\Omega \sqrt{-g}\left[\varphi R + \psi T - U(\varphi,\psi)\right]d^4x + \int_\Omega \sqrt{-g}\mathcal{L}_m d^4x. \quad (12)$$

The scalar field φ is the analogue of a Brans-Dicke type scalar field with parameter $\omega_{BD} = 0$, and with a self-interaction potential U (Faraoni & Belknap-Keet 2017). Additionally to the scalar field φ, in $f(R,T)$ gravity we have a second degree of freedom, corresponding to the arbitrary dependence of the action on T. This second degree of freedom is also related to another scalar field, ψ.

Accordingly to what was previously stated, the action (12) depends now on three independent fields, the metric tensor $g_{\mu\nu}$ and the two scalar fields φ and ψ. By varying the action Eq. (12) with respect to $g_{\mu\nu}$ yields the field equations

$$\varphi R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}(\varphi R + \psi T - U) - (\nabla_\mu\nabla_\nu - g_{\mu\nu}\Box)\varphi = \kappa^2 T_{\mu\nu} - \psi(T_{\mu\nu} + \Theta_{\mu\nu}). \quad (13)$$

It is also possible to derive the above field equation (13) directly by using Eq. (3) via the substitution of the quantities defined in Eqs. (10) and (11), respectively, by taking $\eta = R$ and $\zeta = T$. Furthermore, the variation of Eq. (12) with respect to the scalar fields φ and ψ gives,

$$U_\varphi = R, \qquad U_\psi = T. \quad (14)$$

In addition to this, by using Eqs. (10) and (11), with $\eta = R$ and $\zeta = T$, and the mathematical identity $\nabla^\mu\left[R_{\mu\nu} - (1/2)g_{\mu\nu}R\right] = 0$, we obtain the results that the conservation equation for $f(R,T)$ gravity in the scalar-tensor representation takes the form

$$\nabla^\mu T_{\mu\nu} = \frac{\left(T_{\mu\nu} + \Theta_{\mu\nu}\right)\nabla^\mu\psi}{(\kappa^2 - \psi)} + \frac{\psi\nabla^\mu\Theta_{\mu\nu}}{(\kappa^2-\psi)} - \frac{1}{2}g_{\mu\nu}\frac{R\nabla^\mu\varphi + \nabla^\mu(\psi T - U)}{(\kappa^2-\psi)}. \quad (15)$$

2.2 Cosmological evolution, and cosmological models

We assume that the metric of the Universe is given by the flat, homogeneous and isotropic Friedmann-Lemaitre-Robertson-Walker (FLRW) expression, which in spherical coordinates (t,r,θ,ϕ) is given by

$$ds^2 = -dt^2 + a^2(t)\left[dr^2 + r^2\left(d\theta^2 + \sin^2\theta d\phi^2\right)\right], \quad (16)$$

where $a(t)$ is the scale factor. Moreover, we introduce the Hubble function defined as $H = \dot{a}/a$, where, in the following, an overdot denotes the derivative with respect to the cosmological time t.

We also conjecture that the matter content of the Universe can be represented as an isotropic perfect fluid, with the energy-momentum tensor of the matter $T_{\mu\nu}$ given by

$$T_{\mu\nu} = (\rho + p)u_\mu u_\nu + p g_{\mu\nu}, \quad (17)$$

where ρ is the energy density of the cosmic matter, and p is its pressure. The four-velocity u^μ of the cosmological fluid is normalized according to $u_\mu u^\mu = -1$. As for the matter Lagrangian density we adopt the expression $\mathcal{L}_m = p$ (Bertolami et al. 2008). Our assumption of the isotropy of the matter distribution also implies that on average, and on large cosmological scales, the matter creation process are isotropic. However, particle production phenomena, quantum or classical, may not be necessarily isotropic. Anisotropic matter creation naturally generates inhomogeneities at small cosmological or galactic scales, and leads to the presence of anisotropic stresses, which could play an important, or event determinant role in the structure formation. However, on large cosmological scales, after averaging the matter density over large cosmological volumes, the cosmological dynamics can be described by the usual Friedmann-Lemaître-Robertson-Walker cosmology, in which the matter dynamics decouples completely from that describing particle creation inhomogeneities. Hence, in the following we consider that the possible physical or astrophysical properties inherited by the matter density from a possible anisotropic particle generation process, in the large scale phase, the average cosmological dynamics can be well described as taking place in a homogeneous and isotropic Universe. This assumption is also supported by the comparison of our model with the observational data, indicating that the particle creation model fits well the distance modulus observational data of type Ia supernovae, and thus it correctly describes the observed accelerated phase of the Universe.

Thus, the auxiliary tensor $\Theta_{\mu\nu}$ becomes

$$\Theta_{\mu\nu} = -2T_{\mu\nu} + pg_{\mu\nu}. \tag{18}$$

In a cosmological setting all geometrical and thermodynamical quantities must depend only on the cosmological time t. Consequently, in the present approach, $\rho = \rho(t)$, $p = p(t)$, $\varphi = \varphi(t)$, and $\psi = \psi(t)$, respectively.

Accordingly, the system of the cosmological field equations of the scalar-tensor representation of $f(R,T)$ gravity, representing the generalized Friedmann equations, takes the form (Pinto et al. 2022)

$$3H^2 = 8\pi\frac{\rho}{\varphi} + \frac{3\psi}{2\varphi}\left(\rho - \frac{1}{3}p\right) + \frac{1}{2}\frac{U}{\varphi} - 3H\frac{\dot{\varphi}}{\varphi}, \tag{19}$$

$$2\dot{H} + 3H^2 = -8\pi\frac{p}{\varphi} + \frac{\psi}{2\varphi}(\rho - 3p) + \frac{1}{2}\frac{U}{\varphi} - \frac{\ddot{\varphi}}{\varphi} - 2H\frac{\dot{\varphi}}{\varphi}, \tag{20}$$

$$U_\varphi = 6\left(\dot{H} + 2H^2\right), U_\psi = 3p - \rho, \tag{21}$$

$$\dot{\rho} + 3H(\rho + p) = \frac{3}{8\pi}\left\{-\frac{\dot{\psi}}{2}\left(\rho - \frac{p}{3} + \frac{U_\psi}{3}\right) - \psi\left[H(\rho + p) + \frac{1}{2}\left(\dot{\rho} - \frac{1}{3}\dot{p}\right)\right]\right\}. \tag{22}$$

To indicate the accelerating/decelerating nature of the cosmological expansion we will use the deceleration parameter q, defined as

$$q = \frac{d}{dt}\frac{1}{H} - 1 = -\frac{\dot{H}}{H^2} - 1. \tag{23}$$

From Eqs. (19) and (20), we find the following expression for the deceleration parameter

$$q = \frac{1}{2} + \frac{3\left[4\pi\frac{p}{\varphi} - \frac{\psi}{4\varphi}(\rho - 3p) - \frac{V}{4\varphi} + \frac{\ddot{\varphi}}{2\varphi} + H\frac{\dot{\varphi}}{\varphi}\right]}{8\pi\frac{\rho}{\varphi} + \frac{\psi}{2\varphi}(\rho - 3p) + \frac{V}{2\varphi} - 3H\frac{\dot{\varphi}}{\varphi}}. \tag{24}$$

In order to facilitate the comparison with the observational data, we will study the cosmological evolution in the redshift space, with the redshift coordinate z defined as

$$1 + z = \frac{1}{a}. \tag{25}$$

In the following we will also adopt the thermodynamic interpretation of $f(R,T)$ gravity, as introduced in (Pinto et al. 2022). Particle creation can be formulated in a thermodynamic formalism in which one introduces the creation pressure p_c, given by (Pinto et al. 2022)

$$p_c = \frac{\rho + p}{3H}\frac{\psi}{8\pi + \psi}\left(\frac{d}{dt}\ln\psi + \frac{1}{2}\frac{\dot{\rho} - \dot{p}}{\rho + p}\right), \tag{26}$$

which is related to the particle creation rate Γ through the relation (Lima et al. 2021)

$$p_c = -\frac{\rho + p}{3H}\Gamma. \tag{27}$$

Hence, for the particle creation rate we obtain the expression

$$\Gamma = -\frac{\psi}{8\pi + \psi}\left(\frac{d}{dt}\ln\psi + \frac{1}{2}\frac{\dot{\rho} - \dot{p}}{\rho + p}\right). \tag{28}$$

The expressions of the particle number creation rate Γ and of the creation pressure p_c follow directly from Eq.(15), expressing the non-conservation of the matter energy-momentum tensor. The present model does admit de Sitter type, exponentially expanding solutions. Two such classes of solutions can be obtained, under the assumption of a constant matter density, or, alternatively, in the presence of a time varying matter density. By assuming $H = H_0 =$ constant and $\rho = \rho_0 =$ constant, Eq. (20) gives for the potential the expression

$$U(\phi,\psi) = 12H_0^2\varphi - \rho_0\psi + \Lambda_0, \tag{29}$$

where Λ_0 is an arbitrary integration constant. With this form of the potential, describing de Sitter type expansion, the general solution of the field equations was found in (Pinto et al. 2022) and corresponds to a cosmological dynamics during which both the particle creation rate and the creation pressure decrease exponentially in time.

A second de Sitter type solution can be obtained for dust matter by assuming that the potential has the form (Pinto et al. 2022)

$$U(\phi,\psi) = 12H_0^2\varphi - \frac{1}{2\beta_1}\psi^2, \tag{30}$$

where β_1 is a constant. The derivative of the potential with respect to ψ gives immediately $U_\psi = -\psi/\beta_1$, while from Eq. (20) for $p = 0$ we obtain $U_\psi = -\rho$, thus obtaining $\psi = \beta_1\rho$. For this choice of the potential the matter density decreases exponentially, while the creation rate becomes either a constant, for $\beta_1\rho >> 8\pi$, $\Gamma \approx (9/5)H_0$, or decreases exponentially in the opposite limit as $\Gamma(t) \propto e^{-3H_0 t}$.

In the following we write down the generalized Friedmann equations for two distinct cosmological models, which we will use as toy models for testing the validity of $f(R,T)$ gravity on cosmological scales. The considered models correspond to choices of the potential that extend the de Sitter type forms of the potential (29) and (30) beyond the strict exponential phase, but still describe an accelerating Universe. From a physical point of view, for a Universe with time varying matter density, the potential ψ can be interpreted as an effective matter density. As for φ, as one can see from Eq. (20), it is fully determined by the expansion rate of the Universe, through the derivative of the potential. Hence, from a physical point of view, in a general sense, the potential describes the combined effects of the cosmological expansion rate, and of the matter density on the dynamical evolution of the Universe.

2.2.1 Model 1: $U(\varphi,\psi)=\alpha\varphi+\beta\varphi^2-(1/2\gamma)\,\psi^2$

As a first cosmological model in the scalar-tensor representation of $f(R,T)$ gravity we consider that the self-interacting potential U has a simple quadratic additive quadratic form, with

$$U(\varphi,\psi)=\alpha\varphi+\beta\varphi^2-\frac{1}{2\gamma}\psi^2, \tag{31}$$

where α, β and γ are arbitrary constants to be determined from the observations. For a dust Universe with $p=0$ we have $\psi=\gamma\rho$. This form of the potential generalizes the expression (30) of $U(\varphi,\psi)$ by adding to it an additive quadratic correction term in φ. The deviation of the potential from the de Sitter form leads to an accelerating Universe, with non-exponential scale factor, and with time varying matter density,

After introducing a set of dimensionless and rescaled variables $(h,\tau,r,\xi,\sigma,\epsilon)$, defined as

$$\begin{aligned} H &= H_0 h, \qquad \tau=H_0 t, \qquad \rho=\rho_c r,\\ \alpha &= 16\pi\rho_c\xi, \qquad \beta=8\pi\rho_c\sigma, \qquad \gamma=\frac{8\pi}{\rho_c}\epsilon, \end{aligned} \tag{32}$$

where H_0 is the present-day value of the Hubble function, and $\rho_c=3H_0^2/8\pi$, we obtain the generalized Friedmann equations of Model 1 in the redshift space as given by (Pinto et al. 2022)

$$-\frac{dh(z)}{dz}=\frac{\xi}{(1+z)h(z)}+\frac{\sigma\varphi(z)}{(1+z)h(z)}-2\frac{h(z)}{(1+z)}, \tag{33}$$

$$-\frac{dr(z)}{dz}=-3\frac{r(z)(1+\epsilon r(z))}{(1+z)\,(1+(5\epsilon/2)r(z))}, \tag{34}$$

and

$$\begin{aligned} -\frac{d\varphi(z)}{dz} &= \frac{r(z)}{(1+z)h^2(z)}+\frac{5\epsilon}{4}\frac{r^2(z)}{(1+z)h^2(z)}\\ &+\xi\frac{\varphi(z)}{(1+z)h^2(z)}+\frac{\sigma}{2}\frac{\varphi^2(z)}{(1+z)h^2(z)}-\frac{\varphi(z)}{(1+z)}. \end{aligned} \tag{35}$$

The system of differential equations (33)-(35) must be solved with the initial conditions $h(0)=1$, $r(0)=r_0$, and $\varphi(0)=\varphi_0$, respectively, where $r_0=\rho(0)/rho_c=\Omega_m$ is the density parameter of the baryonic matter. Several relations between the numerical values of the model parameters and the initial conditions can be obtained by using Eq. (21), from which we first obtain

$$U_\varphi=\alpha+2\beta\varphi=6\left(\dot{H}+2H^2\right)=6H^2(1-q), \tag{36}$$

which gives the condition

$$\varphi(0)=\frac{3H_0^2}{\beta}\,(1-q(0))-\frac{\alpha}{2\beta}. \tag{37}$$

For the initial value of ψ we obtain the relation

$$\psi(0)=\beta\rho(0)=\beta\rho_c r(0). \tag{38}$$

2.2.2 Model 2: $U(\varphi,\psi)=\alpha\varphi^n\psi^{-m}+\beta\varphi$

The second cosmological model which we will test observationally in the scalar-tensor representation of $f(R,T)$ gravity, corresponds to a self-interaction potential given by

$$U(\varphi,\psi)=\alpha\varphi^n\psi^{-m}+\beta\varphi, \tag{39}$$

with α, β, n and $m>0$ constants. This form of the potential extends the de Sitter type potential (29) by introducing a multiplicative algebraic coupling term in its structure, which extends the ψ dependence of $U(\varphi,\psi)$. The presence of this term modifies the exponential de Sitter type expansion rate, allows a time varying matter density, but still induces an accelerating expansion of the Universe. In the following we consider again only the pressureless case, $p=0$.

By introducing the set of dimensionless variables $(r,h,\tau,\alpha_0,\beta_0)$, defined by

$$H=H_0 h, \quad \tau=H_0 t, \quad \rho=\rho_c r, \quad \alpha=\rho_c\alpha_0, \quad \beta=\rho_c\beta_0, \tag{40}$$

where $\rho_c=3H_0^2/8\pi$ as before, we obtain the generalized Friedmann equations in the redshift space for Model 2 as (Pinto et al. 2022)

$$\begin{aligned} -\frac{dh(z)}{dz} &= \frac{n\alpha_0}{16\pi}\frac{\varphi^{n-1}(z)}{(1+z)h(z)\psi^m(z)}+\frac{\beta_0}{16\pi(1+z)h(z)}\\ &-2\frac{h(z)}{(1+z)}, \end{aligned} \tag{41}$$

$$\begin{aligned} -\frac{d\varphi(z)}{dz} &= \frac{\beta_0\varphi(z)}{16\pi(1+z)h^2(z)}-\frac{\varphi(z)}{(1+z)}\\ &+\alpha_0 m\frac{\varphi^n(z)}{(1+z)h^2(z)\psi^{m+1}(z)}\\ &+\frac{(1+3m)\alpha_0}{16\pi(1+z)h^2(z)}\frac{\varphi^n(z)}{\psi^m(z)}, \end{aligned} \tag{42}$$

and

$$\begin{aligned} -\frac{d\psi(z)}{dz} &= \frac{\psi(z)}{\varphi(z)}\left[\frac{6h(z)(\psi(z)+8\pi)\varphi(z)}{(1+z)h(z)\,[(3m+1)\psi(z)+16\pi(m+1)]}\right.\\ &\left.-\frac{n(3\psi(z)+16\pi)(1+z)h(z)\frac{d\varphi(z)}{dz}}{(1+z)h(z)\,[(3m+1)\psi(z)+16\pi(m+1)]}\right], \end{aligned} \tag{43}$$

respectively. The system of nonlinear differential equations (41)-(43) must be solved with the initial conditions $h(0)=1$, $\varphi(0)=\varphi_0$, and $\psi(0)=\psi_0$. Once these initial conditions are fixed, the system can be solved numerically.

Note that for this system the initial values of φ_0 and ψ_0 are not independent.

From the relations (21) we obtain first

$$U_\varphi=\alpha n\varphi^{n-1}\psi^{-m}+\beta=6H^2\,(1-q)\,, \tag{44}$$

which gives the following relation between the initial conditions and the model parameters

$$\alpha n\varphi_0^{n-1}\psi_0^{-m}+\beta=6H_0^2\,(1-q(0))\,. \tag{45}$$

The condition $U_\psi=-\rho$ gives the condition

$$\alpha m\varphi^n\psi^{-m-1}=\rho, \tag{46}$$

leading to

$$\alpha m\varphi_0^n\psi_0^{-m-1}=\rho_c r(0). \tag{47}$$

Hence, the condition (45) can be reformulated as

$$\frac{n}{m}\rho_c\frac{\psi_0}{\varphi_0}r(0)+\beta=6H_0^2\,(1-q(0))\,. \tag{48}$$

3 COSMOLOGICAL TESTS

In the present Section, we test the observational viability of the two previously introduced cosmological models in the framework of the scalar-tensor representation of $f(R,T)$ gravity. To this aim, a detailed statistical analysis comparing the theoretical predictions of the $f(R,T)$ gravity models with the cosmological observations

is performed. In other words, the parameters of the two models, namely, $(\xi, \sigma, \epsilon, h_0)$ for Model 1 and $(\alpha_0, \beta_0, \varphi, h_0)$ for Model 2, are constrained by observations, using the Markov Chain Monte Carlo (MCMC) approach. To avoid any possible confusion, we mention that the parameter h_0 is related to the current Hubble rate as $h_0 = H_0/100$. Furthermore, the obtained results are used in the cosmographic analysis of the models, i.e., by investigating the dynamical evolutions of the deceleration, the jerk, and the snap parameters, respectively. Using the cosmographic analysis, one can figure out some general features of the behavior of the cosmological models in scalar-tensor $f(R,T)$ gravity, thus allowing for an easy discrimination of this set of models from the ΛCDM paradigm. The observational datasets used to analyse the cosmological models in scalar-tensor $f(R,T)$ gravity are the Hubble measurement, $H(z)$, and the Pantheon+ data, consisting of 57 and 1701 data points, respectively.

3.1 Methodology and data description

Nowadays, cosmological theories are tested by using different observational data, collected from different cosmological probes, such as Supernovae type Ia, Baryon Acoustic oscillations (BAO), Cosmic Microwave Background (CMB) and Hubble rate measurements, respectively. Confronting theoretical models with observations is the only way to distinguish between observationally supported, and unsupported models. In order to efficiently compare theoretical models with the observational data, a detailed statistical analysis is required, involving a multi-level approach. As cosmology is Bayesian by its very nature, parameters estimation processes are, in general, performed in the Bayesian inference framework,

$$\mathcal{P}(H|D) \propto \mathcal{L}(D|H), \tag{49}$$

where $\mathcal{P}(H|D)$ is the posterior distribution, and $\mathcal{L}(D|H)$ is the likelihood function. In the case of Gaussian errors, the chi-square function, χ^2, and the likelihood function, $\mathcal{L}$, are related according to the relation

$$\chi^2(\theta) = -2\ln\mathcal{L}(\theta). \tag{50}$$

In order to estimate the parameters of a given theoretical model, one has to look for the free parameters vector, θ, which minimizes the chi-square function, or, equivalently, maximizes the posterior distribution. To constrain the cosmological parameters of the $f(R,T)$ gravity models, we minimize the χ^2 function, by using the Markov Chains Monte Carlo method (Nesseris et al. 2017),

$$\chi^2 = (\mathcal{P}_{obs} - \mathcal{P}_{th})^T \cdot \mathbf{C}^{-1} \cdot (\mathcal{P}_{obs} - \mathcal{P}_{th}), \tag{51}$$

where $\mathcal{P}_{obs}$, $\mathcal{P}_{th}$, and $\mathbf{C}$ denote the observed values, the predicted values, and the standard deviation, respectively.

To analyse a set of cosmological models, and in order to figure out the best model describing the observational data, a comparison criterion is usually adopted. In the current analysis, we use the most used corrected Akaike Information Criterion (AIC_c) (Akaike 1974; Vrieze 2012; Tan & Biswas 2012; Rezaei & Malekjani 2021; Arevalo et al. 2017; Tauscher et al. 2018),

$$AIC_c = \chi^2_{min} + 2\mathcal{K}_{\rm f} + \frac{2\mathcal{K}_{\rm f}(\mathcal{K}_{\rm f}+1)}{\mathcal{N}_{\rm t} - \mathcal{K}_{\rm f} - 1}, \tag{52}$$

where χ^2_{min}, $\mathcal{K}_{\rm f}$ and $\mathcal{N}_{\rm t}$ denote the minimum of the chi-square function, corresponding to the best fit values, the number of the free parameters, and the total number of the data points employed in the parameter inference process, respectively. When it comes to classify a group of cosmological models based on their statistical significance, the $AICc$ criterion is strongly recommended.

The model with the lowest AIC_c is the most supported by observations, and it is chosen to be the reference model. To measure how closely models resemble the reference one, we compute the quantity

$$\Delta AIC_c = AIC_{c,model} - AIC_{c,reference}.$$

Models are said to be well supported by observations if $0 < \Delta\mathrm{AIC}_c < 2$, less supported if $4 < \Delta\mathrm{AIC}_c < 7$, and not supported if $\Delta\mathrm{AIC}_c > 10$, with respect to the reference model.

Similar to AIC_c, the Bayesian Information Criterion (BIC) (Burnham & Anderson 2004a,b; Liddle 2007) is also a very popular criterion used for model selection, and it is based on the Bayesian probability principles. The BIC takes into account the goodness of fit and penalizes the complexity of models i.e. higher numbers of free parameters, stronger than the AIC_c does. For a set of considered models, the BIC is defined as :

$$\mathrm{BIC} = \chi^2_{min} + \mathcal{K}_{\rm f}\ln(\mathcal{N}_{\rm t}), \tag{53}$$

where χ^2_{min}, $\mathcal{K}_{\rm f}$, and $\mathcal{N}_{\rm t}$ were previously defined in the AIC_c criterion. Similar to AIC, the difference, ΔBIC, between models and the reference one is more significant than the BIC value itself,

$$\Delta\mathrm{BIC} = \mathrm{BIC}_{model} - \mathrm{BIC}_{reference}. \tag{54}$$

The model is said to be weakly disfavored by the data if $0 < \Delta\mathrm{BIC} \leq 2$, it is said to be disfavored if $2 < \mathrm{BIC} \leq 6$ while it is said to be strongly disfavored if $\Delta\mathrm{BIC} > 6$.

3.2 Data description

In the following, we briefly describe the observational datasets we are going to use to test the two cosmological models introduced within the framework of the scalar-tensor representation of $f(R,T)$ gravity. To obtain the tightest possible constraints on the cosmological parameters, probes from independent sources must be used. Given that, we estimate the free parameters of the scalar-tensor $f(R,T)$ cosmological models by using the $H(z)$ measurements, together with the Pantheon+ data.

3.2.1 The $H(z)$ Dataset

In general, $H(z)$ measurements are obtained from the measurements of BAO in the radial direction of the galaxy clustering (Gaztanaga et al. 2009). They can also be extracted from the differential age method, using the following relation

$$H(z) = -\frac{1}{1+z}\frac{dz}{dt}.$$

In the current study, we use a dataset consisting of 57 points of $H(z)$, shown in Table 3. The dataset consists of 31 data points derived from the differential ages of the galaxies, and from 26 points obtained from BAO. The latter are in general correlated, and one must take into account these correlations by using the corresponding covariance matrix (Moresco et al. 2020). In the literature, the 57 points are assumed most of the time to be uncorrelated, which is not a totally

unacceptable approach, if the correlations between points are negligible (Koussour et al. 2022; Luciano & Liu 2023; Luciano 2023). Therefore, the chi-square corresponding to the $H(z)$ measurements is given by

$$\chi^2_{H(z)} = \sum_{i=1}^{57} \frac{[H_{th}(z_i) - H_{obs}(z_i)]^2}{\sigma^2_{H(z_i)}}, \tag{55}$$

where $H_{\rm th}$, H_{obs} and $\sigma_{H(z_i)}$ stand for model predictions, measured values and standard deviation, respectively.

3.2.2 The Pantheon+ dataset

The observations of the type Ia supernova (SNIa) have played a major role in the discovery of the comic accelerating expansion. So far, SNIa have proved to be one of the most effective tools to study the nature of the component causing the accelerating evolution of the Universe. Several SNIa compilations have been released in recent years, like the Union (Kowalski et al. 2008), Union2 (Amanullah et al. 2010) and Union2.1 (Suzuki et al. 2012), Joint Light-cure Analysis (JLA) (Betoule et al. 2014), Pantheon (Scolnic et al. 2018), and the Pantheon+ (Scolnic et al. 2022) compilations, respectively. The latter has been recently released, and it contains 1701 SNIa data, covering the redshift range $0.001 < z < 2.3$. Moreover, SNIa are incredibly luminous astrophysical objects, and they are assumed to be a perfect standard candles, serving to measure relative distances via the distance modulus.

The chi-square values of the SNIa dataset are given by

$$\chi^2_{Pantheon+} = \vec{D}^T \cdot \mathbf{C}^{-1}_{Pantheon+} \cdot \vec{D}, \tag{56}$$

where $\mathbf{C}_{Pantheon+}$ is the covariance matrix provided with the Pantheon+ data, including both statistical and systematic uncertainties. and

$$\vec{D} = m_{Bi} - M - \mu_{\rm model}, \tag{57}$$

where m_{Bi} and M denote the apparent, and the absolute magnitudes of SNIa, respectively. Furthermore, $\mu_{\rm model}$ denotes the corresponding distance modulus as predicted by an assumed cosmological model, and defined as

$$\mu_{\rm model}(z_i) = 5\log_{10}\frac{D_L(z_i)}{(H_0/c)Mpc} + 25, \tag{58}$$

where H_0 is the current value of the Hubble rate, and D_L is the luminosity distance, expressed, for a flat homogeneous and isotropic FLRW Universe, as follows

$$D_L(z) = (1+z)H_0\int_0^z \frac{dz'}{H(z')}. \tag{59}$$

Unlike the Pantheon sample, the degeneracy between the absolute magnitude M and H_0 is broken in the Pantheon+ dataset, due to the rewriting of the vector $\vec{D}$, appearing in Eq. (56), in terms of the distance moduli of SNIa in the Cepheid hosts, which allows to constrain M independently via the relations

$$\vec{D}'_i = \begin{cases} m_{Bi} - M - \mu_i^{\rm Ceph} & i \in \text{ Cepheid hosts} \\ m_{Bi} - M - \mu_{\rm model}(z_i) & \text{otherwise}, \end{cases} \tag{60}$$

where $\mu_i^{\rm Ceph}$ stands for the distance modulus corresponding to the Cepheid host of the $i^{\rm th}$ SNIa, measured independently with Cepheid calibrators. Therefore, Eq. (56) can be rewritten as

$$\chi^2_{SN} = \vec{D'}^T \cdot \mathbf{C}^{-1}_{\rm Pantheon+} \cdot \vec{D'} \tag{61}$$

The total chi-square function is therefore given by

$$\chi^2_{Tot} = \chi^2_{H(z)} + \chi^2_{Pantheon+}. \tag{62}$$

3.3 Statistical results

In this Section, we present the findings resulting from our analysis of the two cosmological models in the scalar-tensor representation of $f(R,T)$ gravity. We put under test the two models, Model 1 and 2, described by the differential equation systems Eqs. (33)-(35) and Eqs. (41)-(43), respectively. Since these equations describing the two cosmological models do not have an exact analytical solution, numerical solutions are required. After defining the initial conditions, by using the relations and the constraints on the initial conditions and model parameters as given by Eqs. (37), (38), (47) and (48), respectively, and computing the numerical solutions for both models, we have constrained the free parameters by using the Hubble measurements, and the Pantheon+ data, respectively. The priors for the cosmological parameters (Hubble function, distance modulus, matter density) have been assumed based on the previous results obtained in the framework of the ΛCDM model.

The Table 1 presents the MCMC results regarding the two cosmological models under consideration of the scalar-tensor representation of $f(R,T)$ gravity. The Table 1 shows the prior of the free parameters used in running the MCMC analysis, their best fit, and their mean values. The MCMC results are then used the test the statistical significance of each model with respect to the ΛCDM model. To this aim, we use the AIC_c criterion, Eq. (52), to classify these models according to their ability of describing the cosmological data.

According to Table 2, the first model of $f(R,T)$ gravity is the most favoured by observations, as it corresponds to the lowest value of $AICc$. As a result, it must be considered as the reference model, i.e. $\Delta AIC_c = 0$. However, in the current study, we still keep ΛCDM to be the reference model, as it is almost always the case in the present-day literature. Therefore, negative deviations from zero, $\Delta AIC_c < 0$, stand for models more favored over ΛCDM, while positive deviations, $\Delta AICc > 0$, stand for models less favored than ΛCDM.

In Table 2, we summarize the results of the statistical analysis, revealing what might be the capital result of the whole paper, i.e., *the fact that Model 1 is statistically better* than ΛCDM, as $\Delta AIC_{c,\rm Model\ 1} = -5.20541$. In other words, observations favor Model 1 over the Standard Model of cosmology. On the other hand, Model 2 is less favored than ΛCDM, but it still has considerable observational support as $\Delta AIC_{c,\rm Model\ 2} = 1.97916$.

However, according to the BIC criterion ΛCDM is statistically better, mainly due its simplicity as it is described with less parameters compared to $f(R,T)$ gravity models. In addition, the BIC analysis reports less evidence against the $f(R,T)$ Model 1 than $f(R,T)$ Model 2. In other words, $f(R,T)$ Model 1 is more favored by data and consistent with ΛCDM than the $f(R,T)$ Model 2.

The 1D and 2D posterior distributions, at 68.3% (1σ) and 95.4% (2σ) confidence levels, are depicted in Fig. 1 and Fig. 2 for Model 1 and Model 2, respectively. The first model presents a positive correlation between (ξ, h) and (σ, h). A positive correlation is also seen between the (β_0, h) and (ρ_0, h) parameters of the second model, *which might be an important signature to address the* H_0 *tension problem within the context of* $f(R,T)$ *gravity*.

Additionally, as one can see from Fig. 3, both cosmological models in scalar-tensor $f(R,T)$ gravity fit well the Hubble measurement, as

Model	Parameter	Prior	Best fit	Mean
ΛCDM	$\Omega_{\rm m}$	$[0.001, 1]$	$0.279816^{+0.0108}_{-0.0108}$	$0.279816^{+0.0108}_{-0.0108}$
	h_0	$[0.4, 1]$	$0.698281^{+0.0065}_{-0.0065}$	$0.698487^{+0.0065}_{-0.0065}$
	M	$[-21, -17]$	$-19.3637^{+0.0178}_{-0.0178}$	$-19.3632^{+0.0178}_{-0.0178}$
$f(R,T)$ Model 1	ξ	$[0, 3]$	$1.42545^{+0.0351535}_{-0.0351535}$	$1.430^{+0.035}_{-0.035}$
	σ	$[-3, 3]$	$-0.633484^{+0.1427}_{-0.1427}$	$-0.80^{+0.16}_{-0.13}$
	ϵ	$[-2, 2]$	$0.775255^{+0.2494}_{-0.2494}$	$0.44^{+0.24}_{-0.28}$
	h_0	$[0.4, 1]$	$0.691558^{+0.0068}_{-0.0068}$	$0.6910^{+0.0069}_{-0.0069}$
	M	$[-21, -17]$	$-19.3666^{+0.0173}_{-0.0173}$	$-19.369^{+0.018}_{-0.018}$
$f(R,T)$ Model 2	α_0	$[0, 3000]$	$1323.78^{+66.4980}_{-66.4980}$	1306.91^{+74}_{-45}
	β_0	$[0, 100]$	$77.7069^{+0.8321}_{-0.8321}$	$77.74^{+0.88}_{-0.88}$
	φ_0	$[0, 6]$	$3.1181^{+0.1697}_{-0.1697}$	$3.11913^{+0.14}_{-0.17}$
	h_0	$[0.4, 1]$	$0.693981^{+0.0060}_{-0.0060}$	$0.6947^{+0.0065}_{-0.0065}$
	M	$[-21, -17]$	$-19.3775^{+0.0167}_{-0.0167}$	$-19.375^{+0.018}_{-0.018}$

Table 1. Summary of the best fit and mean values of the free cosmological parameters of the $f(R,T)$ cosmological models.

much as does ΛCDM, particularly in the recent past, for low redshift values. In Fig. 4, the relative differences between both models and ΛCDM are depicted. The Figure clearly shows that Model 1 behaves exactly like the ΛCDM model up to $z = 1$, while Model 2 does so only up to $z = 0.5$.

3.4 Cosmographic analysis

In view of the existence of a considerable number of alternatives theoretical models describing the current cosmic acceleration, an approach relying on the scale factor derivatives, dubbed cosmography (Visser 2004, 2005; Cattoën & Visser 2008; Visser & Cattoen 2010), has proved its extreme efficiency in comparing models to each other and discriminating between the various dark energy fluid types proposed from a theoretical perspective. In other words, the cosmographic analysis offers a useful way to compare a large number of cosmological models in the space of the scale factor derivatives.

In the framework under consideration, in the following we consider the deceleration, the jerk and the snap parameters, defined respectively as

$$q(z) = -\frac{1}{a}\frac{d^2a}{dt^2}\left[\frac{1}{a}\frac{da}{dt}\right]^{-2}, \tag{63}$$

$$j(z) = \frac{1}{a}\frac{d^3a}{dt^3}\left[\frac{1}{a}\frac{da}{dt}\right]^{-3}, \tag{64}$$

$$s(z) = \frac{1}{a}\frac{d^4a}{dt^4}\left[\frac{1}{a}\frac{da}{dt}\right]^{-4}. \tag{65}$$

Deceleration parameter. In Fig. 5, we plot the evolution of the deceleration parameter with respect to the redshift z. The current values of the deceleration parameter are -0.43, -0.57 and -0.61 for Model 1, ΛCDM and Model 2, respectively. The two $f(R,T)$ models behave qualitatively as ΛCDM, but they predict a deceleration twice as higher as the one predicted by ΛCDM. Furthermore, the transition from a decelerating to an accelerating expansion takes place first in Model 1, then in Model 2, and lastly in the ΛCDM model.

Jerk parameter. In Fig. 6, the evolution of the second cosmographic parameter, $j(z)$, with respect to the redshift is shown. The two analysed $f(R,T)$ models show a similar (different) behaviour of the jerk parameter for redshifts $z > 4$ (for $z < 4$). Furthermore, compared to ΛCDM, both models start with lower values for $z < 1$, before crossing the ΛCDM curve, i.e., $j(z) = 1$, and evolving toward the constant asymptotic value ~ 2.5.

Snap parameter. In Fig. 7, the evolution of the snap parameter in terms of the redshift is depicted. Both $f(R,T)$ models show a very similar behaviour as the snap takes a negative value in the past, and reaches in the present the value $s = 0$, corresponding to ΛCDM.

Statefinder diagnostic $\{j, s_f\}$**.** The Statefinder is a purely geometrical diagnostic method, allowing to characterize dark energy properties in a model independent way (Sahni et al. 2003; Alam et al. 2003; Sahni et al. 2008). In addition, the Statefinder is related to the EoS parameter and to its first time derivative, and therefore it is useful to differentiate between different dark energy types. Furthermore, the Statefinder diagnostic is also constructed with the help of the scale factor and its time derivatives.

The jerk parameter, j, is defined in Eq. (64), and the parameter s_f is given by the linear combination of the jerk and the deceleration parameters as

$$s_f(z) = \frac{j(z) - 1}{3\,(q(z) - 1/2)}. \tag{66}$$

To figure out the properties of the effective DE induced by the scalar-tensor representations of the cosmological models in $f(R,T)$ gravity, as well as to shed light on the dissimilarities between the models under study, and ΛCDM, we plot the Statefinder pair, $\{j, s_f\}$, in Fig. 8, and the parametric plot of the jerk and the deceleration parameters, $\{q, j\}$, in Fig. 9. Both figures show a significant difference between the cosmological models in the scalar-tensor representations of $f(R,T)$ gravity, and ΛCDM.

Fig. 8 shows that both models start with values in the range $j > 1$ and $s_f < 0$ in the past, indicating the Chaplygin gas type dark energy model (Gorini et al. 2003), and evolve to take values between $j < 1$ and $s_f > 0$, indicating the quintessence domain (Zlatev et al. 1999; Brax & Martin 2000) by crossing the intermediate ΛCDM fixed point $\{j = 1, s_f = 0\}$ and then return to the Chaplygin gas domain by passing the interim fixed ΛCDM point at late times.

In Fig. 9, ΛCDM separates the Chaplygin gas like (top-half) and the quintessence like (bottom half) regions. Both $f(R,T)$ models evolve from the Chaplygin gas like to the quintessence like region. Unlike Model 1, Model 2 crosses the point $\{j = 1, q = 0.5\}$,

Model	$\chi^{2\ min}_{\rm tot}$	$\chi^2_{\rm red}$	$\mathcal{K}_{\rm f}$	AIC_c	ΔAIC_c	BIC	ΔBIC
ΛCDM	1656.4528	0.943848	3	1662.4664	0	1678.87	0
$f(R,T)$ Model 1	1647.2268	0.9396	5	1657.2610	−5.20541	1684.59	5.71786
$f(R,T)$ Model 2	1654.4114	0.94376	5	1664.4456	1.97916	1691.77	12.9025

Table 2. Summary of $\chi^{2\ min}_{\rm tot}$, $\chi^2_{\rm red}$, AIC_c, ΔAIC_c, BIC and ΔBIC for ΛCDM and $f(R,T)$ models.

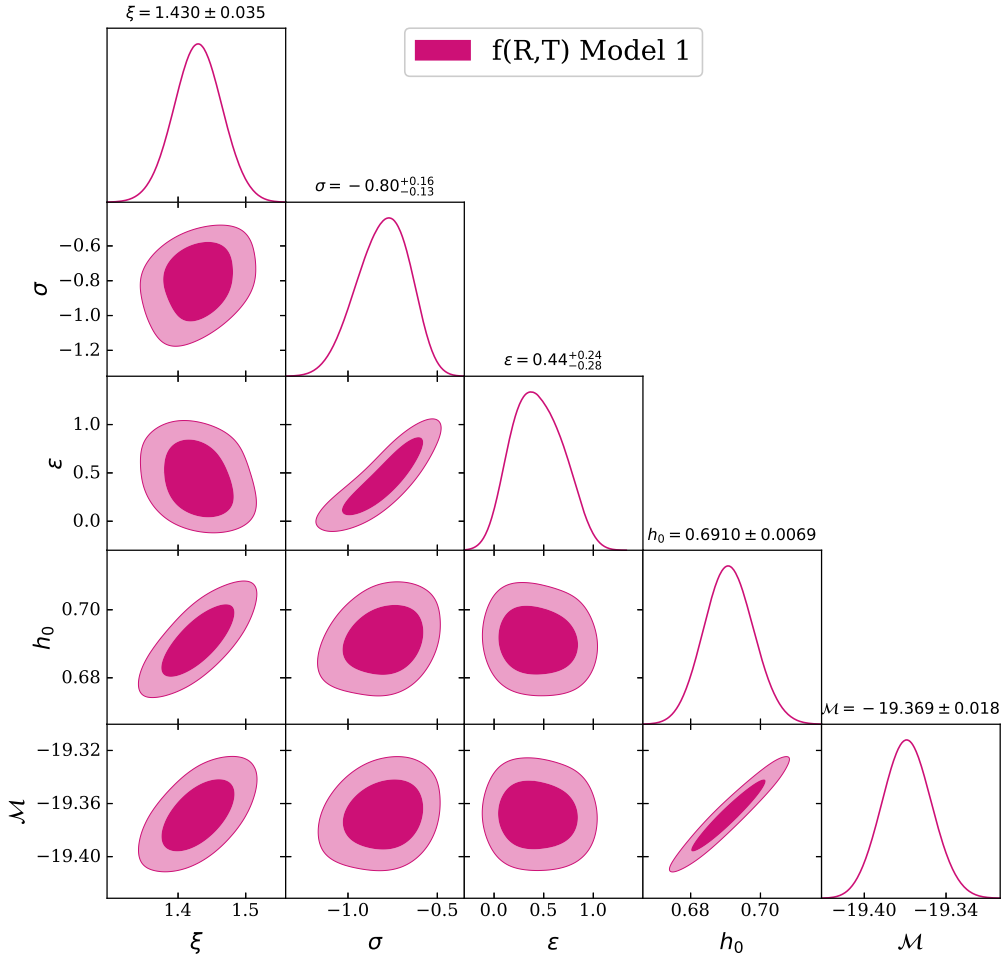


Figure 1. MCMC confidence contours at 1σ and 2σ, obtained after constraining the $f(R,T)$ cosmological Model 1 with the SNIa+H(z) data.

corresponding to the Standard Cold Dark Matter cosmology (SCDM). Moreover, the two models keep evolving in the Chaplygin gas like region, before re-entering again in the quintessence like region, crossing the steady state - the de Sitter (dS) solution, i.e., $\{j = 1, q = -1\}$.

3.5 Cosmological quantities

Matter densities. Now we proceed to a comparative study of the relevant physical parameters of the cosmological models of the $f(R,T)$ gravity and ΛCDM model. We consider first the evolution of the matter density parameter $r(z)$, presented in Fig. 10. The Figure shows that the ordinary matter density is a monotonically increasing function of the redshift in all models, starting with an initial value[1] $r(0) = 1$ for $f(R,T)$ gravity models and $r(0) \approx 0.28$ for ΛCDM.

Nevertheless, at higher redshifts, important differences begin to appear between models, as the first model of $f(R,T)$ gravity (i.e. Model 1) predicts a higher value of matter density than Model 2, and ΛCDM. The huge amount of matter density predicted by Model 1 in the past (i.e., at higher redshifts) can be explained by an intensive production of matter through the gravitational mechanism. The same behavior is also shown in Fig. 11 and Fig. 13.

On the other hand, a small offset resulting from the chosen initial conditions (i.e. $r(0) = 1$) is separating matter densities of the $f(R,T)$ gravity cosmological models and ΛCDM. Nonetheless, at

[1] Other choices of initial conditions could be adopted to compute the numerical solution.

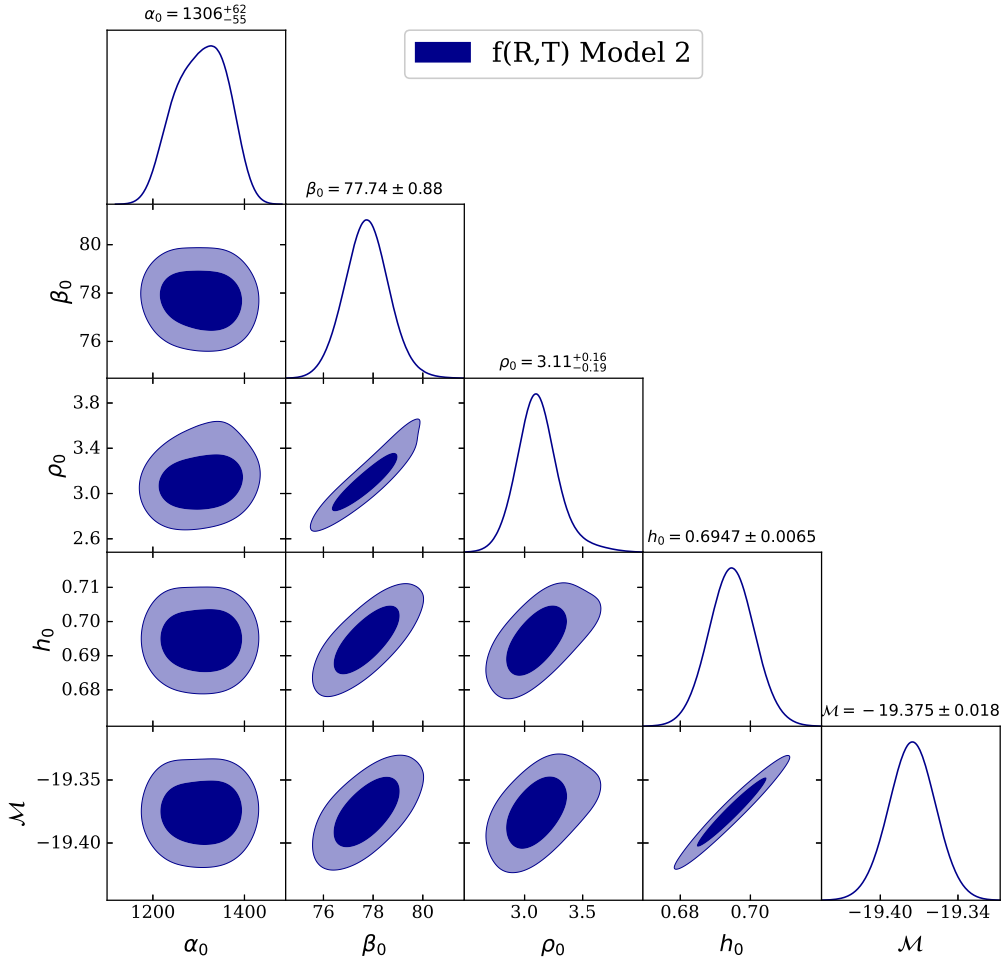


Figure 2. MCMC confidence contours at 1σ and 2σ, obtained after constraining the $f(R,T)$ cosmological Model 2 with SNIa+H(z) data.

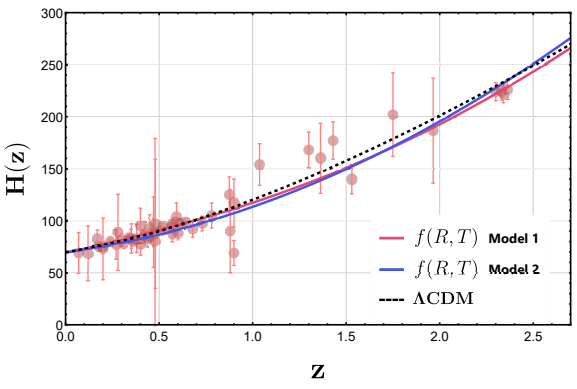


Figure 3. The evolution of the Hubble function H(z) of the $f(R,T)$ gravity and of the ΛCDM cosmological models as a function of the redshift z against the Hubble data measurements.

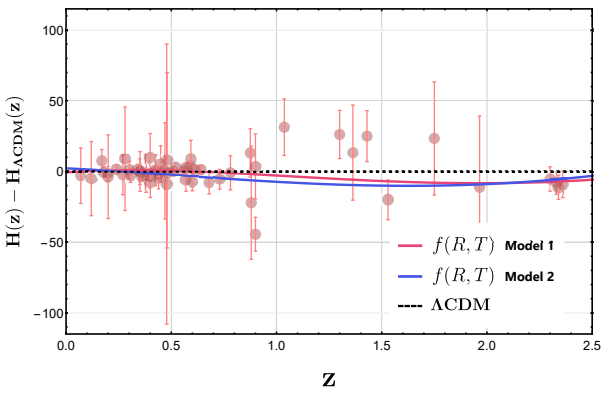


Figure 4. The variation of the difference between the $f(R,T)$ cosmological models, and the ΛCDM model, as a function of the redshift z against the Hubble measurements.

$z = 1.4$, this offset increases in Model 1, and disappears in Model 2, leading, beyond that redshift, to a similar evolution of the matter density in both Model 2 and ΛCDM.

Creation pressures and particle creation rates. In the thermodynamic interpretation of $f(R,T)$ gravity, the non-conservation of the matter energy-momentum tensor is related to matter creation processes, which are described by two effective thermodynamic quantities, the creation pressure, and the particle creation rate, respectively, and which are essentially related to matter generation processes. The creation pressure for Model 1 is presented in Fig. 11. The creation pressure is negative in the redshift range $z \in (0, 10)$, indicating a continuous presence of matter creation . As for Model 2, the creation pressure is depicted in Fig. 12. In the late stages of the

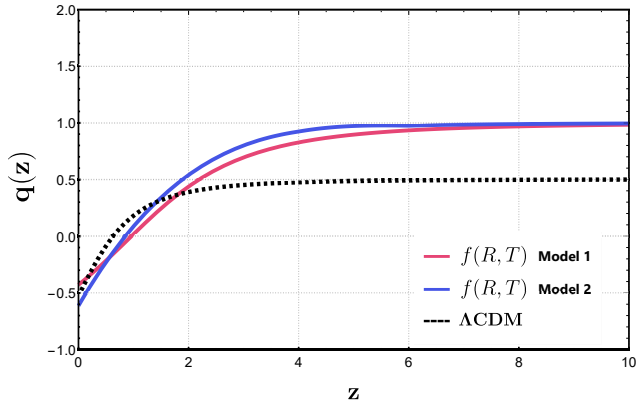


Figure 5. Evolution of the deceleration parameter as a function of the redshift z in the $f(R,T)$ and in ΛCDM cosmologies.

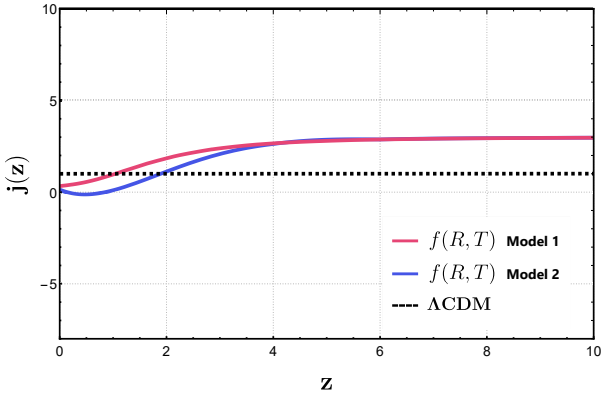


Figure 6. Evolution of the jerk parameter as a function of the redshift z in the $f(R,T)$ and in ΛCDM cosmological models.

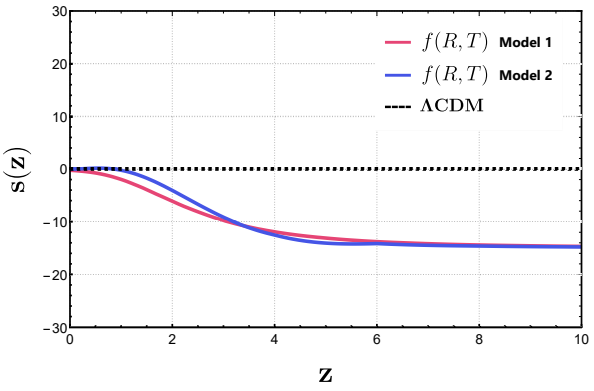


Figure 7. The variation of the snap parameter s as a function of the redshift z in $f(R,T)$ and in ΛCDM models.

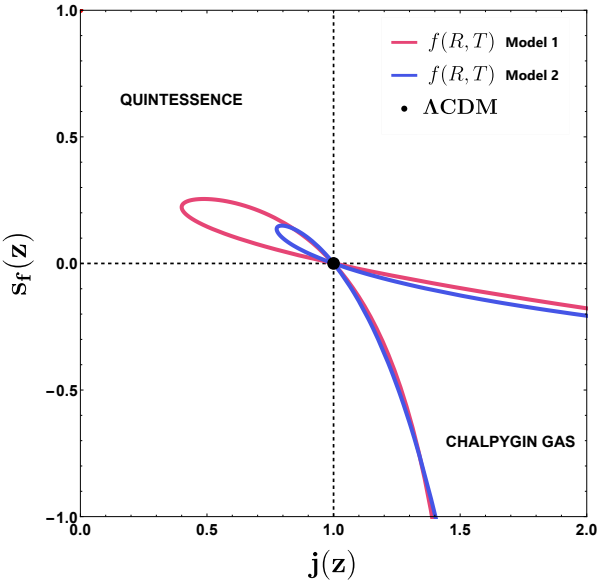


Figure 8. The variation of the snap parameter s_f as a function of the jerk parameter j in $f(R,T)$ and in ΛCDM models.

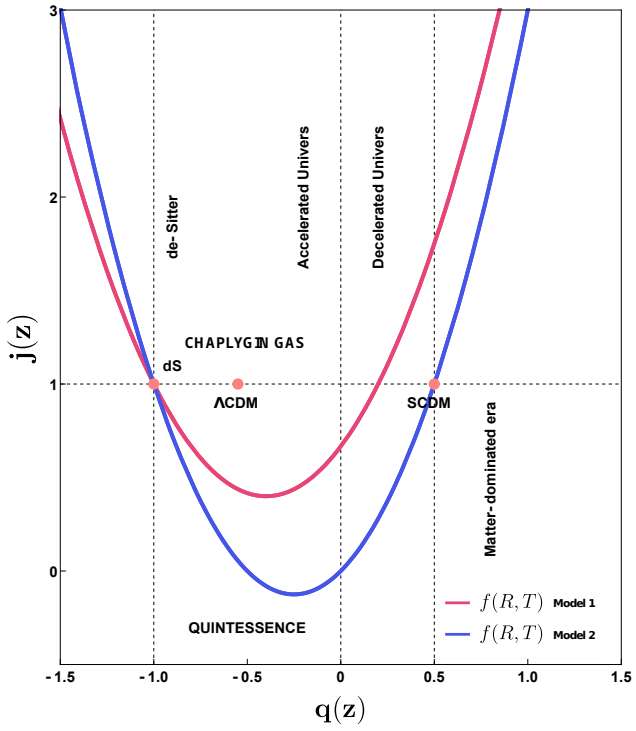


Figure 9. The variation of the jerk parameter j as a function of the deceleration parameter q in $f(R,T)$ and ΛCDM models.

cosmological evolution, in the redshift range $z \in (0,2)$, the creation pressure is roughly a constant, and it can mimic a cosmological constant. For $z > 2$ the creation pressure is positive, indicating the presence of a matter annihilation process, and increases rapidly with the redshift (decreases in time). Still, even a very small constant creation pressure can trigger the accelerated expansion of the Universe. The particle creation rates for the two $f(R,T)$ cosmological models are presented in Fig. 13. The behavior of Γ closely follows the behavior of the creation pressure. For Model 1, Γ increases rapidly with the redshift, decreasing in time. Hence, the accelerating expansion of the Universe is triggered by a significant increase in the matter production rate, which reaches its maximum at the present time. On the other hand, the dynamics of the Universe for Model 2 is controlled by a very small (close to zero) particle creation/annihilation rate. For $z > 4$, the particle creation rate becomes negative, indicating the beginning of a particle annihilation process. Furthermore, for redshifts higher than 4, $z > 4$, due to particle annihilation, Model 2 violates the second law of thermodynamics.

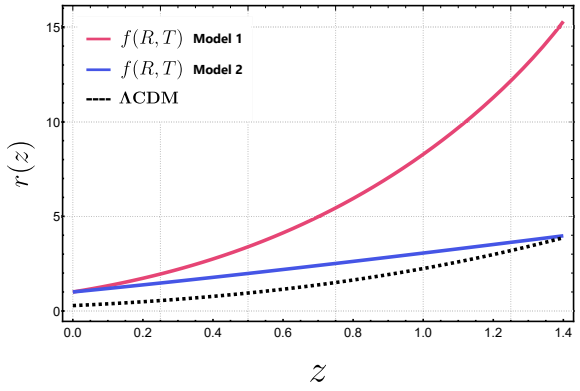


Figure 10. The dynamical evolution of the energy density of matter r, as a function of the redshift z, for both cosmological models of the scalar-tensor representation of $f(R,T)$ gravity, and for the ΛCDM model.

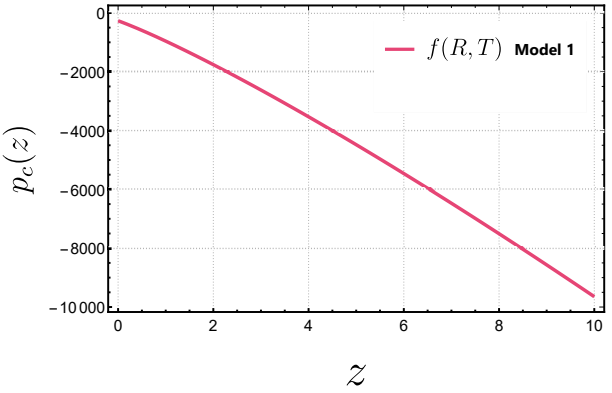


Figure 11. The evolution of the creation pressure p_c as a function of redshift z for Model 1. The initial conditions used for the numerical integration of the generalized Friedmann equations are $h(0) = 1$, $r(0) = 1$ and $\varphi(0) = -0.005$.

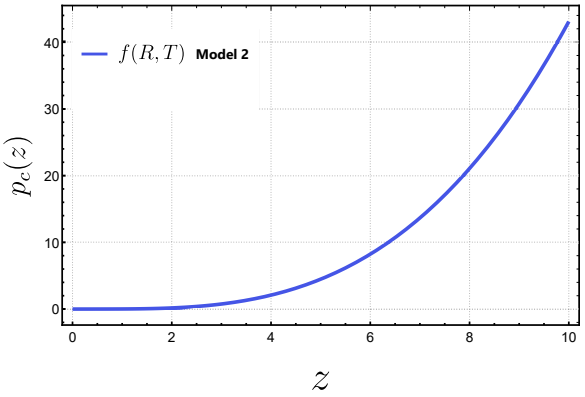


Figure 12. The evolution of the creation pressure p_c as a function of redshift z for Model 2. The initial conditions used for the numerical integration of the generalized Friedmann equations are $h(0) = 1$, $r(0) = 1$, $n = 1$, $m = 2$ and $\psi(0) = \left(m\alpha_0\varphi_0^n\right)^{\frac{1}{1+m}}$.

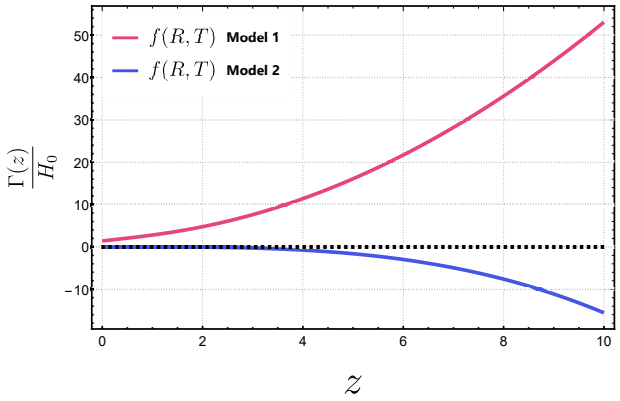


Figure 13. The evolution of the matter creation rate Γ/H_0 as a function of redshift z. The initial conditions adopted for the numerical solution are $h(0) = 1$, $r(0) = 1$ and $\varphi(0) = -0.005$.

$Om(z)$ ***diagnostic.*** The $Om(z)$ diagnostic tool is a very important theoretical method for distinguishing alternative cosmological models from ΛCDM. To be more precise, $Om(z)$ diagnostic is used to determine the nature of the cosmological fluid under study, i.e., either it is a phantom fluid (Dahmani et al. 2023; Mhamdi et al. 2023; Bouali et al. 2021, 2019), a quintessence one, or a cosmological constant like.

The $Om(z)$ parameter is defined as (Sahni et al. 2008)

$$Om(z) = \frac{(H(z)/H_0)^2 - 1}{(1+z)^3 - 1}. \tag{67}$$

In the case of the ΛCDM model, the function $Om(z)$ remains constant, and equal to the current matter density parameter Ω_{m0}. Furthermore, for cosmological models with a constant EoS, i.e., ω = constant, a positive slope of $Om(z)$ refers to a phantom behavior, whereas a negative slope refers to a quintessence-like behavior. In the case of the standard ΛCDM cosmology, and in the presence of a dynamical dark energy fluid, described by a linear barotropic EoS, with the parameter of the equation of state denoted by $\omega(z)$, the first Friedmann equation is written as

$$\left(\frac{H(z)}{H0}\right)^2 = \Omega_{m0}(1+z)^3 + (1-\Omega_{m0}) \times \mathrm{Exp}\left[3\int_0^z \frac{1+\omega(z')}{1+z'}dz'\right]. \tag{68}$$

Thus, for a constant ω, we obtain

$$\left(\frac{H(z)}{H0}\right)^2 = \Omega_{m0}(1+z)^3 + (1-\Omega_{m0})(1+z)^{3(1+\omega)}. \tag{69}$$

Then, it easily follows that,

$$Om(z) = \frac{\Omega_{m0}(1+z)^3 + (1-\Omega_{m0})(1+z)^{3(1+\omega)} - 1}{(1+z)^3 - 1}. \tag{70}$$

For the ΛCDM model, the EoS parameter is $\omega = -1$, and as a result

$$Om(z) = \Omega_{m0}. \tag{71}$$

The evolutions of the $Om(z)$ function with respect to the redshift in both $f(R,T)$ cosmological models, and in the ΛCDM cosmology, are presented in Fig. 14.

The $Om(z)$ functions for the three considered cosmological models are different. For ΛCDM, $Om(z)$ is an absolute constant, while

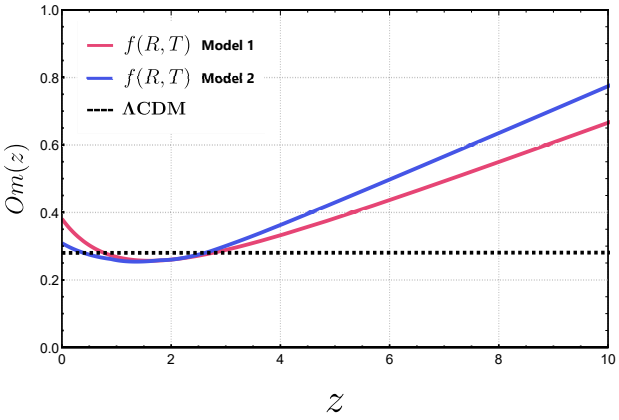


Figure 14. The evolution of the $Om(z)$ function in the two $f(R,T)$ cosmological models, and in the ΛCDM model.

in the range $z < 2$, for the $f(R,T)$ cosmological models, $Om(z)$ is a decreasing function, i.e., it has a negative slop, which indicates a quintessence like behavior. On the other hand, for $z > 2$, the $Om(z)$ is an increasing function for the $f(R,T)$ models, i.e., with a positive slope, indicating the presence of a phantom like behavior.

4 CONCLUSIONS AND FINAL REMARKS

In the present paper, as a first step in our study, we have reviewed the scalar-tensor representation of $f(R,T)$ gravity, which can be reformulated in terms of two independent scalar functions φ and ψ, and in the presence of a self-interaction potential $U(\varphi,\psi)$ (Pinto et al. 2022). Similarly to its metric version, the energy-momentum tensor of the matter is not conserved in this formalism. From a physical point of view, we interpret the non-conservation of the energy-momentum tensor as implying the presence of particle production/annihilation processes, which can be elegantly describe by means of the physical and mathematical approach of the irreversible thermodynamics of open systems. The cosmological evolution and the generalized Friedmann equations can also be obtained in a straightforward manner in the scalar-tensor representation, and the cosmic behavior is essentially determined by the (generally unknown) functional form of $U(\varphi,\psi)$. We have also assumed that the non-conservation of the matter energy-momentum tensor, which also appears in the scalar-tensor formulation, implies a particle creation/annihilation, which can be described through the formalism of the irreversible thermodynamics of open systems.

To study the viability of such a mathematical setup, we have tested observationally the implications of scalar-tensor $f(R,T)$ gravity on cosmological scales. We have considered two specific cosmological models, already introduced, and investigated in (Pinto et al. 2022), denoted by $f(R,T)$ Model 1, and $f(R,T)$ Model 2, respectively. These models correspond to two distinct choices of the self-interaction potential $U(\varphi,\psi)$ of the theory. In the first case we have adopted a simple additive quadratic structure for U, while the second case corresponds to a multiplicative-additive algebraic structure.

The system of the generalized Friedmann cosmological equations considered in the present investigation of the scalar-tensor formulation of $f(R,T)$ gravity is a strongly nonlinear system of differential equations. Hence, the behavior of its solutions essentially depends on the initial conditions used for φ and ψ for the numerical integration of the system. Small changes of the initial conditions may lead to large variations in the cosmological behavior, which could not be compatible with the observational data. However, the initial conditions for the two scalar fields are not independent, but they are determined by the initial conditions of the cosmological parameters, or, more exactly, by the initial value of the Hubble function and of the deceleration parameter. Hence, the range of variation of the initial conditions of the scalar fields is restricted by the adopted cosmological model, and they must satisfy a set of constraints for each system of differential equations describing the cosmic expansion. These conditions also allow us to significantly restrict the priors of the model parameters during the numerical fitting.

To constrain the aforementioned models observationally, we have used a set of 57 data points of Hubble measurements, together with the latest release of Pantheon compilation, i.e., Pantheon+. After solving numerically the generalized Friedmann equations for the two models, the best and the mean values of the model parameters were extracted by the help of the Markov Chain Monte Carlo analysis. A comparative study between the $f(R,T)$ gravity cosmological models, and the ΛCDM model has been performed by means of the corrected Aikaike Information Criterion. The results of this comparison are shown in Table 2. The latter Table clearly shows that the preference order of models, according to the observations is Model 1, ΛCDM, and then Model 2. Thus, in other words, it turns out that Model 1 fits the observational data better than the ΛCDM model, indicating that $f(R,T)$ gravity holds a greater potential to explain the present-day dynamics of the Universe than standard GR.

Using the best fit parameters, we have also studied the various cosmological and physical quantities relevant for the description of the Universe. The behavior of the deceleration parameter shows that in the past, the Universe, as described by the $f(R,T)$ cosmological models, evolves with a larger deceleration parameter than the one predicted by the ΛCDM model. The acceleration of the Universe begins at a much higher redshift, of the order of $z \approx 4$, as compared to the redshift $z \approx 2$, at which the ΛCDM Universe begins to accelerate. The differences between the previously noted behaviors can be explained by the presence of the effective creation/annihilation processes of $f(R,T)$ gravity. Nevertheless, the two models enter in the accelerated state of expansion, with $q < 0$, at almost the same redshift, $z \approx 0.5$. This dynamical evolution can be explained by considering the variation of the particle creation rate. When the rate of creation/annihilation becomes insignificant (i.e. at late times), the Universe enters an accelerated phase of expansion a bit earlier in both Model 1 and 2 than in ΛCDM. In other words, the accelerated expansion of the Universe can be fully explained by a particle production mechanism without assuming the existence of any exotic form of dark energy. Furthermore, the statefinder and $Om(z)$ diagnostics show that the effective fluids characterising the studied $f(R,T)$ cosmological models in the present are quintessence like, as indicated in Fig. 8 and Fig. 14, respectively.

An important and interesting question is the physical nature of the matter particles that are created due to the energy transfer from geometry to matter. We may conjecture that these particles contribute to, or fully represent, the dark matter content of the present-day Universe. The physical nature of the dark matter particle is not yet known, despite the intensive experimental and observational effort dedicated to finding them. One theoretically attractive possibility is that dark matter may consist of ultra-light bosonic particles, with masses in the range of $m \in \left(10^{-24} - 10^{-22}\right)$ eV (Hui et al. 2017). From a physical point of view, such ultra-light dark matter particles created from the gravitational field could represent a pseudo Nambu-Goldstone boson. Other types of ultra-light dark matter particle that could be generated by the gravitational field are the axions, hypothetical particles whose

existence is suggested by elementary particle physics, having masses of the order of $m \leq 10^{-22}$ eV (Park et al. 2012). It is important to note that very small mass particles, like axions or Nambu-Goldstone bosons, can be generated even by weak gravitational fields. Therefore, the creation of significant amounts of very low mass dark matter particles in the early Universe, from strong gravitational fields, cannot be excluded, and this process may be active in the present-day Universe, thus triggering its accelerated evolution.

Another physical possibility for the nature of the dark matter particle creation from geometry can be related to the Scalar Field Dark Matter (SFDM) models (Martinez-Medina et al. 2015). In these particular types of dark matter models, dark matter is considered as a real scalar field that couples minimally to gravity. The mass of the scalar field particle is assumed to have a value of the order of $m < 10^{-21}$ eV, of the same order of magnitude as the mass of the axions or Nambu-Goldstone bosons. Such particles can also be generated in the framework of the scalar-tensor representation of the $f(R,T)$ gravity, due to the natural presence of scalar fields in the theoretical formalism. The scalar field dark matter particles, as well as the other bosonic type dark matter particles, have the important property that at very low temperatures they can form a Bose-Einstein Condensate, a special physical state in which all particles are in the same quantum ground state (Boehmer & Harko 2007; Harko 2014a; Harko & Lobo 2015). From a physical point of view there is a close relationship between the Scalar Field Dark Matter and the Bose-Einstein Condensate dark matter models. In the open irreversible thermodynamic model of particle creation discussed in the present paper, the creation of the dark matter particles can also take place in the form of dynamical scalar fields, originating from the fields φ and ψ. Moreover, in the present approach we assume that the two scalar fields φ and ψ do not fully decay into dark or ordinary matter. Therefore, dark matter in the form of a scalar field could be a direct result of the particle creation dynamics during the early phase in the evolution of the Universe.

The present day cosmological observations, as well as the present knowledge of the elementary particle physics cannot constrain accurately the basic predictions of the present phenomenological particle creation model. However, even without a detailed and precise knowledge of the physical processes determining the cosmological evolution, the scalar-tensor model of the modified $f(R,T)$ gravity theory allows us to make some predictions about the dynamical aspects of the Universe. In the two models we have considered the presence of the scalar fields is fully responsible for the matter content of the Universe, and for its large scale evolution.

As a conclusion and according to AIC_c and BIC, it can be said that at least one of the $f(R,T)$ gravity cosmological models is statistically and observationally more supported than the ΛCDM model. Furthermore, the presence of matter creation in these models, described by the formalism of irreversible thermodynamics of open systems, can fully explain the dynamical evolution of the Universe, and, perhaps, the formation of cosmic structure in the early Universe, provided that the particle creation rate, or, equivalently, the creation pressure, is not zero. Unlike GR, modified theories of gravity that admit non-minimal geometry-matter couplings can provide a phenomenological description of particle production, either in a matter dominated or in an empty Universe. Thus, the results presented in this work can open some new windows for the understanding of the intricate properties of the cosmic dynamics, and of the role of the matter in an evolving Universe.

ACKNOWLEDGEMENTS

We would like to thank the anonymous referee for comments and suggestions that helped us to significantly improve our manuscript. The work of TH is supported by a grant of the Romanian Ministry of Education and Research, CNCS-UEFISCDI, project number PN-III-P4-ID-PCE-2020-2255 (PNCDI III). FSNL acknowledges support from the Fundação para a Ciência e a Tecnologia (FCT) Scientific Employment Stimulus contract with reference CEECINST/00032/2018, and funding from the research grant CERN/FIS-PAR/0037/2019. FSNL and MASP acknowledge support from the Fundação para a Ciência e a Tecnologia (FCT) research grants UIDB/04434/2020 and UIDP/04434/2020, and through the FCT project with reference PTDC/FIS-AST/0054/2021 ("BEYond LAmbda"). MASP also acknowledges support from the Fundação para a Ciência e a Tecnologia (FCT) through the Fellowship UI/BD/154479/2022.

DATA AVAILABILITY

This is entirely theoretical work, and all of the results presented in the manuscript are derived from the equations. We did not generate any original data during the course of this study, nor did we analyse any third-party data in this article.

REFERENCES

Abbott B. P., et al., 2016, Physical Review Letters, 116, 061102
Akaike H., 1974, IEEE transactions on automatic control, 19, 716
Alam S., [BOSS] e., 2017, Mon. Not. Roy. Astron. Soc., 470, 2617
Alam U., Sahni V., Deep Saini T., Starobinsky A., 2003, Monthly Notices of the Royal Astronomical Society, 344, 1057
Alvarenga F., De La Cruz-Dombriz A., Houndjo M., Rodrigues M., Sáez-Gómez D., 2013, Physical review D, 87, 103526
Amanullah R., et al., 2010, The Astrophysical Journal, 716, 712
Arevalo F., Cid A., Moya J., 2017, The European Physical Journal C, 77, 1
Avelino P., et al., 2016, Symmetry, 8, 70
Aviles A., Bravetti A., Capozziello S., Luongo O., 2013, Physical Review D, 87, 044012
Bautista J. E., et al., 2017, Astronomy & Astrophysics, 603, A12
Bertolami O., Boehmer C. G., Harko T., Lobo F. S., 2007, Physical Review D, 75, 104016
Bertolami O., Lobo F. S., Páramos J., 2008, Physical Review D, 78, 064036
Betoule M., et al., 2014, Astronomy & Astrophysics, 568, A22
Beylin V., Khlopov M., Kuksa V., Volchanskiy N., 2020, Universe, 6, 196
Bian L., Liu X., Xie K.-P., 2021, Journal of High Energy Physics, 2021, 1
Blake C., et al., 2012, Mon. Not. Roy. Astron. Soc., 425, 405
Boehmer C. G., Harko T., 2007, Journal of Cosmology and Astroparticle Physics, 2007, 025
Borriello A., Salucci P., 2001, Monthly Notices of the Royal Astronomical Society, 323, 285
Bouali A., Albarran I., Bouhmadi-López M., Ouali T., 2019, Physics of the Dark Universe, 26, 100391
Bouali A., Albarran I., Bouhmadi-López M., Errahmani A., Ouali T., 2021, Physics of the Dark Universe, 34, 100907
Bouali A., Chaudhary H., Hama R., Harko T., Sabau S. V., Martín M. S., 2023, The European Physical Journal C, 83, 121
Bouhmadi-Lopez M., Capozziello S., Cardone V. F., 2010, Physical Review D, 82, 103526
Brax P., Martin J., 2000, Physical Review D, 61, 103502
Burnham K. P., Anderson D. R., 2004a, A practical information-theoretic approach, 2
Burnham K. P., Anderson D. R., 2004b, Sociological methods & research, 33, 261
Capozziello S., 2002, International Journal of Modern Physics D, 11, 483

Capozziello S., Cardone V., Salzano V., 2008, Physical Review D, 78, 063504
Capozziello S., Cardone V., Farajollahi H., Ravanpak A., 2011, Physical Review D, 84, 043527
Capozziello S., D'Agostino R., Luongo O., 2019, International Journal of Modern Physics D, 28, 1930016
Capozziello S., D'Agostino R., Luongo O., 2020, Acta Phys Polon Supp, 13, 271
Carroll S. M., 2001, Living reviews in relativity, 4, 1
Cattoën C., Visser M., 2008, Physical Review D, 78, 063501
Cebrián S., 2023, in Journal of Physics: Conference Series. p. 012004
Chimento L., Forte M. I., 2008, Phys. Lett. B, 666, 205
Chuang C.-H., Wang Y., 2013, Mon. Not. Roy. Astron. Soc., 435, 255
Chuang C.-H., et al., 2013, Monthly Notices of the Royal Astronomical Society, 433, 3559
Clifton T., Ferreira P. G., Padilla A., Skordis C., 2012, Physics reports, 513, 1
Dahmani S., Bouali A., El Bojaddaini I., Errahmani A., Ouali T., 2023, General Relativity and Gravitation, 55, 22
De Felice A., Tsujikawa S., 2010, Living Reviews in Relativity, 13, 1
Delubac T., et al., 2015, Astronomy & Astrophysics, 574, A59
Di Valentino E., et al., 2021, Classical and Quantum Gravity, 38, 153001
Einstein A., 1915, Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften, pp 844–847
Einstein A., 1986, Cosmological Constants, p. 16
Faraoni V., Belknap-Keet S. D., 2017, Phys. Rev. D, 96, 044040
Farias I., Moraes P., 2021, arXiv preprint arXiv:2108.09332
Font-Ribera A., et al., 2014, Journal of Cosmology and Astroparticle Physics, 2014, 027
Gaztanaga E., Cabre A., Hui L., 2009, Mon. Not. Roy. Astron. Soc., 399, 1663
Ghilencea D., 2021, The European Physical Journal C, 81, 510
Ghilencea D., 2022, The European Physical Journal C, 82, 23
Ghilencea D., 2023, The European Physical Journal C, 83, 176
Gonçalves T. B., Rosa J. L., Lobo F. S. N., 2022a, The European Physical Journal C, 82, 418
Gonçalves T. B., Rosa J. L., Lobo F. S., 2022b, Physical Review D, 105, 064019
Gorini V., Kamenshchik A., Moschella U., 2003, Physical Review D, 67, 063509
Haghani Z., Harko T., Lobo F. S., Sepangi H. R., Shahidi S., 2013, Physical Review D, 88, 044023
Haghani Z., Harko T., Shahidi S., 2018, Physics of the Dark Universe, 21, 27
Harko T., 2014a, Phys. Rev. D, 89, 084040
Harko T., 2014b, Physical Review D, 90, 044067
Harko T., Lobo F. S., 2010, The European Physical Journal C, 70, 373
Harko T., Lobo F. S., 2014, Galaxies, 2, 410
Harko T., Lobo F. S. N., 2015, Phys. Rev. D, 92, 043011
Harko T., Lobo F. S., 2020, International Journal of Modern Physics D, 29, 2030008
Harko T., Lobo F. S., Nojiri S., Odintsov S. D., 2011, Physical Review D, 84, 024020
Harko T., Lobo F. S., Minazzoli O., 2013, Physical Review D, 87, 047501
Harko T., Lobo F. S., Otalora G., Saridakis E. N., 2014a, Physical Review D, 89, 124036
Harko T., Lobo F. S., Otalora G., Saridakis E. N., 2014b, Journal of Cosmology and Astroparticle Physics, 2014, 021
Harko T., Lobo F. S., Mimoso J. P., Pavón D., 2015, The European Physical Journal C, 75, 1
Harko T., Koivisto T. S., Lobo F. S., Olmo G. J., Rubiera-Garcia D., 2018, Physical Review D, 98, 084043
Harko T., Lobo F. S., Saridakis E. N., 2021, Universe, 7, 227
Hawking S. W., Ellis G. F., 2023, The large scale structure of space-time. Cambridge university press
Hawking S. W., Penrose R., 1970, Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences, 314, 529
Houndjo M., 2012, International Journal of Modern Physics D, 21, 1250003
Hui L., Ostriker J. P., Tremaine S., Witten E., 2017, Phys. Rev. D, 95, 043541
Koivisto T., 2006, Classical and Quantum Gravity, 23, 4289
Koussour M., Shekh S., Bennai M., 2022, International Journal of Modern Physics A, 37, 2250184
Kowalski M., et al., 2008, The Astrophysical Journal, 686, 749
Lebedev O., 2021, Progress in Particle and Nuclear Physics, 120, 103881
Lee S., 2021, arXiv preprint arXiv:2108.06043
Lee S., 2023, arXiv preprint arXiv:2301.06947
Liddle A. R., 2007, Monthly Notices of the Royal Astronomical Society: Letters, 377, L74
Lima J., Trevisani S., Santos R., 2021, Physics Letters B, 820, 136575
Luciano G. G., 2023, Physics of the Dark Universe, 41, 101237
Luciano G. G., Liu Y., 2023, Symmetry, 15, 1129
Martin J., 2012, Comptes Rendus Physique, 13, 566
Martinez-Medina L. A., Robles V. H., Matos T., 2015, Phys. Rev. D, 91, 023519
Mhamdi D., Bargach F., Dahmani S., Bouali A., Ouali T., 2023, General Relativity and Gravitation, 55, 11
Moresco M., et al., 2012, Journal of Cosmology and Astroparticle Physics, 2012, 006
Moresco M., et al., 2016, Journal of Cosmology and Astroparticle Physics, 2016, 014
Moresco M., Jimenez R., Verde L., Cimatti A., Pozzetti L., 2020, The Astrophysical Journal, 898, 82
Nesseris S., Pantazis G., Perivolaropoulos L., 2017, Phys. Rev. D, 96, 023542
Odintsov S. D., Sáez-Gómez D., 2013, Physics Letters B, 725, 437
Oka A., Saito S., Nishimichi T., Taruya A., Yamamoto K., 2014, Mon. Not. Roy. Astron. Soc., 439, 2515
Overduin J. M., Wesson P. S., 2004, Physics Reports, 402, 267
Park C.-G., Hwang J.-c., Noh H., 2012, Phys. Rev. D, 86, 083535
Perlmutter S., et al., 1999, The Astrophysical Journal, 517, 565
Persic M., Salucci P., Stel F., 1996, Monthly Notices of the Royal Astronomical Society, 281, 27
Piedipalumbo E., Moglie E. D., Cianci R., 2015, International Journal of Modern Physics D, 24, 1550100
Pinto M. A. S., Harko T., Lobo F. S. N., 2022, Physical Review D, 106, 044043
Pires N., Santos J., Alcaniz J., 2010, Physical Review D, 82, 067302
Pizza L., 2015, Physical Review D, 91, 124048
Ratsimbazafy A. L., Loubser S. I., Crawford S. M., Cress C. M., Bassett B. A., Nichol R. C., Väisänen P., 2017, Mon. Not. Roy. Astron. Soc., 467, 3239
Rezaei M., Malekjani M., 2021, The European Physical Journal Plus, 136, 219
Riess A. G., et al., 1998, The Astronomical Journal, 116, 1009
Rosa J. a. L., 2021, Phys. Rev. D, 103, 104069
Rosa J. L., Kull P. M., 2022, The European Physical Journal C, 82, 1154
Rosa J. L., Marques M. A., Bazeia D., Lobo F. S., 2021, The European Physical Journal C, 81, 1
Sahni V., Saini T. D., Starobinsky A. A., Alam U., 2003, Journal of Experimental and Theoretical Physics Letters, 77, 201
Sahni V., Shafieloo A., Starobinsky A. A., 2008, Physical Review D, 78, 103502
Salucci P., Martins C. F., Lapi A., 2011, arXiv preprint arXiv:1102.1184
Schwarzschild K., 1916, Sitzungsber.Preuss.Akad.Wiss.Berlin (Math.Phys.), 1916, 189
Scolnic D. M., et al., 2018, Astrophys. J., 859, 101
Scolnic D., et al., 2022, The Astrophysical Journal, 938, 113
Shabani H., Farhoudi M., 2014, Physical Review D, 90, 044031
Shomer A., 2007, arXiv preprint arXiv:0709.3555
Sotiriou T. P., Faraoni V., 2010, Reviews of Modern Physics, 82, 451
Stern D., Jimenez R., Verde L., Kamionkowski M., Stanford S. A., 2010, Journal of Cosmology and Astroparticle Physics, 2010, 008
Suzuki N., et al., 2012, The Astrophysical Journal, 746, 85
Tan M., Biswas R., 2012, Monthly Notices of the Royal Astronomical Society, 419, 3292
Tauscher K., Rapetti D., Burns J. O., 2018, Journal of Cosmology and Astroparticle Physics, 2018, 015
Teppa Pannia F. A., Perez Bergliaffa S. E., Manske N., 2019, The European Physical Journal C, 79, 1
Visser M., 2004, Classical and Quantum Gravity, 21, 2603

Visser M., 2005, General Relativity and Gravitation, 37, 1541
Visser M., Cattoen C., 2010, in , Dark Matter In Astrophysics And Particle Physics. World Scientific, pp 287–300
Vrieze S. I., 2012, Psychological methods, 17, 228
Wang Y., et al., 2017, Monthly Notices of the Royal Astronomical Society, 469, 3762
Weinberg S., 1989, Reviews of modern physics, 61, 1
Yousaf Z., Bamba K., Bhatti M. Z.-u.-H., 2016, Physical Review D, 93, 124048
Zhang C., Zhang H., Yuan S., Liu S., Zhang T.-J., Sun Y.-C., 2014, Research in Astronomy and Astrophysics, 14, 1221
Zlatev I., Wang L., Steinhardt P. J., 1999, Physical Review Letters, 82, 896

APPENDIX

In this Appendix, we present the Table of a total of 57 data points. Among them, 31 points are derived using the differential age technique, while the remaining 26 points are obtained through the use of BAO and other methods, all within the redshift range of $0.07 \leq z \leq 2.42$. The following Table represented in this Section describes 57 data points, which provide values of the Hubble parameter, $H(z)$, along with their corresponding errors, σ_H. Additionally, each data point is accompanied by references for further information.

This paper has been typeset from a TEX/LATEX file prepared by the author.

Observational measurements of Hubble parameter

z	$H(z)$	σ_H	Ref.	z	$H(z)$	σ_H	Ref.
0.070	69	19.6	(Stern et al. 2010)	1.363	160	33.6	(Lee 2023)
0.120	68.6	26.2	(Stern et al. 2010)	1.965	186.5	50.4	(Lee 2023)
0.480	97	62	(Stern et al. 2010)	0.24	79.69	2.99	(Gaztanaga et al. 2009)
0.880	90	40	(Stern et al. 2010)	0.34	83.8	3.66	(Gaztanaga et al. 2009)
0.170	83	8	(Chimento & Forte 2008)	0.43	86.45	3.97	(Gaztanaga et al. 2009)
0.270	77	14	(Chimento & Forte 2008)	0.30	81.7	6.22	(Oka et al. 2014)
0.400	95	17	(Chimento & Forte 2008)	0.31	78.18	4.74	(Wang et al. 2017)
0.900	117	23	(Chimento & Forte 2008)	0.36	79.94	3.38	(Wang et al. 2017)
0.90	69	12	(Chimento & Forte 2008)	0.40	82.04	2.03	(Wang et al. 2017)
1.300	168	17	(Chimento & Forte 2008)	0.44	84.81	1.83	(Wang et al. 2017)
1.430	177	18	(Chimento & Forte 2008)	0.48	87.79	2.03	(Wang et al. 2017)
1.530	140	14	(Chimento & Forte 2008)	0.52	94.35	2.64	(Wang et al. 2017)
1.750	202	40	(Chimento & Forte 2008)	0.56	93.34	2.3	(Wang et al. 2017)
0.1791	75	4	(Moresco et al. 2012)	0.59	98.48	3.18	(Wang et al. 2017)
0.1993	75	5	(Moresco et al. 2012)	0.64	98.82	2.98	(Wang et al. 2017)
0.3519	83	14	(Moresco et al. 2012)	0.35	82.7	9.1	(Chuang & Wang 2013)
0.593	104	13	(Moresco et al. 2012)	0.38	81.5	1.9	(Alam & [BOSS] 2017)
0.6797	92	8	(Moresco et al. 2012)	0.51	90.4	1.9	(Alam & [BOSS] 2017)
0.7812	105	12	(Moresco et al. 2012)	0.61	97.3	2.1	(Alam & [BOSS] 2017)
0.8754	125	17	(Moresco et al. 2012)	0.44	82.6	7.8	(Blake et al. 2012)
1.037	154	20	(Moresco et al. 2012)	0.60	87.9	6.1	(Blake et al. 2012)
0.200	72.9	29.6	(Zhang et al. 2014)	0.73	97.3	7.0	(Blake et al. 2012)
0.280	88.8	36.6	(Zhang et al. 2014)	0.57	87.6	7.8	(Chuang et al. 2013)
0.3802	83	13.5	(Moresco et al. 2016)	0.57	96.8	3.4	(Lee 2021)
0.4004	77	10.2	(Moresco et al. 2016)	2.30	224	8.6	(Delubac et al. 2015)
0.4247	87.1	11.2	(Moresco et al. 2016)	2.33	224	8	(Bautista et al. 2017)
0.4497	92.8	12.9	(Moresco et al. 2016)	2.34	222	8.5	(Delubac et al. 2015)
0.4783	80	99	(Moresco et al. 2016)	2.36	226	9.3	(Font-Ribera et al. 2014)
0.470	89	34	(Ratsimbazafy et al. 2017)				

Table 3. The table shows the 57 data points of the Hubble parameter considered by the current paper.

MAUSAM, **74**, 4 (October 2023), 1015-1028

MAUSAM

DOI : https://doi.org/10.54302/mausam.v74i4.5903
Homepage: https://mausamjournal.imd.gov.in/index.php/MAUSAM

UDC No. 502.2 : 551.588 (540.23)

Poor air quality as an important predictor of climate change in Delhi

GAURAV YADAV, GEETA SINGH and S. D. ATTRI*

Department of Environmental Engineering, Delhi Technological University, New Delhi – 110 042, India

**India Meteorological Department, MoES, New Delhi – 110 003, India*

(*Received 23 May 2022, Accepted 24 January 2023*)

e mail : geeta.singh@dce.ac.in

सार —**जलवायवी परिस्थितियों में निरंतर परिवर्तन ने दुनिया भर के लोगों के लिए एक चुनौतीपूर्ण स्थिति पैदा कर दी है। बढ़ते तापमान और आर्द्रता का अधिक आबादी वाले शहरों पर सबसे बुरा प्रभाव पड़ता है, और शहरी वातावरण में खराब वायु गुणवत्ता जलवायवी परिवर्तन को मुख्य रूप से प्रभावित करती है। दुनिया भर के सबसे अधिक प्रदूषित शहरों में वायु प्रदूषण हॉटस्पॉट की सूची में दिल्ली सबसे ऊपर है, जिसे इस अध्ययन के लिए चुना गया है। इस अध्ययन में मानदंड वायु प्रदूषकों और मौसम संबंधी मापदंडों के बीच संबंध का आकलन किया गया। यह परिकल्पना की गई थी कि मानदंड वायु प्रदूषक दिल्ली में दैनिक डेटासेट** (01 **जनवरी**, 2015 - 31 **दिसंबर**, 2021) **और औसत वार्षिक डेटासेट** (2000 **से** 2021) **के दौरान तापमान और सापेक्ष आर्द्रता** (**जलवायु परिवर्तन के स्तंभ**) **में परिवर्तन का सकारात्मक पूर्वानुमान करेंगे। यह अध्ययन** EVIEWS 12 **का उपयोग करके मॉडल अन्वेषण और गुणांक अनुमान में इलास्टिक नेट-एप्लाइड नियमितीकरण का उपयोग करता है। यह देखा गया कि चयनित अध्ययन अवधि के दौरान, अधिकांश मानदंड वायु प्रदूषकों ने दिल्ली की जलवायवी परिस्थितियों में परिवर्तन को बढ़ाने में महत्वपूर्ण भूमिका निभाई। यह शोध उपलब्ध साहित्य की मदद से वायु प्रदूषण और जलवायु परिवर्तन के बीच अंतर संबंध को और स्पष्ट करता है।**

ABSTRACT. The continuous change in climatic conditions has created a challenging situation for people living all over the World. The rising temperature and humidity have the worst impacts on the cities with high populations, and poor air quality in an urban environment significantly affects climatic variables. Delhi, which tops the list of air pollution hotspots among the top polluted cities around the World, is selected for this study. This study assessed a correlation between criteria air pollutants and meteorological parameters. It was hypothesized that criteria air pollutants would positively predict the change in temperature and relative humidity (pillars of climate change) during the daily dataset (January 01, 2015 - December 31, 2021) and average annual dataset (2000 to 2021) in Delhi. This study uses elastic net-applied regularization in model exploration and coefficient estimation using EVIEWS 12. It was observed that during the selected study period, most of the criteria air pollutants played an important part in increasing the changes in the climatic conditions of Delhi. This research further explains the interlinkage between air pollution and climate change with the help of available literature.

Key words – Air Pollutants, Climate Change, Temperature, Relative humidity, Elastic Net Regression.

1. Introduction

Due to population increase, growing numbers of automobiles, fuel use, inadequate transportation networks, poor land use patterns, industrialization, and insufficient environmental legislation, India has been particularly vulnerable to air pollution and climate change over the previous two decades (Ravindra *et al*., 2016; Sharma & Mauzerall, 2022; Yarlagadda *et al*., 2022). New Delhi, India's capital, is one of the worst-affected cities. It has the country's highest level of particulate matter pollution and very abnormal meteorological conditions (Biswas *et al*., 2011; Krishan *et al*., 2019; Marlier *et al*., 2016; Sharma *et al*., 2018). Before the outbreak of SARS-COV-2 infection in 2020, the average annual $PM_{2.5}$concentration in India was 58.1 $\mu g/m^3$ in 2019, with Delhi's average $PM_{2.5}$ concentration being 98.6 $\mu g/m^3$, which lowered down to 93.23 in 2020 (due to lockdown) and which again increased to 103.33 in 2021. Similarly, the average annual relative humidity in 2019 was 48.69%, which increased to 53.25% in 2021 (Krishan *et al*., 2019). All of these above stats indicate that the situations are getting worse yearly. The weather has a significant impact on air quality in Delhi, *i.e*., meteorological variables such as temperature,

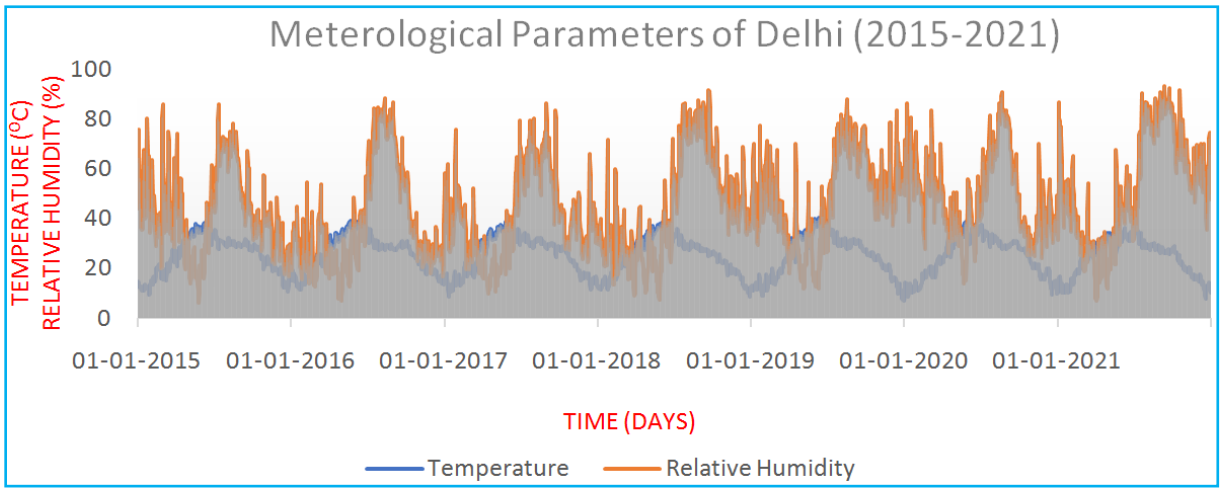


Fig. 1. Daily meteorological conditions in Delhi

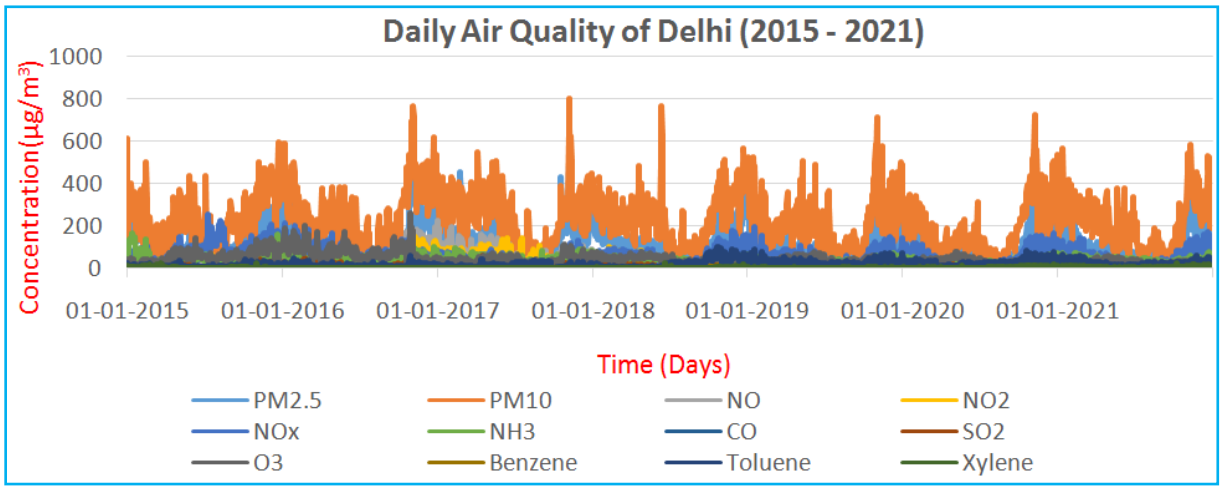


Fig. 2. Daily Air quality in Delhi

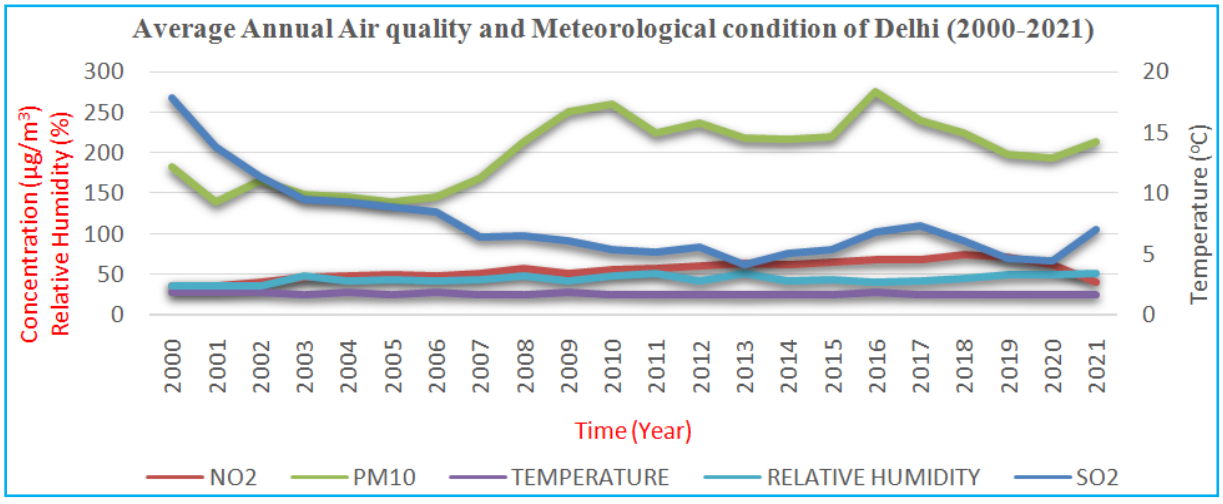


Fig. 3. Annual Concentration of Ecological Conditions in Delhi

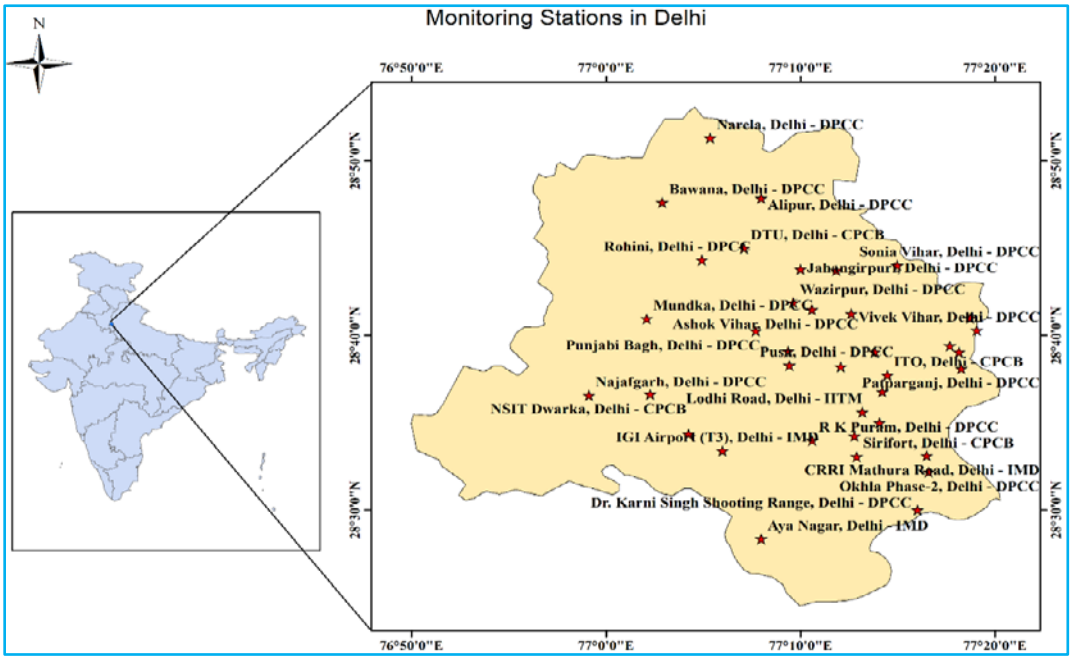


Fig. 4. Study area with 38 CAAQMS stations

humidity, wind characteristics and vertical mixing can influence pollutant emission, transport, dispersion, chemical change and deposition (Das *et al*., 2021; Deswal & Chandna, 2010; Guttikunda, 2010; Guttikunda & Gurjar, 2012; Krishan *et al*., 2019; Ramsey *et al*., 2014; Sonwani *et al*., 2021; Tiwari *et al*., 2014; Trivedi *et al*., 2014; Verma & Desai, 2008). Further, if the weather is significantly associated with air pollutants, the fluctuating concentrations of pollutants will also show some impact on the change in weather and climate (Saha *et al*., 2014). Fig. 1 represents the daily change in the temperature and relative humidity of Delhi during (2015- 2021); Fig. 2 represents the concentration of air pollutants on a daily 24-hour mean basis and, Fig. 3 shows the perfect example of a change in criteria air pollutants concentration and meteorological conditions in Delhi over past 22 years on an annual basis. All these figures showed that the change in the temperature and relative humidity trend is very similar to the transformation of some air pollutants. This indicates that both meteorological variables and air pollutants influence each other and share a positive association. Temperature and Relative humidity are the pillars of climate change which further is projected to exacerbate air pollution in numerous densely populated areas by altering atmospheric ventilation and dilution, precipitation and other removal mechanisms and atmospheric chemistry (Depietri *et al*., 2012; Godde *et al*., 2021). The bad air quality will have a direct influence on human health and the change of climate. Climate change affects air quality and hence impacts human health.

So far, researchers have considered only meteorological parameters to conclude the association between meteorological parameters and air pollutants. (Barzeghar *et al*., 2022; Dandotiya *et al*., 2019; Deswal & Chandna, 2010; Ding *et al*., 2021; Ghosh *et al*., 2017; Guo *et al*., 2022; Gupta *et al*., 2004; Guttikunda & Gurjar, 2012; Heydarzadeh *et al*., 2022; Ilten & Selici, 2008; Jayamurugan *et al*., 2013; Khedairia & Khadir, 2012; Kucbel *et al*., 2017; Kuzu & Saral, 2017; Li *et al*., 2017; Mahanta *et al*., 2021; Mahapatra *et al*., 2014; Ramsey *et al*., 2014; Sati & Mohan, 2018; Trivedi *et al*., 2014; Verma & Desai, 2008; Xue & Liu, 2014; Yadav *et al*., 2015). At the same time, most of the studies were done at the bivariate level, which has the drawback of not considering the presence or effect of other independent variables between the two being investigated. This study examined the impact of criteria air pollutants on climate change in the Delhi region. It was hypothesized that all the criteria air pollutants (PM_{10}, $PM_{2.5}$, NO, NO_2, NO_X, NH_3, O_3, CO, SO_2, C_6H_6, C_7H_8, C_8H_{10}) would positively predict the change in temperature and relative humidity. Elastic net regression analysis was used to test this hypothesis because it is a model with minimal crucial independent variables. Additionally, it punishes the fitting of unwanted independent variables. However, none of the studies which used elastic-net has been published for finding an association between air pollutants and meteorological conditions in Delhi. Many statisticians proved that elastic-net produces accurate results with reference to the lucidity and precision of regression models (Edouard Grave,

2011). Elastic-net regression provides the most uncompromising model with perfect prediction when predicting a dependent variable with numerous predictors (Kim *et al*., 2016). Through cross-validation, it was also discovered that a model founded by elastic-net regression predicts accurate results out of the bounds of data required for regression analysis. (McNeish, 2015).

2. Materials and methods

2.1. *Study area*

The study analyzedthe eleven most polluted districts in the Delhi (28.7041° N, 77.1025° E) region, as shown in Fig. 4. These districts have emerged as important hubs for the commercial, industrial, medical and educational sectors, attracting people from all over the country. These districts are struggling to handle air pollution and its impacts due to the increase in population. In all these districts, the effect of change in meteorological conditions over a significant period has been highly observed.

2.2. *Data collection*

(*i*) For model 1, daily 24-hour mean concentration data of criteria air pollutants, temperature and relative humidity (of 38 CAAQMS sites) was collected from (https://app.cpcbccr.com/ccr) operated by CPCB (https://cpcb.nic.in) as shown in Fig. 4. The daily dataset was collected between (January 01, 2015 - December 31, 2021) by using the above-mentioned sites.

(*ii*) For model 2, the dataset for the past 22 years was collected from (http://www.cpcbenvis.nic.in/air_quality _data.html), where annual average data was stored under NAMP (National Air Quality Monitoring Programme). For this study,annualaverage data for the year 2000 to 2021 was collected from seven monitoring sites. (*i.e*., ITO, Janakpuri, Nizamuddin, Pitampura, Shahdara, Shahazada Bagh, and Siri Fort).

2.3. *Data analysis*

The hypothesis tests were conducted to find whether the criteria air pollutants carry an essential influence on the change in Temperature and relative humidity in the eleven most polluted districts lying under the Delhi region.

(*i*) For Model 1, the daily 24-hour mean Temperature and relative humidity in all eleven districts of Delhi were used as the dependent variables, and 24-hour mean air quality was taken as predictors to find how much variance each parameter was created on the change in the climate.

(*ii*) For Model 2, the annual average temperature and relative humidity were used as dependent variables and the yearly average concentration of (SO_2, NO_2 and PM_{10}) was taken as predictors to find how much variance each parameter was created on the change in the climate.

(*iii*) This study's statistical analyses were performed using EVIEWS 12 (www.eviews.com).

2.3.1. *Elastic Net Regression*

Elastic net-applied regularization has been used in model exploration and coefficient estimation (Zou & Hastie, 2005) as shown in equation (1), where α is the mixing parameter between ridge ($\alpha = 0$) and lasso ($\alpha = 1$).

$$L_{\text{enet}}\left(\hat{\beta}\right) = \frac{\sum_{i-1}^{n}\left(y_i - x_i'\hat{\beta}\right)^2}{2n} + \lambda\left(\frac{1-\alpha}{2}\sum_{j=1}^{m}\hat{\beta}_j^2 + \alpha\sum_{j=1}^{m}\left|\hat{\beta}_j\right|\right) \quad (1)$$

By minimizing regression coefficients to zero, the L1-norm (lasso) of penalty creates a sparse model. The L2-norm (ridge) of penalty removes the constraint on the number of selected variables, stimulates grouping, and stabilizes the L1 regularization route (Kuang *et al*., 2015). Elastic net reduces the regression coefficients by combining the L1-norm (lasso) and L2-norm (ridge) penalties; $\alpha = 0.5$ was utilized, which is a midpoint among L1 and L2. Additionally, to reduce the threat of overfitting, elastic net regression was done with K-Fold cross-validation to measure the mean square error. The complete dataset was randomly split into ten folds. For every test, one block of split datasets (90% of the entire dataset) was used to calculate coefficients.

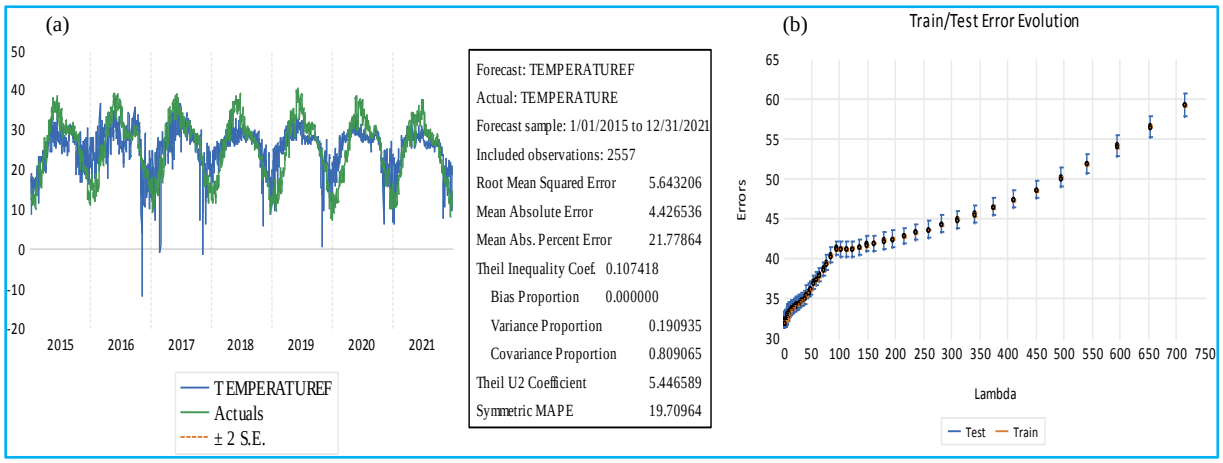
Furthermore, the predictive performance of our prediction model was calculated with the remaining data block (10% of the entire dataset). For each test, a separate set of coefficients quantity was calculated while changing λ. Lambda (λ) at minimum error is used to calculate the simplest model with high prediction accuracy and regularized beta coefficients. To estimate accurate algorithm OLS with 500 maximum iterations and 0.0001 convergence was used.

3. Results and discussion

3.1. *Analysis output of daily data model (2015-2021), including all twelve air pollutants*

3.1.1. *Association among criteria air pollutants and temperature*

The stepwise procedure of elastic net regression was performed to find an association between predictors and



Figs. 5(a&b). Temperature model output after testing {graph (a), X – axis denotes Time (Years) and (b) Y - axis denotes temperature (°C)}

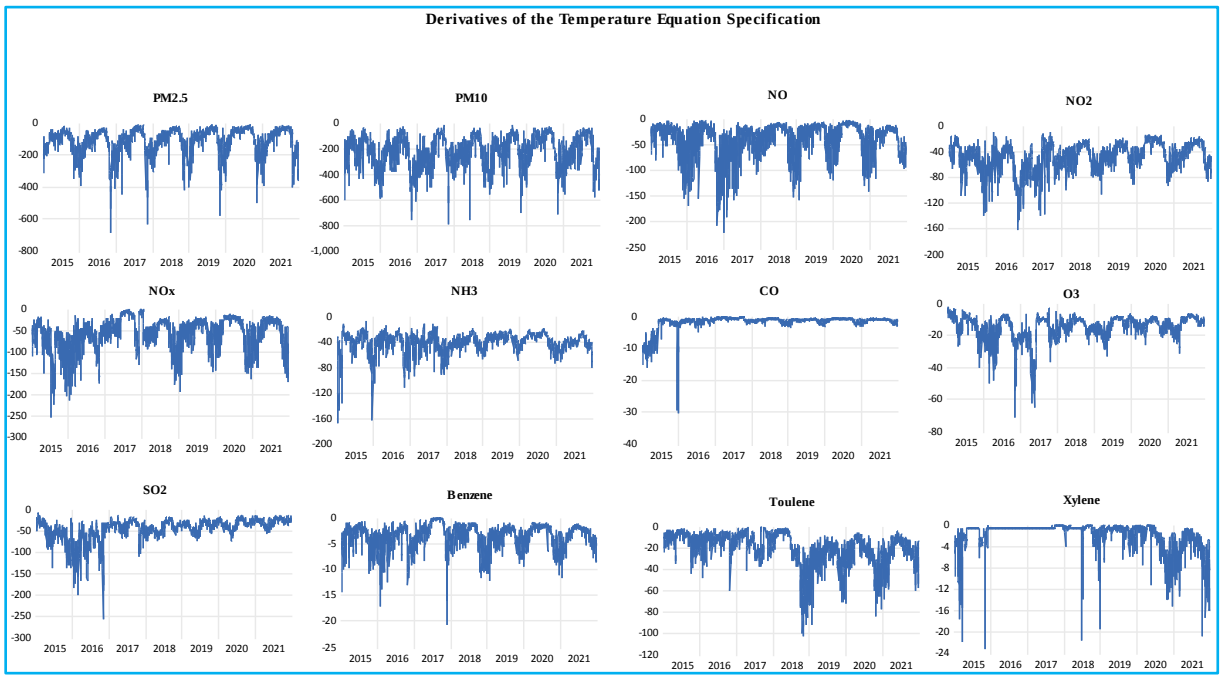
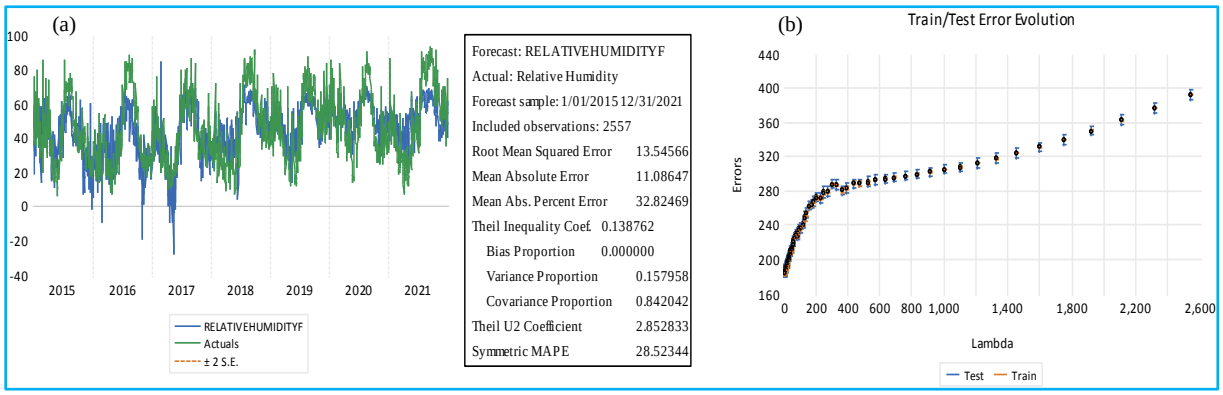


Fig. 6. Daily temperature equation derivatives

dependent variables. After running the model, the predicted results demonstrate that air pollutants account for 46.29% of the variance in change in daily temperature. Similarly, Lambda (λ) at minimum error comes out to be 0.0716, RMSE (Root Mean Square Error) came as 5.64, and MAE is 4.42, as shown in Fig. 5. The results indicate that the air pollutants influencing Delhi's temperature change are portrayed in equation (2). The equation derivatives of temperature are displayed in Fig. 6.

$$\text{Daily Temperature}_{\text{Delhi}} = 25.6+0.027*PM_{10}+0.025*NO_2+0.085*O_3-0.17*CO+0.084*SO_2+0.025*NO_2-0.076PM_{2.5}-0.014*NO_x-0.03*NO-0.013*NH_3-0.74*C_6H_6+0.064*C_7H_8+0.082*C_8\,H_{10} \quad (2)$$



Figs. 7(a&b). Relative humidity model output after testing {graph (a), X- axisdenotes Time (Years) and (b) Y- axis denotes relative humidity}

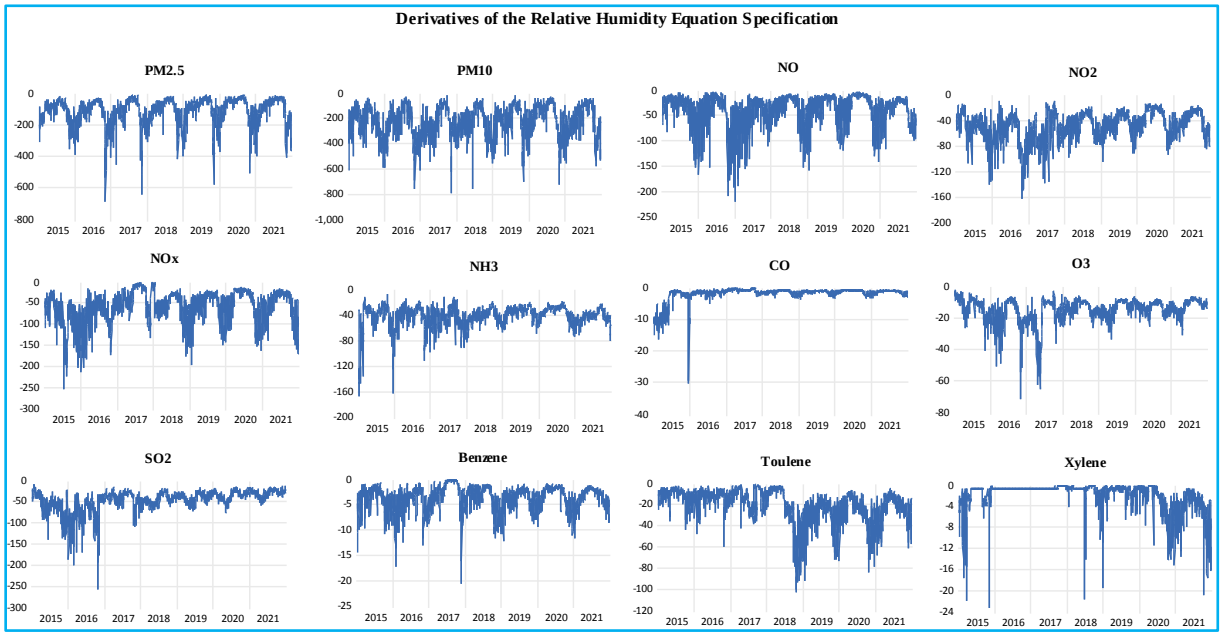
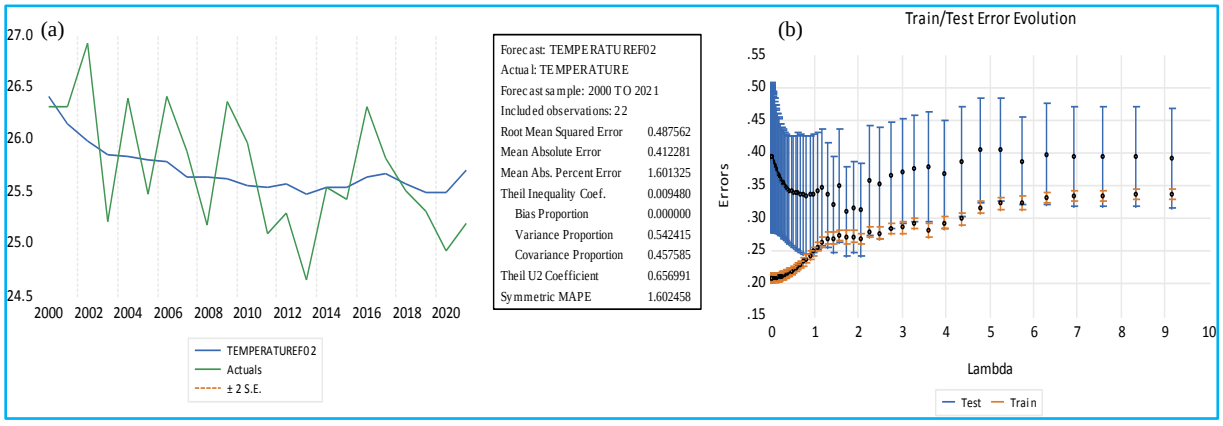


Fig. 8. Daily relative humidity equation derivatives



Figs. 9(a&b). Annual average temperature model output {graph (a), X-axis denotes Time (Years) and (b) Y- axis denotes temperature (°C)}

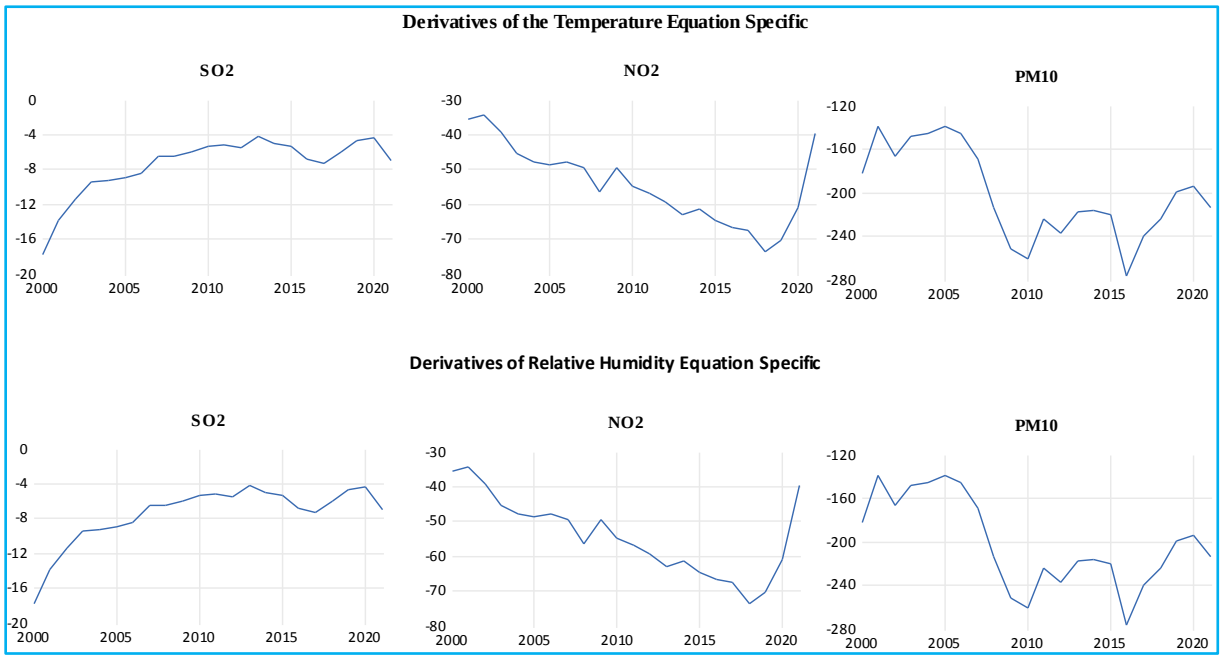
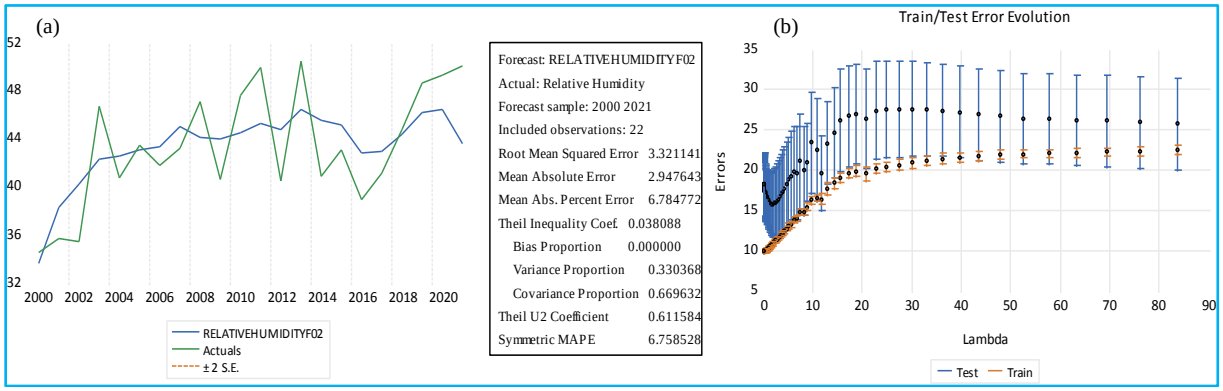


Fig. 10. Annual average temperature and relative humidity equation derivatives



Figs. 11(a&b). Annual average relative humidity output {graph (a), X -denotes Time (Years) (b) Y-axis denotes relative humidity}

3.1.2. *Association among criteria air pollutants and relative humidity*

The stepwise procedure of elastic net regression was performed to find an association between predictors and dependent variables. After running the model, the predicted results demonstrate that air pollutants account for 53.36% of the variance in change in daily relative humidity. Similarly, Lambda (λ) at minimum error comes out to be 0.2544, RMSE (Root Mean Square Error) came as 13.54 and MAE is 11.08, as shown in Fig. 7. The results indicate that the air pollutants influencing Delhi's humidity change are portrayed in Equation(3). The equation derivatives of relative humidity are displayed in Fig. 8.

$$\text{Daily Relative Humidity}_{\text{Delhi}} = 78.58+0.072*PM_{2.5}+0.056*NO_x-0.133*NO_2-0.11*O_3-0.621*CO-0.97*SO_2+0.025*NO_2-0.102PM_{10}+0.11*NO-0.028*NH_3-0.58*C_6\ H_6+0.27*C_7\ H_8-0.35*C_8\ H_{10} \quad (3)$$

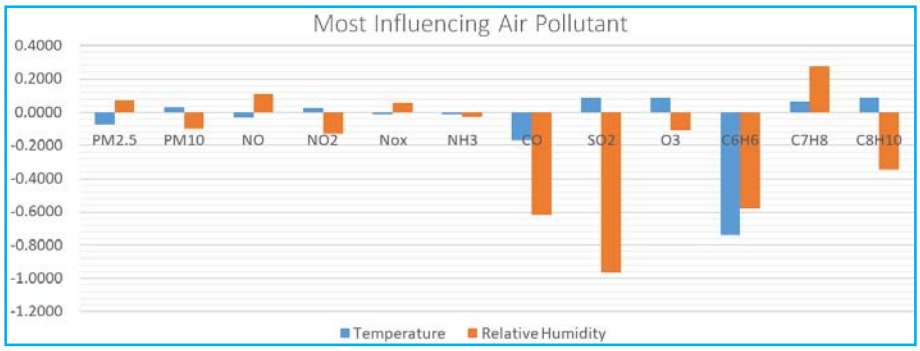


Fig. 12. Most influencing pollutant during 2015 to 2021

3.2. *Analysis output of the annual average model (2000-2021), including three criteria air pollutants*

3.2.1. *Association among criteria air pollutants and temperature*

The stepwise procedure of elastic net regression was performed to find an association between predictors and dependent variables. After running the model, the predicted results demonstrate that air pollutants account for 30.51% of the variance in change in average annual temperature. Similarly, Lambda (λ) at minimum error comes out to be 1.715, RMSE (Root Mean Square Error) came as 0.487, and MAE is 0.412, as shown in Fig. 9.Looking at the unique individual contributions of the predictors (pollutants), the results show that the air pollutants which influenced the temperature change in Delhi are portrayed in Equation 4. The equation derivatives of temperature are displayed in Fig. 10.

$$\text{AnnualTemperature}_{\text{Delhi}} = 25.31 + 0.0649 * SO_2 + 0.0023 * PM_{10} - 0.0016 * NO_2 \quad (4)$$

3.2.2. *Association among criteria air pollutants and relative humidity*

The stepwise procedure of elastic net regression was performed to find an association between predictors and dependent variables. After running the model, the predicted results demonstrate that air pollutants account for 51.7% of the variance in change in daily relative humidity. Similarly, Lambda (λ) at minimum error comes out to be 1.847, RMSE (Root Mean Square Error) came as 3.321, and MAE is 2.94, as shown in Fig. 11. Looking at the unique individual contributions of the predictors (pollutants), the results indicate that the air pollutants influencing the humidity change in Delhi are portrayed in Equation 5. The equation derivatives of relative humidity are displayed in Fig. 10.

$$\text{Annual Relative Humidity}_{\text{Delhi}} = 54.122 + 0.00042 * NO_2 - 0.977 * SO_2 - 0.016 * PM_{10} \quad (5)$$

3.3. *Impact of air pollutants on climate change*

3.3.1. *Based on the daily dataset*

Findings that appeared after running the best available method are portrayed in Fig. 12 showing the footprint of air pollutants on temperature and relative humidityfrom 2015 to 2021. The results from this study indicate that $PM_{2.5}$, PM_{10}, NO, NO_2, NO_x, SO_2, and O_3 showed a high correlation with the meteorological variables.Through all these findings of this study, it is seen that most of the pollutants in our atmosphere have significantly positive impacts on temperature and relative humidity. It is well known that both temperature and relative humidity are inversely proportional. Hence, the pollutants showing a positive association with Temperature and negative association with relative humidity work as an influencer for the climate change and *vice versa* (Khan *et al*., 2022; Sein *et al*., 2022; Uzuazor *et al*., 2021). Table 1 indicates the beta values that came out from this study. In the third column, indications are given. Green indication means the pollutant has significantly influenced climate change, and red indicates that there is no role that particular pollutant has played.

3.3.2. *Based on the annual average dataset*

Findings that appeared after running the best available method show the influences of air pollutants on the temperature and relative humidity change from 2000 to 2021. The results appeared after running the model with

TABLE 1

Climate change indication on daily basis

Pollutants	Temperature	Relative Humidity	Association
$PM_{2.5}$	-0.0763	0.0719	
PM_{10}	0.0274	-0.1029	
NO	-0.0302	0.1111	
NO_2	0.0259	-0.1335	
NOx	-0.0150	0.0562	
NH_3	-0.0134	-0.0281	
CO	-0.1697	-0.6212	
SO_2	0.0844	-0.9672	
O_3	0.0860	-0.1112	
C_6H_6	-0.7397	-0.5823	
C_7H_8	0.0641	0.2737	
C_8H_{10}	0.0826	-0.3490	

TABLE 2

Climate change indication on an annual basis

Pollutants	Temperature	Relative Humidity	Association
SO_2	0.0649	-0.977	
NO_2	-0.0016	0.00042	
PM_{10}	0.0023	-0.016	

the average annual dataset it was observed that all three criteria air pollutants whose data has been collected are showing significant association in enhancing the process of climatic change in Delhi over the past 22 years. It was found that SO_2, NO_2 and PM_{10} have demonstrated positive and negative associations with temperature and relative humidity, satisfying the condition of both being inversely proportional and the role of each pollutant in increasing the impact of change in the weather of Delhi. Table 2 indicates the beta values that came out from this study. In the third column, green indicates that the pollutant has significantly influenced climate change, and red indicates no role that particular pollutant has played.

3.4. *Association among pollutants and climatic variables*

In the below sections, a comparison between the findings with the results of previously available literature will be discussed, the effects of the combination of pollutants and climate change on human health.

3.4.1. *Particulate Matters ($PM_{2.5}$ and PM_{10}) and Meteorological Variables (Temp, RH)*

Particulate matter (PM) is currently one of the most severe urban pollutants due to its known adverse effects on human health, climate change, ecosystem damage, and its role in impaired visibility. It has been observed from previous studies that long- and short-term exposure to particulate matters can lead to premature death, non-fatal cardiovascular events, and respiratory effects (Hassan *et al.*, 2022; Jain & Barthwal, 2022; Liao *et al.*, 2022; Priyan *et al.*, 2022; Sasmita *et al.*, 2022; Sharma *et al.*, 2022; Sisani *et al.*, 2022). In this study, both $PM_{2.5}$ [with temperature (β = -0.076, R^2 = 46.29%) and relative humidity (β = 0.0719, R^2 = 53.36%)] and PM_{10} [with temperature (β = 0.0274, R^2 = 46.29%) and relative humidity (β = -0.1029, R^2 = 53.36%)] shown significant association in influencing the climate change process. Additionally, $PM_{10,}$ when assessed with the yearly dataset of the past 22 years, showed a significant association with climate change, which further indicates that particulate

matter is playing a vital role in enhancing the process of climatic change in Delhi. Previous studies also observed a high association between particulate matter and meteorological parameters, *i.e.*, (Chithra & Shiva Nagendra, 2014; Goyal & Jaiswal, 2010; "IMPACT OF METEOROLOGICAL FACTORS ON $PM_{2.5}$ IN CHENNAI," 2017; Li *et al*., 2017; Maleki & Molaei, 2022; Sharma & Sharma, 2016. The government has to take additional preventive measures to stop particulate matter emissions from controlling the high pollution levels and even climate change.

3.4.2. *Oxides of Nitrogen (NO, NO_2 and NO_X) and Meteorological Variables (Temp, RH)*

Nitrogen dioxide (NO_2) forms when nitrogen oxide (NO) and other nitrogen oxides (NO_x) react with other chemicals in the air to form nitrogen dioxide, which means they are associated with each other (Kumar *et al*., 2015; Tiwari *et al*., 2015).

(*i*) Delhi has recorded an increase of 125% in NO2 pollution in the past year (https://www.greenpeace.org/india/en/). NO_2 is also an indirect greenhouse gas. It forms photochemical reactions in the atmosphere and produces tropospheric ozone. NO_2 exposure can have a negative influence on people's health at any age, damaging the respiratory, circulatory, and neural impulses, as well as the brain, leading to a rise in hospital admissions and mortality (Boningari & Smirniotis, 2016; Erickson *et al*., 2020; Kjellstrom *et al*., 2002). In this study, NO_2 [with temperature (β = 0.0259, R^2 = 46.29%), and relative humidity (β = -0.1335, R^2 = 53.36%)] showed significant association in influencing the climate change process. Additionally, the annual study of the past 22 years also states that NO_2 plays a vital role in the temperature change in Delhi. The most significant sources of NO_2 pollution in Delhi are motor vehicles and fossil fuel industries. Governments, local governments and city planners must begin the transition from privately owned automobiles to an efficient, clean, and safe public transportation system powered by renewable energy, which will not only assist in reducing pollution but also to decrease the rate of climate change.

(*ii*) Both NO (with temperature (β = -0.032, R^2 = 46.29%) and relative humidity (β = 0.111, R^2 = 53.36%) and NO_x [with temperature (β = -0.0150, R^2 = 46.29%), and relative humidity (β = 0.0562, R^2 = 53.36%)] shown significant association in influencing the climate change process. Previous studies indicating association among these both including NO_2, *i.e*., (Cheng & Lam, 2000; Ghosh *et al*., 2017; Jiang *et al*., 2005; Khiem *et al*., 2010; Lin *et al*., 2021; Mahapatra *et al*., 2014; Sharma & Sharma, 2016).

3.4.3. *Ammonia (NH_3), Carbon Monoxide (CO), and Meteorological Variables (Temp, RH)*

(*i*) Ammonia presents for a shorter period in the atmosphere. Hence it has a global warming potential of zero which even matches with the findings of this study NH_3 [with temperature (β = -0.0134, R^2 = 46.29%) and relative humidity (β = -0.0281, R^2 = 53.36%)], had not shown any significant association in influencing the climate change process (Majeed *et al*., 2020).

(*ii*) On the other hand, CO does not absorb terrestrial thermal IR energy strongly enough, which indicates that it is not a greenhouse gas. In this study, CO [with temperature (β = -0.1679, R^2 = 46.29%) and relative humidity (β = -0.6212, R^2 = 53.36%)] had not shown significant association in influencing the climate change process (Majeed *et al*., 2020).

3.4.4. *Ground-level Ozone (O_3), Sulphur dioxide (SO_2) and Meteorological Variables (Temp, RH)*

(*i*) Ground-level ozone levels tend to high in summer due to the elevated heat and sunlight. With the help of the available dataset, the link between ground-level ozone and temperature was assessed, and it was found that the Pearson Correlation came out as (r = 0.12). Recently, in many studies, it has been found that ground-level ozone is highly associated with heat waves on hot days. It also indicates that ozone also helps with climatic change.

(*ii*) On the other hand, SO_2 is not a direct greenhouse gas. It is considered an indirect greenhouse gas because it forms aerosols that contribute to the planet's cooling and warming when coupled with elemental carbon. From this study, it is observed that SO_2 [with temperature (β = 0.0844, R^2 = 46.29%) and relative humidity (β = -0.9672, R^2 = 53.36%)] showed significant association in influencing the climate change process after forming aerosols in Delhi region. These have also been justified with the second model in which 22 years of annual average data was assessed (Abdul-Wahab, 2006; Calkins *et al*., 2016; Ghosh *et al*., 2017; Gvozdić *et al*., 2011; Kim & Lim, 2005; Sezer Turalıoğlu *et al*., 2005; Unnikrishnan & Madhu, 2019; Xue & Yin, 2014; Yang *et al*., 2022).

3.4.5. *Benzene, Toluene, Xylene, and Meteorological Variables (Temp, RH)*

In this study, both C_6H_6 [with temperature (β = -0.7397, R^2 = 46.29%) and relative humidity (β = -0.5823, R^2 = 53.36%)] and C_7H_8 [with temperature

(β = 0.0641, R^2 = 46.29%) and relative humidity (β = 0.2737, R^2 = 53.36%)], has not shown positive association on the climate change process. Only C_8H_{10} [with temperature (β = 0.0826, R^2 = 46.29%) and relative humidity (β = -0.3490, R^2 = 53.36%)] are shown to have a significant association in influencing the climate change process, making it an indirect greenhouse gas. Additionally, this study has observed that Toluene positively affects both temperature and relative humidity, indicating it is a strong predictor in influencing temperature and relative humidity.

4. Conclusion

A daily dataset of twelve air pollutants (2015-2021), an average annual dataset with three criteria air pollutants (2000-2021), and the best regression model were used to assess the impact of air pollutants on the temperature and relative humidity. The results were obtained in which $PM_{2.5}$, PM_{10}, NO, NO_2, NO_x, SO_2, O_3 and C_8H_{10} showed a significant positive association in enhancing the process of climatic change in the Delhi region. Additionally, NH_3, CO, C_6H_6 and C_7H_8 showed a negative association with climatic change. Toluene, in this study, appeared as a pollutant that has shown a positive relationship with both temperature and relative humidity. It can be studied whether it shows the same behavior in other regions. From all these findings, it has been concluded most of the pollutants were showing significant association in enhancing the rate of climate change in Delhi. The Government of Delhi has to make effective policies to control the emission of pollutants in this region, which will further help minimize air pollution's influence on climate change and human health.

Disclaimer : The contents and views expressed in this study are the views of the authors and do not necessarily reflect the views of the organizations they belong to.

References

Abdul-Wahab, S. A., 2006, "The role of meteorology on predicting SO_2 concentrations around a refinery: A case study from Oman",*Ecological Modelling*, **197**, 1-2, 13-20. https://doi.org/10.1016/j.ecolmodel.2006.02.021.

Barzeghar, V., Hassanvand, M. S., Faridi, S., Abbasi, S. and Gholampour, A., 2022, "Long-term trends in ambient air pollutants and the effect of meteorological parameters in Tabriz, Iran", *Urban Climate*, **42**, 101119. https://doi.org/10.1016/j.uclim.2022.101119.

Biswas, J., Upadhyay, E., Nayak, M. and Yadav, A. K., 2011, "An Analysis of Ambient Air Quality Conditions Over Delhi, India From 2004 to 2009", *Atmospheric and Climate Sciences*, **01**, 04, 214-224. https://doi.org/10.4236/acs.2011.14024.

Boningari, T. and Smirniotis, P. G., 2016, "Impact of nitrogen oxides on the environment and human health : Mn-based materials for the NOx abatement", *Current Opinion in Chemical Engineering*, **13**, 133-141. https://doi.org/10.1016/j.coche.2016.09.004.

Calkins, C., Ge, C., Wang, J., Anderson, M. and Yang, K., 2016, "Effects of meteorological conditions on sulfur dioxide air pollution in the North China plain during winters of 2006-2015", *Atmospheric Environment*, **147**, 296-309. https://doi.org/10.1016/j.atmosenv.2016.10.005.

Cheng, S. and Lam, K. C., 2000, "Synoptic typing and its application to the assessment of climatic impact on concentrations of sulfur dioxide and nitrogen oxides in Hong Kong", *Atmospheric Environment*, **34**, 4, 585-594. https://doi.org/10.1016/S1352-2310(99)00194-6.

Chithra, V. S. and Shiva Nagendra, S. M., 2014, "Impact of outdoor meteorology on indoor PM_{10}, $PM_{2.5}$ and PM_1 concentrations in a naturally ventilated classroom", *Urban Climate*, **10**, 77-91. https://doi.org/10.1016/j.uclim.2014.10.001.

Dandotiya, B., Jadon, N. and Sharma, H. K., 2019, "Effects of Meteorological Parameters on Gaseous Air Pollutant Concentrations in Urban Area of Gwalior City, India", *Environmental Claims Journal*, **31**, 1, 32-43. https://doi.org/10.1080/10406026.2018.1507508.

Das, M., Das, A., Sarkar, R., Mandal, P., Saha, S. and Ghosh, S., 2021, "Exploring short term spatio-temporal pattern of $PM_{2.5}$ and PM_{10} and their relationship with meteorological parameters during COVID-19 in Delhi", *Urban Climate*, **39**, 100944. https://doi.org/10.1016/j.uclim.2021.100944.

Depietri, Y., Renaud, F. G. and Kallis, G., 2012, "Heat waves and floods in urban areas: a policy-oriented review of ecosystem services", *Sustainability Science*, **7**, 1, 95-107. https://doi.org/10.1007/s11625-011-0142-4.

Deswal, S. and Chandna, P., 2010, "Effects of meteorological conditions on concentration of air pollutants in Delhi", *International Journal of Environment and Pollution*, **42**, 1/2/3, 58. https://doi.org/10.1504/IJEP.2010.034226.

Ding, J., Dai, Q., Li, Y., Han, S., Zhang, Y. and Feng, Y., 2021, "Impact of meteorological condition changes on air quality and particulate chemical composition during the COVID-19 lockdown", *Journal of Environmental Sciences*, **109**, 45-56. https://doi.org/10.1016/j.jes.2021.02.022.

Edouard Grave, G. R. O. F. B., 2011, "Trace Lasso: a trace norm regularization for correlated designs", *Advances in Neural Information Processing Systems*, 24.

Erickson, L. E., Newmark, G. L., Higgins, M. J. and Wang, Z., 2020, "Nitrogen oxides and ozone in urban air: A review of 50 plus years of progress", *Environmental Progress & Sustainable Energy*, **39**, 6. https://doi.org/10.1002/ep.13484.

Ghosh, D., Lal, S. and Sarkar, U., 2017, "Variability of tropospheric columnar NO_2 and SO_2 over eastern Indo-Gangetic Plain and impact of meteorology", *Air Quality, Atmosphere & Health*, **10**, 5, 565-574. https://doi.org/10.1007/s11869-016-0451-y.

Godde, C. M., Mason-D'Croz, D., Mayberry, D. E., Thornton, P. K. and Herrero, M., 2021, "Impacts of climate change on the livestock food supply chain; a review of the evidence", *Global Food Security*, **28**, 100488. https://doi.org/10.1016/j.gfs.2020.100488.

Goyal, P. and Jaiswal, N., 2010, "Effects of meteorological parameters on RSPM concentration in urban Delhi", *International Journal of Environment and Waste Management*, **5**, 3/4, 237. https://doi.org/10.1504/IJEWM.2010.032006.

Guo, Q., Wu, D., Yu, C., Wang, T., Ji, M. and Wang, X., 2022, "Impacts of meteorological parameters on the occurrence of air pollution episodes in the Sichuan basin", *Journal of Environmental*

Sciences, **114**, 308-321. https://doi.org/10.1016/j.jes.2021.09.006.

Gupta, A. K., Patil, R. S. and Gupta, S. K., 2004, "Influence of meteorological factors on air pollution concentration for a coastal region in India", *International Journal of Environment and Pollution*, **21**, 3, 253. https://doi.org/10.1504/IJEP.2004.004193.

Guttikunda, S., 2010, "Simple Interactive Models for Better Air Quality Role of Meteorology on Urban Air Pollution Dispersion : A 20yr Analysis for Delhi, India", www.urbanemissions.info.

Guttikunda, S. K. and Gurjar, B. R., 2012, "Role of meteorology in seasonality of air pollution in megacity Delhi, India", *Environmental Monitoring and Assessment*, **184**, 5, 3199-3211. https://doi.org/10.1007/s10661-011-2182-8.

Gvozdić, V., Kovač-Andrić, E. and Brana, J., 2011, "Influence of Meteorological Factors NO_2, SO_2, CO and PM_{10} on the Concentration of O_3 in the Urban Atmosphere of Eastern Croatia", *Environmental Modeling& Assessment*, **16**, 5, 491-501. https://doi.org/10.1007/s10666-011-9256-4.

Hassan, H., Kumar, P. and Kakosimos, K. E., 2022, "The impact of local fugitive particulate matter and emission inventories on air quality and health in dry and arid areas",*S cience of The Total Environment*, **824**, 153799. https://doi.org/10.1016/j.scitotenv.2022.153799.

Heydarzadeh, H., Jafari, H. and Karimi, S., 2022, "Effects of meteorological parameters and fuel composition on the air pollution production from motor vehicles", *Environmental Monitoring and Assessment*, **194**, 4, 236. https://doi.org/10.1007/s10661-022-09866-0.

Ilten, N. and Selici, A. T., 2008, "Investigating the impacts of some meteorological parameters on air pollution in Balikesir, Turkey", *Environmental Monitoring and Assessment*, **140**, 1-3, 267-277. https://doi.org/10.1007/s10661-007-9865-1.

Jain, S. and Barthwal, V., 2022, "Health impact assessment of auto rickshaw and cab drivers due to exposure to vehicular pollution in Delhi : an integrated approach", *Environmental Science and Pollution Research*, **29**, 4, 5124-5133. https://doi.org/10.1007/s11356-021-16058-9.

Jayamurugan, R., Kumaravel, B., Palanivelraja, S. and Chockalingam, M. P., 2013, "Influence of Temperature, Relative Humidity and Seasonal Variability on Ambient Air Quality in a Coastal Urban Area", *International Journal of Atmospheric Sciences*, 2013, 1-7. https://doi.org/10.1155/2013/264046.

Jiang, Hay and Fisher, 2005, "Effects of meteorological conditions on concentrations of nitrogen oxides in Auckland", *Weather and Climate*, **24**, 15. https://doi.org/10.2307/26169672.

Khan, A. R., Abedin, S. and Khan, S., 2022, "Interaction of Temperature and relative humidity for growth of COVID-19 cases and death rates", *Environmental Research Letters*, **17**, 3, 034048. https://doi.org/10.1088/1748-9326/ac4cf2.

Khedairia, S. and Khadir, M. T., 2012, "Impact of clustered meteorological parameters on air pollutants concentrations in the region of Annaba, Algeria", *Atmospheric Research*, **113**, 89-101. https://doi.org/10.1016/j.atmosres.2012.05.002.

Khiem, M., Ooka, R., Huang, H., Hayami, H., Yoshikado, H. and Kawamoto, Y., 2010, "Analysis of the Relationship between Changes in Meteorological Conditions and the Variation in Summer Ozone Levels over the Central Kanto Area", *Advances in Meteorology*, 2010, 1-13. https://doi.org/10.1155/2010/349248.

Kim, J. H. and Lim, J. W., 2005, "Multivariate Analysis between Air Pollutants and Meteorological Variables in Seoul", *Key Engineering Materials*, 277-279, 492-496. https://doi.org/10.4028/www.scientific.net/KEM.277-279.492.

Kim, M. H., Banerjee, S., Park, S. M. and Pathak, J., 2016, "Improving risk prediction for depression via Elastic Net regression - Results from Korea National Health Insurance Services Data", *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, 2016, 1860-1869.

Kjellstrom, T. E., Neller, A. and Simpson, R. W., 2002, "Air pollution and its health impacts: the changing panorama", *Medical Journal of Australia*, **177**, 11, 604-608. https://doi.org/10.5694/j.1326-5377.2002.tb04982.x.

Krishan, M., Jha, S., Das, J., Singh, A., Goyal, M. K. and Sekar, C., 2019, "Air quality modelling using long short-term memory (LSTM) over NCT-Delhi, India", *Air Quality, Atmosphere & Health*, **12**, 8, 899-908. https://doi.org/10.1007/s11869-019-00696-7.

Kuang, D., Choo, J. and Park, H., 2015, "Nonnegative Matrix Factorization for Interactive Topic Modeling and Document Clustering", In Partitional Clustering Algorithms (215-243). Springer International Publishing. https://doi.org/10.1007/978-3-319-09259-1_7.

Kucbel, M., Corsaro, A., Švédová, B., Raclavská, H., Raclavský, K. and Juchelková, D., 2017, "Temporal and seasonal variations of black carbon in a highly polluted European city: Apportionment of potential sources and the effect of meteorological conditions", *Journal of Environmental Management*, **203**, 1178-1189. https://doi.org/10.1016/j.jenvman.2017.05.038.

Kumar, A., Singh, D., Singh, B. P., Singh, M., Anandam, K., Kumar, K. and Jain, V. K., 2015, "Spatial and temporal variability of surface ozone and nitrogen oxides in urban and rural ambient air of Delhi-NCR, India", *Air Quality, Atmosphere & Health*, **8**, 4, 391-399. https://doi.org/10.1007/s11869-014-0309-0.

Kuzu, S. L. and Saral, A., 2017, "The effect of meteorological conditions on aerosol size distribution in Istanbul", *Air Quality, Atmosphere & Health*, **10**8, 1029-1038. https://doi.org/10.1007/s11869-017-0491-y.

Li, X., Ma, Y., Wang, Y., Liu, N. and Hong, Y., 2017, "Temporal and spatial analyses of particulate matter (PM_{10} and $PM_{2.5}$) and its relationship with meteorological parameters over an urban city in northeast China", *Atmospheric Research*, **198**, 185-193. https://doi.org/10.1016/j.atmosres.2017.08.023.

Liao, J., Liu, Y., Steenland, K., Pillarisetti, A., Thompson, L. M., Dey, S., Balakrishnan, K. and Clasen, T., 2022, "Child Survival and Early Lifetime Exposures to Ambient Fine Particulate Matter in India : A Retrospective Cohort Study", *Environmental Health Perspectives*, **130**, 1. https://doi.org/10.1289/EHP8910.

Lin, C., Lau, A. K. H., Fung, J. C. H., Song, Y., Li, Y., Tao, M., Lu, X., Ma, J. and Lao, X. Q., 2021, "Removing the effects of meteorological factors on changes in nitrogen dioxide and ozone concentrations in China from 2013 to 2020", *Science of The Total Environment*, **793**, 148575. https://doi.org/10.1016/j.scitotenv.2021.148575.

Mahanta, S. K., Patil, S. S., Mahanta, B., Biswas, K., Sahu, R., Panda, B. S. and Mallik, M. R., 2021, "Influence of Meteorological Variables on Ambient Air Pollutants of a Coastal District in Eastern India", *Oriental Journal Of Chemistry*, **37**, 1, 194-203. https://doi.org/10.13005/ojc/370127.

Mahapatra, P. S., Panda, S., Walvekar, P. P., Kumar, R., Das, T. and Gurjar, B. R., 2014, "Seasonal trends, meteorological impacts,

and associated health risks with atmospheric concentrations of gaseous pollutants at an Indian coastal city", *Environmental Science and Pollution Research*, **21**, 19, 11418-11432. https://doi.org/10.1007/s11356-014-3078-2.

Majeed, H. T., Nassif, W. G. and Tawfeek, Y. Q., 2020, "Effects of Meteorological Parameters on Surface Concentration of Carbon Monoxide over Iraq Synoptic Analysis of Potential Evaporation and Soil Moisture", View project Climate system View project Effects of Meteorological Parameters on Surface Concentration of Carbon Monoxide over Iraq. *Journal of Green Engineering*, **10**, 9, 5927-5940. https://doi.org/10.13140/RG.2.2.32807.98720.

Maleki, R. and Molaei, I., 2022, "The effect of meteorological parameters on $PM_{2.5}$ concentration changes in 2018 (Case study: Tehran)", *Journal of Air Pollution and Health*. https://doi.org/10.18502/japh.v6i4.8586.

Marlier, M. E., Jina, A. S., Kinney, P. L. and DeFries, R. S., 2016, "Extreme Air Pollution in Global Megacities", *Current Climate Change Reports*, **2**, 1, 15-27. https://doi.org/10.1007/s40641-016-0032-z.

McNeish, D. M., 2015, "Using Lasso for Predictor Selection and to Assuage Overfitting: A Method Long Overlooked in Behavioral Sciences", *Multivariate Behavioral Research*, **50**, 5, 471-484. https://doi.org/10.1080/00273171.2015.1036965.

Priyan R, S., Peter, A. E., Menon, J. S., George, M., Nagendra, S. M. S. and Khare, M., 2022, "Composition, sources, and health risk assessment of particulate matter at two different elevations in Delhi city", *Atmospheric Pollution Research*, **13**, 2, 101295. https://doi.org/10.1016/j.apr.2021.101295,

Ramsey, N. R., Klein, P. M.and Moore, B., 2014, "The impact of meteorological parameters on urban air quality", *Atmospheric Environment*, **86**, 58-67. https://doi.org/10.1016/j.atmosenv.2013.12.006.

Ravindra, K., Sidhu, M. K., Mor, S., John, S. and Pyne, S., 2016, "Air Pollution in India : Bridging the Gap between Science and Policy", *Journal of Hazardous, Toxic, and Radioactive Waste*, **20**, 4. https://doi.org/10.1061/(ASCE)HZ.2153-5515.0000303.

Saha, U., Talukdar, S., Jana, S. and Maitra, A., 2014, "Effects of air pollution on meteorological parameters during Deepawali festival over an Indian urban metropolis", *Atmospheric Environment*, **98**, 530-539. https://doi.org/10.1016/j.atmosenv.2014.09.032.

Sasmita, S., Kumar, D. B. and Priyadharshini, B., 2022, "Assessment of sources and health impacts of PM_{10} in an urban environment over eastern coastal plain of India", *Environmental Challenges*, **7**, 100457. https://doi.org/10.1016/j.envc.2022.100457.

Sati, A. P. and Mohan, M., 2018, "The impact of urbanization during half a century on surface meteorology based on WRF model simulations over National Capital Region, India", *Theoretical and Applied Climatology*, **134**, 1-2, 309-323. https://doi.org/10.1007/s00704-017-2275-6.

Sein, Z. M. M., Ullah, I., Iyakaremye, V., Azam, K., Ma, X., Syed, S. and Zhi, X., 2022, "Observed spatiotemporal changes in air temperature, dew point temperature and relative humidity over Myanmar during 2001-2019", *Meteorology and Atmospheric Physics*, **134**, 1, 7. https://doi.org/10.1007/s00703-021-00837-7.

SezerTuralıoğlu, F., Nuhoğlu, A. and Bayraktar, H., 2005, "Impacts of some meteorological parameters on SO_2 and TSP concentrations in Erzurum, Turkey", *Chemosphere*, **59**, 11, 1633-1642. https://doi.org/10.1016/j.chemosphere.2005.02.003.

Sharma, A. K., Baliyan, P. and Kumar, P., 2018, "Air pollution and public health: the challenges for Delhi, India", *Reviews on Environmental Health*, **33**, 1, 77-86. https://doi.org/10.1515/reveh-2017-0032.

Sharma, D. and Mauzerall, D., 2022, "Analysis of Air Pollution Data in India between 2015 and 2019", *Aerosol and Air Quality Research*, **22**, 2, 210204. https://doi.org/10.4209/aaqr.210204.

Sharma, P., Peshin, S. K., Soni, V. K., Singh, S., Beig, G. and Ghosh, C., 2022, "Seasonal dynamics of particulate matter pollution and its dispersion in the city of Delhi, India", *Meteorology and Atmospheric Physics*, **134**, 2, 28. https://doi.org/10.1007/s00703-021-00852-8.

Sharma, R. C. and Sharma, N., 2016, "Influence of Some Meteorological Variables on PM_{10} and NO_x in Gurgaon, Northern India", *American Journal of Environmental Protection*, **4**, 1, 1-6. https://doi.org/10.12691/env-4-1-1.

Sisani, F., di Maria, F. and Cesari, D., 2022, "Environmental and human health impact of different powertrain passenger cars in a life cycle perspective. A focus on health risk and oxidative potential of particulate matter components", *Science of The Total Environment*, **805**, 150171. https://doi.org/10.1016/j.scitotenv.2021.150171.

Sonwani, S., Saxena, P. and Shukla, A., 2021, "Carbonaceous Aerosol Characterization and Their Relationship With Meteorological Parameters During Summer Monsoon and Winter Monsoon at an Industrial Region in Delhi, India", *Earth and Space Science*, **8**, 4. https://doi.org/10.1029/2020EA001303.

Tiwari, S., Bisht, D. S., Srivastava, A. K., Pipal, A. S., Taneja, A., Srivastava, M. K. and Attri, S. D., 2014, "Variability in atmospheric particulates and meteorological effects on their mass concentrations over Delhi, India", *Atmospheric Research*, 145-146, 45-56. https://doi.org/10.1016/j.atmosres.2014.03.027.

Tiwari, S., Dahiya, A. and Kumar, N., 2015, "Investigation into relationships among NO, NO_2, NO_x, O_3 and CO at an urban background site in Delhi, India", *Atmospheric Research*, **157**, 119-126. https://doi.org/10.1016/j.atmosres.2015.01.008.

Trivedi, D. K., Ali, K. and Beig, G., 2014, "Impact of meteorological parameters on the development of fine and coarse particles over Delhi", *Science of The Total Environment*, **478**, 175-183. https://doi.org/10.1016/j.scitotenv.2014.01.101.

Unnikrishnan, R. and Madhu, G., 2019, "Comparative study on the effects of meteorological and pollutant parameters on ANN modelling for prediction of SO_2", *SN Applied Sciences*, **1**, 11, 1394. https://doi.org/10.1007/s42452-019-1440-1.

Uzuazor Siloko, I., Eghonghon Ukhurebor, K., Akpevwe Siloko, E., Enoyoze, E. and Ikpotokin, O., 2021, "The Interactions between Temperature and Relative Humidity: Results for Benin City, Nigeria using Statistical Analysis", *Current Applied Science and Technology*. https://doi.org/10.55003/cast.2022.01.22.009.

Verma, S. S. and Desai, B., 2008, "Effect of Meteorological Conditions on Air Pollution of Surat City #", *In J. Int. Environmental Application & Science* (Vol. **3**, Issue 5).

Xue, D. and Liu, Q., 2014, "Investigating the Impacts of Meteorological Parameters on Air Pollution in Shanghai", *Applied Mechanics and Materials*, 675-677, 643-646. https://doi.org/10.4028/www.scientific.net/AMM.675-677.643.

Xue, D. and Yin, J., 2014, "Meteorological influence on predicting surface SO_2 concentration from satellite remote sensing in Shanghai, China", *Environmental Monitoring and Assessment*, **186**, 5, 2895-2906. https://doi.org/10.1007/s10661-013-3588-2.

Yadav, S., Praveen, O. D. and Satsangi, P. G., 2015, "The effect of climate and meteorological changes on particulate matter in Pune, India", *Environmental Monitoring and Assessment*, **187**, 7, 402. https://doi.org/10.1007/s10661-015-4634-z.

Yang, X., Zeng, G., Iyakaremye, V. and Zhu, B., 2022, "Effects of different types of heat wave days on ozone pollution over Beijing-Tianjin-Hebei and its future projection", *Science of The Total Environment*, **837**, 155762. https://doi.org/10.1016/j.scitotenv.2022.155762.

Yarlagadda, B., Smith, S. J., Mignone, B. K., Mallapragada, D., Randles, C. A. and Sampedro, J., 2022, "Climate and air pollution implications of potential energy infrastructure and policy measures in India", *Energy and Climate Change*, **3**, 100067. https://doi.org/10.1016/j.egycc.2021.100067.

Zou, H. and Hastie, T., 2005, "Regularization and variable selection via the elastic net", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**, 2, 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.

Applied Biochemistry and Biotechnology
https://doi.org/10.1007/s12010-023-04741-1

ORIGINAL ARTICLE

Potato Peel Waste as an Economic Feedstock for PHA Production by *Bacillus circulans*

Sonika Kag[1] · Pravir Kumar[1] · Rashmi Kataria[1]

Accepted: 15 September 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract
Polymers of hydroxy alkanoates (PHA), also known as biodegradable, biocompatible plastic, are potential alternatives to petrochemical-based plastics. PHA is synthesized by microbes in their cytoplasm in the form of inclusion bodies in stress conditions such as nitrogen, oxygen, and phosphorus with excessive amounts of carbon. Sugar extracted from potato peel in the form of hydrolysate was employed as a carbon source for PHA production after acidic hydrolysis. The acid hydrolysis conditions are optimized for dilute acid concentrations and temperatures. The highest sugar-yielding condition (2% 15 min at 121 ℃) was used for submerged fermentation for PHA production by *Bacillus circulans* MTCC 8167. Fourier transform infrared spectroscopy, nuclear magnetic resonance, and differential scanning calorimetry were used for polymer characterization. Gas chromatography coupled with mass spectrometry confirmed the monomers such as hexadecenoic acid 3-hydroxy, methyl esters, pentadecanoic acid 14 methyl esters, and tetradecanoic acid 12-methyl esters. Crotonic acid assay was used for quantification of PHA and it was found highest (0.232 ± 0.04 g/L) at 37 °C and 36 h of incubation. Hence, potato peel waste could be a potential feedstock for waste to valuable production.

Keywords Cell dry weight · Fermentation · Bioplastics · Potato peel hydrolysate · Reducing sugars

Introduction

Plastics play an indispensable role in our day-to-day lives due to their quality and diverse applications. The increasing human population has led to the deposition of commercial plastics in the form of microplastics, which have adversely affected the marine environment. Consequently, many countries are now focusing on sustainable alternative materials with properties similar to those of petroleum-based synthetic plastics [20]. Polyhydroxybutyrate (PHB) is a homopolymer and one of the most extensively studied types of polyhydroxyalkanoates (PHAs). These biopolymers are synthesized as intracellular inclusion

✉ Rashmi Kataria
rashmikataria@gmail.com

[1] Department of Biotechnology, Delhi Technological University (DTU), Shahbad Daulatpur Village, Bawana Road, Delhi 110042, India

bodies by various bacterial species under adverse conditions and exhibit similarities to commercial plastics [50].

Based on their monomeric subunits, PHAs can be classified into two categories: short-chain-length (SCL) PHAs, which contain 3 to 5 carbons, and medium-chain-length (MCL) PHAs, with structures containing 6 to 14 carbons [11]. Structurally, PHAs can be found in the form of homopolymers (P(HB)) or heteropolymers (P(HB-CO-HV)). These biopolymers have various industrial and biomedical applications, including drug delivery carriers, dental implants, and the synthesis of biodegradable bioplastics [39].

However, the application of these biodegradable plastics has been challenging at an industrial level due to the high production costs [39]. This challenge can be mitigated by using cost-effective substrates for microorganisms that accumulate PHAs [37]. Sugar-rich wastes from different sources, such as wheat straw, rice straw, corn straw, potato peels, whey protein, and glycerol, are gaining attention for renewable production [16]. Wheat, rice, maize, and potatoes are among the most consumed crops worldwide, leading to substantial peel waste generation from both industries and households. Potato peel waste, for instance, contains 30–40% glucose in the form of starch, which can contribute to environmental pollution if disposed of in open areas [24].

Starch hydrolysis yields monomeric hexose sugar in the form of glucose, which serves as the preferred carbon source for energy generation within bacterial cells.

Starch hydrolysis can be achieved through acid hydrolysis or by using amylase and glucoamylase enzymes [6]. While enzymatic hydrolysis is a preferable and rapid method for generating reducing sugars, the use of commercial enzymes may not be cost-effective for large-scale processes [27]. An alternative option for glucose generation is the use of dilute acid, which conserves both energy and time. This method can be particularly useful for various industrially important chemicals derived from potato peel waste, such as organic acids [29], biofuels [2, 27], carbohydrate polymers [1], and bioplastics [8]. Several studies are underway to enhance product yield using different approaches, including advanced pretreatment strategies and simultaneous saccharification and fermentation [41]. However, microbial production of PHA from potato peel waste remains relatively unexplored.

In the present work, potato peel waste was subjected to hydrolysis using dilute acid, and the extracted sugar was utilized for bacterial polymer production. The production medium for the bacterial strain *Bacillus circulans* MTCC 8167 was prepared by adding sugar hydrolysate, 0.5% NaCl, and 0.1% tryptone. PHA was detected and characterized through a crotonic acid assay, GC–MS, FTIR, NMR, and DSC.

Material and Methods

Potato Peel Collection and Primarily Processing

A potato peel sample was collected from the Delhi Technological University (DTU) canteen and hostels. Processing of the sample was done by washing followed by drying at 40 ℃ until constant weight. The dried sample was grinded to reduce particle size and sieved through a 1-mm mesh.

Bacterial Culture

An isolated bacterial culture, *Bacillus circulans* MTCC 8167, was procured from the Microbial Type Culture Collection (MTCC) in Chandigarh, India. Culture was maintained on Luria Bertani (LB) media containing, yeast extract (5 g/L), peptone (10 g/L), NaCl (10 g/L), and agar (15 g/L), the pH was adjusted to 7. The culture media after inoculation were incubated at 37 ℃ with 180 rpm for the growth studies.

Composition Analysis and Acid Hydrolysis of Potato Peel Waste

One gram of dried powder of potato peel was subjected to composition analysis in a triplicate. Starch analysis was done using the spectrophotometric method [28]; protein analysis was done by multiplying plant factor 6.25 with obtained nitrogen content after three-step procedure digestion, distillation, and titration as given in literature [45]; and lignin analysis was done using the protocol established by National Renewable Energy Laboratory [44]. Structural carbohydrate such as cellulose and hemicellulose analysis was done gravimetrically using chlorite method [51]. In which holocellulose was determined with $NaClO_2$ treatment, cellulose extraction from holocellulose was done using 17.5% NaOH treatment and hemicellulose was estimated by subtracting cellulose from holocellulose.

Potato peel biomass is composed of starch, cellulose, lignin, protein, etc. [7, 13]. To release fermentable sugar from polysaccharides, a proper treatment strategy is required before subjecting the biomass to microbial fermentation [41]. Dilute acid treatment at high temperatures is done to convert the potato peel powder into reducing sugar. For chemical hydrolysis, 1 L of 3.2 M HCl stock solution was prepared by mixing 275.86 mL of 10% HCl in 724.13 mL of distilled water. Different hydrochloric acid (HCl) concentrations (0.5, 0.75, 1, 1.25, 1.5, and 2%) were prepared from stock solution and pretreatment was done at temperatures (50, 100, and 121 ℃). Ten percent solid loading (10 g potato peel in 100 mL of aqueous phase) was done in each combination. Hydrolysate obtained after hydrolysis was cooled at room temperature and subjected to detoxification. In the process of detoxification, dry calcium carbonate was added to acidic hydrolysate, and the pH of the sample was monitored throughout the time to reach 6.8–7 [3]. After neutralization, centrifugation was done to get clear hydrolysate and followed by filtration to obtain the sugar-rich extract. Sugar analysis was performed by the DNS method [35].

Analytical Methods

Sugar Analysis, PHA Production, Quantification, and Characterization

Sugar analysis of liquid hydrolysate was done by di-nitro salicylic acid given by Miller [35] briefly, 1 mL of DNS reagent was added to 1 mL pretreated liquid hydrolysate and heated at 100 ℃ for 10 min, and reducing sugar concentration was determined by UV visible spectrophotometer (GENESYS 50) at 540 nm.

After optimization of acid hydrolysis conditions, the reducing sugar (10 g/L), from the highest yielding condition (2% HCl at autoclave condition), was applied as a carbon source for the cultivation of *Bacillus circulans.* Hydrolysate was filtered through a 0.22-μm pore-sized syringe filter diameter of 25 mm made up of Avantor material limited. Sodium chloride (5 g/L) and organic nitrogen source tryptone (1 g/L) were added after autoclaving in sterilized hydrolysate. Bacterial culture was prepared by inoculation of a single colony of bacteria in a culture

medium in a 250-mL flask at 37 °C in a shaking incubator at 180 rpm. The culture was taken from the mid-exponential phase for 1% inoculum for the hydrolysate fermentation. The growth pattern of *Bacillus circulans* was observed in terms of cell dry weight, PHA production, and sugar consumption at different time intervals such as 24, 36, and 48 h [15]. Residual sugar was analyzed at 540 nm using a thermo-scientific UV visible spectrophotometer (GENESYS 50).

After fermentation, the culture suspensions were subjected to centrifugation (microprocessor-based laboratory centrifuge) at 6000 rpm for 15 min at room temperature. Cell pellets were washed with deionized water and then dried in the oven (Matrix Scientific) at 55 ℃ until the constant weight of cell dry weight in each condition was recorded. The experiment was done in a triplicate manner. The extraction of PHA was done with 1:20 ratios of sample and chloroform by heating at 60 ℃ for 120 min in a water bath. After cooling the extractives, precipitation was done with a ten-fold volume of ice-chilled methanol as described by Shah [43].

The polymer obtained from the above method was estimated by the crotonic acid method. PHA is heated with H_2so_4, it is converted into crotonic acid, and its concentration is determined by spectrophotometric assay. The stock solution of crotonic acid was prepared by dissolving 0.1 mg of crotonic acid into 1 mL of sulfuric acid. Different concentrations of 2, 5, 7, 8, 9, and 10 µg/mL for crotonic acid were prepared from the stock solution with a working volume of 1 mL. Sulfuric acid was used as blank, and absorbance was recorded at 235 nm. The standard graph was plotted for concentration vs. absorbance. From the standard curve concentration of the test sample (PHA) was determined [23].

Biopolymer (PHA) was analyzed by acidic methanolysis. The 20 mg dried biomass was subjected to methanolysis by the addition of 2 mL $CHCl_3$, 1.7 mL CH_3OH, and 0.3 mL H_2SO_4. The mixture was heated at 100 °C for 2 h and 20 min of duration. A 2-µL sample from the bottom organic layer was used for GCMS analysis (Shimadzu QP 2010) and the injection temperature of the column was maintained at 260 °C. Monomers were analyzed by the NIST library as previously done by Mahato et al. [31].

The functional group of PHA from *Bacillus circulans* was analyzed by FT-IR with a spectral range of 4000–400 cm^{-1} (Perkin Elmer (Frontier Shelton CT08484)) as described in the literature [32]. The thermal degradation behavior of extracted PHA powder was checked using a differential scanning calorimeter (DSC) with an instrument make of Perkin Elmer model 8000**.** Nitrogen gas was used for purging with a flow rate of 10 mL/min. 10 mg PHA was subjected to a thermal profile from – 10 to 300 °C and the heating rate was 10 °C/min [19]. To check chemical properties ^{1}H NMR was performed. Briefly, 6 mg of pure PHA was dissolved in 1 mL $CDCl_3$ (denatured chloroform) and spectra were recorded on 500 MHz Bruker spectrophotometer [32].

Result and Discussion

Composition Analysis

The results of the composition analysis of potato peel waste are shown in Table 1. It was observed that a high amount of starch (63.34%), followed by protein (14%), lignin (7.12%), and nitrogen (2.27%), is present in potato peel waste. In a study by Remedios and Domingues [41], 46% starch, 10% protein, ash 7%, and lignin 4% were found in potato peel samples. In a literature Lima et al. [30] reported 32.4% cellulose and 10% hemicellulose content from potato peel. However, the composition of potato peel varies in different varieties of potatoes and peeling methods [42].

Table 1 Composition of potato peel waste

Parameters	% (w/w)
Protein	14 ± 0.40
lignin	7.12 ± 0.22
Starch	63.34 ± 0.20
Cellulose	9.42 ± 0.22
Hemicellulose	2.18 ± 0.22
Others	3.94 ± 0.35

Acid Hydrolysis

As a result of acid hydrolysis, the complexity of peel biomass decreases, and the amount of fermentable sugar increases. Different concentrations of dilute acid (HCl) and time variation for hydrolysate development were applied in this study which resulted in a varied amount of reducing sugar. Based on the sugar yields comparative results are shown in Fig. 1 and Tables 2 and 3. In the acid hydrolysis process, a water molecule is added through a nucleophilic reaction due to the breakdown of chemical bonds of starch, and a change in porosity occurs which leads to the release of sugar upon hydrolysis [10]. Experimental conditions used in this study such as 50 ℃ and 100 ℃ at different acid ranges were found insufficient for maximum reducing sugar extraction because sugar yield is directly proportional to hydrolysis temperature at a certain limit [46]. In an autoclave (at 121 °C) with 2% acid concentration, 71.23 ± 0.46% of sugar was extracted, which is the highest in comparison to other hydrolysis conditions (0.5, 0.75, 1, 1.25, and 1.5% HCl concentration). The reason behind this observation is that soft biomass liberates high sugar at mild acidic conditions and high temperatures which are previously observed in the literature [47]. Similarly,

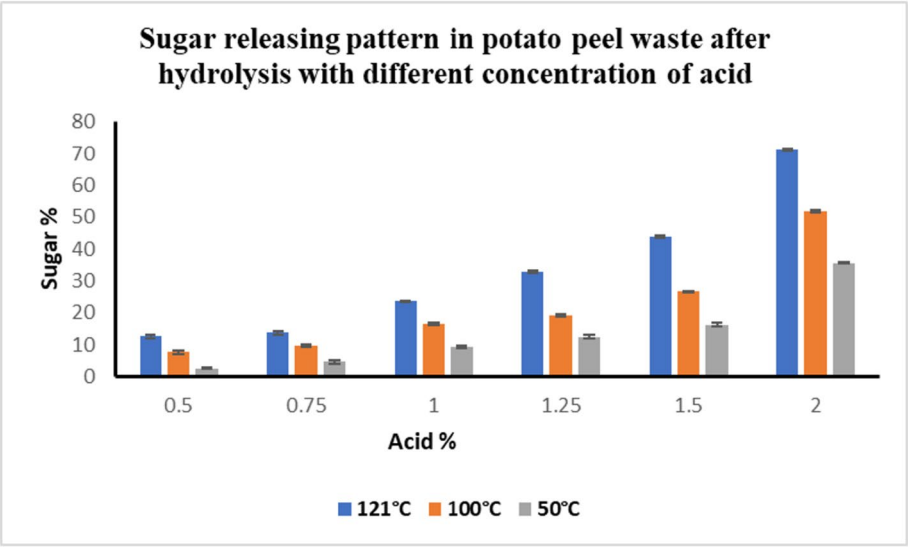


Fig. 1 Sugar releasing pattern in potato peel waste after hydrolysis with different concentrations of acid

Table 2 Reducing sugar yield (%) after acid hydrolysis

Acid %	Sugar % at 121 °C	Sugar % at 100 °C	Sugar % at 50 °C
0.5	12.47 ± 0.40	7.61 ± 0.52	2.58 ± 0.15
0.75	13.71 ± 0.54	9.69 ± 0.30	4.55 ± 0.46
1	23.58 ± 0.17	16.44 ± 0.29	9.28 ± 0.51
1.25	32.77 ± 0.36	19.06 ± 0.38	12.41 ± 0.53
1.5	43.66 ± 0.34	26.59 ± 0.14	16.22 ± 0.46
2	71.23 ± 0.46	51.84 ± 0.29	35.50 ± 0.20

in a report, 23 g/L of sugars using 1% sulfuric acid pretreatment in autoclave condition was obtained [33]. In another study [18], 1–5% HCl was used to hydrolyze potato peel and they reported 62 g/L glucose from 5% acid concentration at 90 ℃ for 120 min reaction time. Martín et al. [34] reported almost similar amount (72%) conversion of glucose from cassava stems where in the first step, cassava stems are subjected to acid pretreatment with 0.6% H_2SO_4 and then enzymatic hydrolysis was with a combination of amylase and glucoamylase. Though high temperatures of more than 121 ℃ and increased acid concentration lead to the accumulation of toxic material such as phenolics due to the degradation of sugars [46]. Though the goal of this study was to extract maximum sugar for PHA production and therefore 2% acid was used which leads to liberating 71.23 ± 0.46% reducing sugar.

PHA Production by Bacillus circulans MTCC 867 in Potato Peel Media

The PHA production by *Bacillus circulans* in potato peel hydrolysate-containing media was observed in three different time intervals (24, 36, and 48 h). The physical growth parameters play a vital role in cell dry weight accumulation and PHA production. At 24 h of incubation, biomass was minimum (0.926 ± 0.01), as actively dividing cells utilize sugar only for growth purposes. Cell biomass was found at maximum (1.36 ± 0.02) up to 36 h as secondary metabolites (PHA) produced during the late stationary phase of the growth cycle when nutrients are in the limit. The stationary phase is signal when PHA synthesis enzymes such as 3-keto thiolase, acetoacetyl-CoA reductase, and PHA synthase start converting PHA from acetyl-CoA which is commonly knowns as inclusion bodies [12]. A depletion of cell biomass (1.09 ± 0.06) until 48 h was observed as storage granules are utilized as a source of energy by the bacteria [32].

Table 4 and Fig. 2 show representation of cell dry weight and PHA accumulation pattern in different time intervals. All three-production media (24, 36, and 48 h) are further subjected to extraction by solvent extraction method. Different studies have shown the PHA accumulation

Table 3 Reducing sugar yield (mg/gram of potato peel waste) after acid hydrolysis

Acid %	Sugar mg/g of potato peel at 121 °C	Sugar mg/g of potato peel at 100 °C	Sugar mg/g of potato peel at 50 °C
0.5	125.55 ± 0.31	76.11 ± 0.22	25.84 ± 0.16
0.75	137.15 ± 0.18	96.94 ± 0.13	45.57 ± 0.40
1	235.84 ± 0.2	164.48 ± 0.29	92.89 ± 0.15
1.25	327.77 ± 0.23	190.64 ± 0.17	124.11 ± 0.21
1.5	436.61 ± 0.11	265.97 ± 0.30	162.28 ± 0.35
2	712.33 ± 0.30	518.46 ± 0.22	355.01 ± 0.20

Table 4 Cell dry weight (g/L) at different time intervals in potato peel hydrolysate media

Time (h)	Cell dry weight (g/L)	PHA g/L	Residual sugar (g/L)
24	0.926 ± 0.01	0.1 ± 0.031	1.38 ± 0.05
36	1.36 ± 0.02	0.232 ± 0.041	0.8 ± 0.21
48	1.09 ± 0.06	0.184 ± 0.015	0.11 ± 0.14

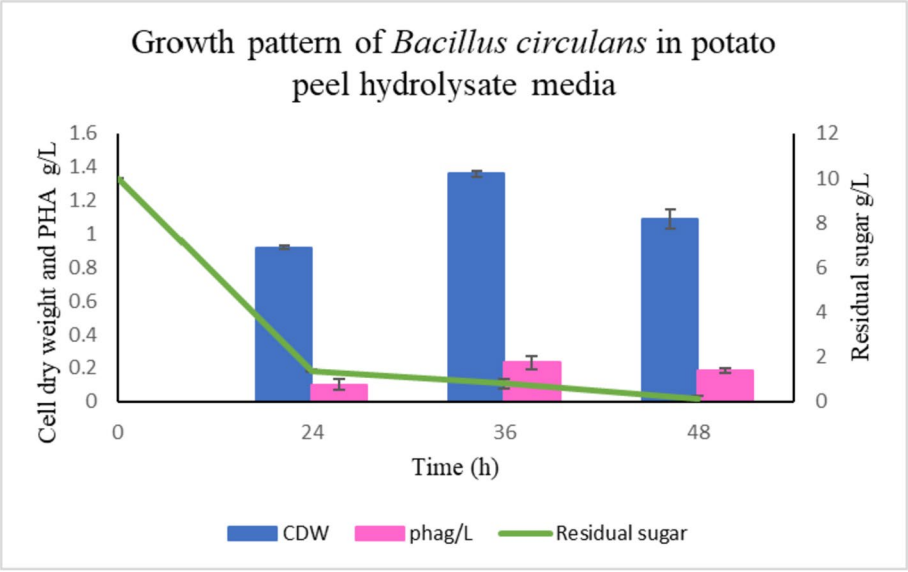


Fig. 2 Growth pattern of *Bacillus circulans* in potato peel hydrolysate media

ability of a variety of *Bacillus species* by utilizing different organic wastes as carbon sources [5, 23, 49]. Subsequently, these carbon sources were applied in discrete combinations of other nutrient components, which showed the high PHA accumulation in the bacterial cytoplasm. For instance, in an optimization study [40], high yield of PHA (46.57%) was observed in *Bacillus endophyticus*; however, bacteria were cultivated in media containing salts such as Na_2HPO_4, KH_2PO_4, and sucrose 40 g/L was used as a carbon source. As several studies observed, PHA enhancement by optimizing the desirable C:N ratio [4, 48, 52] or by use of genetically engineered microbes [21]. But in the present study, wild-type *Bacillus circulans* was taken and carbon was 10 g/L used from waste potato peels, and other growth-promoting micronutrients were not added in production media. A comparison of PHA production by different strains of *Bacillus* based on different parameters including polymer type, yield, and utilized carbon source is shown in Table 5. From the comparative studies, explained in above-mentioned table, it is observed that potato peel waste can be a good carbon source for an adequate amount of PHA production by *Bacillus circulans*.

Table 5 PHA production by different species of *Bacillus*

Bacteria	Type of polymer	Carbon source	Yield (%)	Reference
Bacillus megaterium	PHB	DSMZ media containing glycerol as a carbon source	14.11	Hiremath and Sura [23]
Bacillus cereus	PHB	Grape residue	18.79	Andler et al. [5]
Bacillus sp.	PHB	Mineral salt media with 2% glucose	20	Hassan et al. [22]
Bacillus megaterium	PHA	Food waste-derived volatile fatty acid	10	Vu et al. [49]
Bacillus circulans	PHA	Potato peel waste	23	Present study

GC–MS Analysis

By GC–MS analysis medium chain length, PHA monomers such as hexa decanoic acid 3-hydroxy, methyl ester, pentadecanoic acid 14- methyl -esters, and tetra decanoic acid, 12- methyl esters were identified with area % 4.25, 1.94, and 7.43%, respectively, using potato peel waste as a carbon source. However, scl PHA was reported by *B. cereus* SS105 [32]. Moreover Choonut et al. [14] reported five different types of mcl PHA monomers reported by *Bacillus theroaylovorans*-related strains. The type of monomer accumulation inside the bacterial cell depends on the type of carbon source [9].

Quantification of PHA

Cell dry weight (1 mg) was transformed into crotonic acid by adding 10 mL H_2SO_4 and absorbance was recorded at 235 nm against H_2SO_4 blank. Obtained PHA was found highest at 36 h of incubation 0.232 ± 0.04 g/1000 mL of production media. However, at 24 h and 48 h, it was found 0.1 ± 0.041 and 0.184 ± 0.015 g/1000 mL of media, respectively. In a study 0.5 g/L of PHA was extracted from the *Bacillus marcorestinctum* and mineral salt media containing NH_2PO_4, KH_2PO_4, $MgSO_4$,$(NH_4)_2SO_4$, $CaCl_2$, NaCl, NH_4Fe, citrate, mixture of trace elements, 2% cane molasses along with 10 g/L glucose were used for PHA production [36]. In literature Choonut et al. [14] reported 0.41 ± 0.01 g/L of PHA by thermotolerant bacteria *Bacillus thermoamylovorans*, isolated from palm mill effluent where the MS media (mineral salt) enriched with $Na_2HPO_4.12H_2O$ (9 g/L), KH_2PO_4 (1.5 g/L), nitrogen stress in the form of 0.1 g/L NH_4Cl, and $MgSO.7H_2O$ were taken as the culture media. However, in the present study, 1% sugar from potato peel waste, NaCl, and 0.1 g/L tryptone were used in production media which is very economical as compared to above-mentioned study.

Characterization of PHA

FTIR

Characterization of extracted PHA from *Bacillus circulans* was done with FTIR analysis. The distribution of functional groups present in FTIR spectra was performed based on existing literature [31, 43]. PHA shows the band at 3437.13 cm^{-1} represents the hydroxyl group, the band at 2925.31 cm^{-1} represents C–H which shows the presence of methane and methylene groups, and sharp band at position 1724.25 cm^{-1} indicates

the presence of ester bond (C = O) which is a characteristic feature of polyhydroxyalkanoates [25]. The 1460.02 cm^{-1} represents CH2, band position at 1280.21 corresponds to C–O, and 1184.10 cm^{-1} corresponds to C–O–C (Fig. 3). Similar pattern of FTIR spectroscopy was obtained by *Bacillus megaterium BBST*$_4$ in a study done by Porras et al. [38]. Almost similar pattern of FTIR peaks was found in PHA accumulated by *Bacillus cereus SS105* in which the spectral peak at 3264 cm^{-1} represents the O–H group of polymer; for CH_3 group peak, found at 2973 cm^{-1}, an outcome of changes of crystalline structure and 1736 cm^{-1} band was observed due to the presence of ester bond [32]. Hence, it is confirmed that the deposited polymer inside *Bacillus circulans* MTCC 8167 is PHA.

DSC

A differential scanning calorimetry technique is done to check the melting temperature (T_m) of extracted PHA. The peak obtained in the DSC curve at 165.24 °C indicates the crystalization of the PHA polymer (Fig. 4). Thermal characterization of PHA provides detail about the self-life of PHA and its stability. DSC of standard PHA ranges between 165 and 170 °C. Polymer degradation at low temperatures may be because of impurities. Although, the rate of degradation of polymer is species specific and may change upon using different extraction methods [25]. Cueva-Almendras [17] checked the DSC curve of PHA from isolated *Bacillus thuringiensis* and they found T_m at 166.92 °C close to the curve observed in

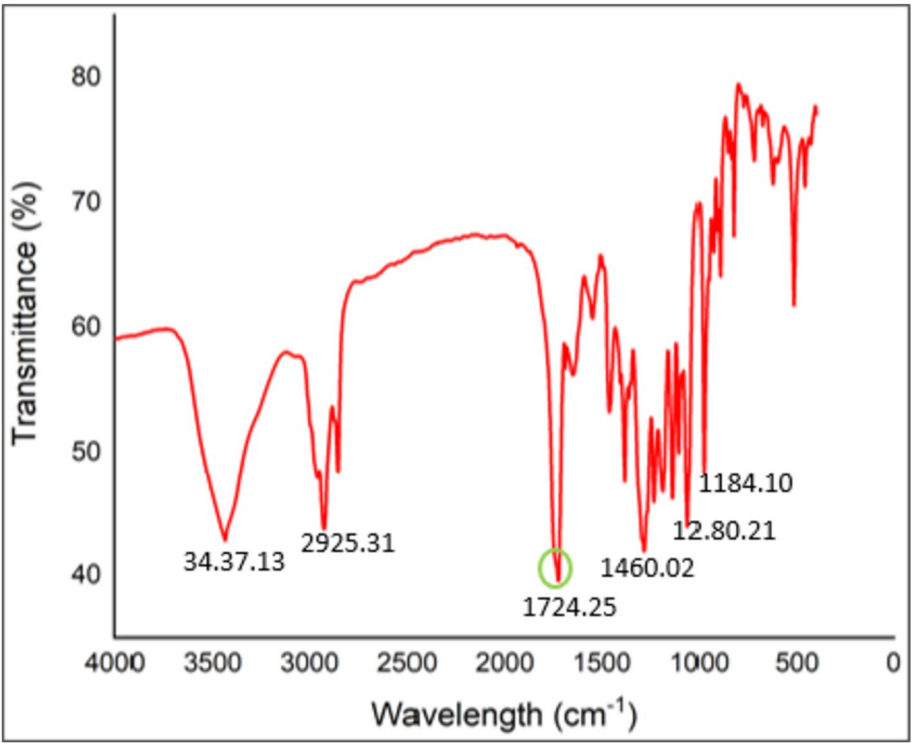


Fig. 3 FTIR spectra of purified PHA polymer

this study. Vu et al. [49] reported 145.43 °C T_m, which is lower than the present report from polymer accumulated by *Bacillus megaterium* using volatile fatty acid as a carbon source. Hassan et al. [22] observed the Tm at 175.9 °C of a polymer obtained from the *Bacillus sp*. which is higher from the present study.

NMR

The presence of methyl and methylene groups is characteristic feature of PHA monomer. The intensity of signals obtained from proton NMR spectra of extracted biopolymer indicated the presence of PHA polymer. The doublet peak at 1.37 is an indication of the presence of the CH_3 group, while the doublet quadruplet at the signal at 2.5 ppm is a characteristic feature of the presence of the CH_2 group in the sample; moreover, peak intensity between 5.28 and 5.3 ppm is an indication of CH group. The peak at 7.2 ppm is attributed to solvents (denatured chloroform). By comparing the peak intensities with previous work, a similar pattern of proton NMR spectra was reported by Raghu and Divyashree [40] in PHA accumulated inside *Bacillus flexus* when it is grown on castor oil as a carbon source. Ensifer [26] identified the same type of peak pattern of proton NMR of PHA polymer accumulated by a soil isolate *Ensifer sp.* Moreover, the pattern of NMR obtained by the studies [22, 45] confirms the polymer obtained from *Bacillus circulans* using potato peel waste in this study is PHA (Fig. 5).

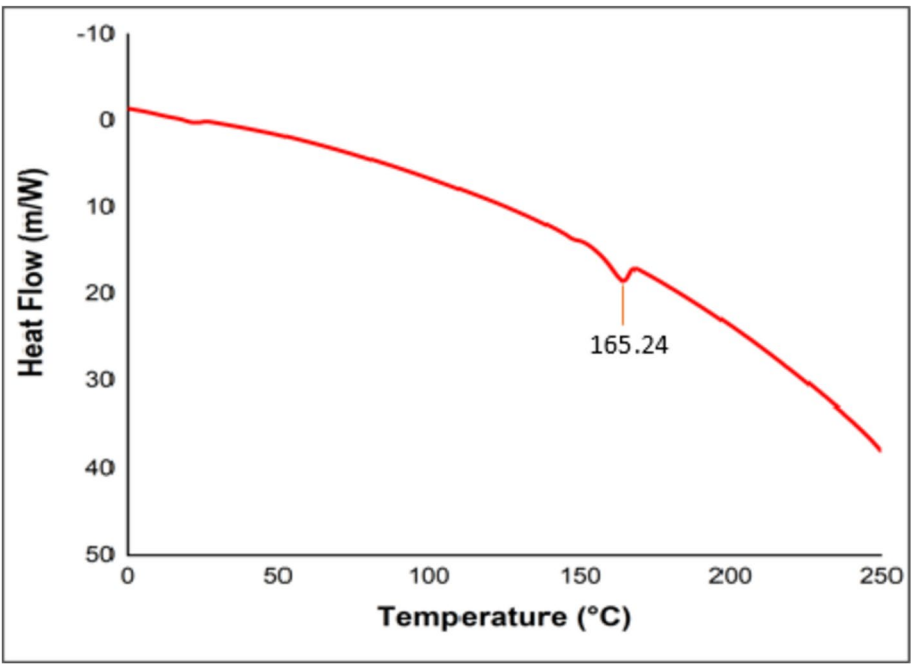


Fig. 4 DSC thermogram of purified PHA polymer

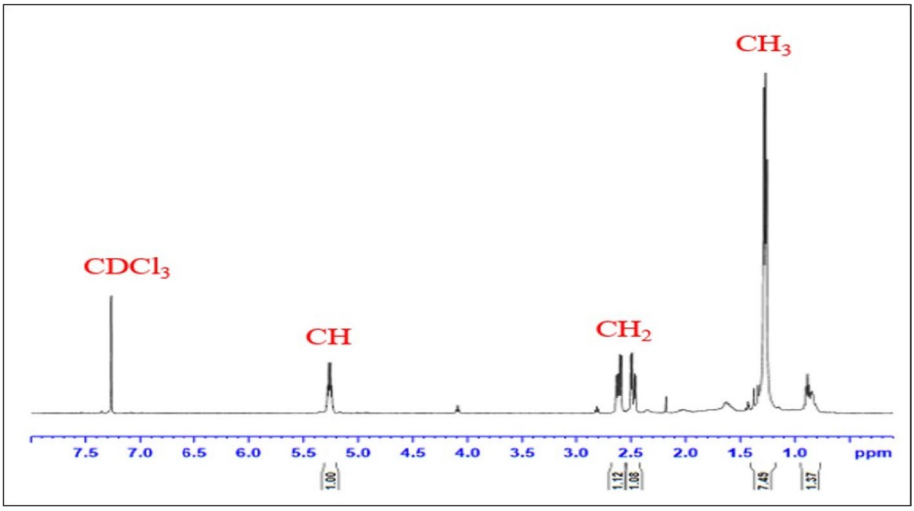


Fig. 5 NMR spectroscopy of purified PHA polymer

Conclusion and Future Prospective

PHA is a promising biopolymer for replacing traditional plastics. The extent of PHA deposition by microbes depends on several factors, including the microbial species' potency and its ability to utilize available carbon sources. Potato peel waste is utilized to liberate sugars, and the resulting reducing sugars serve as a carbon source for PHA accumulation by *Bacillus circulans*. During a 36-h incubation period, a substantial amount of PHA (0.232 ± 0.04 g/L) was extracted, and its characteristics were further analyzed using FTIR, DSC, and NMR. While achieving a desirable amount of CDW (cell dry weight) remains a challenge, this obstacle can potentially be overcome by optimizing growth parameters, such as the ideal quantity of carbon and nitrogen, to achieve a higher dry cell weight. Hence, potato peel waste shows promise as a cost-effective alternative carbon source for PHA production.

Acknowledgements The authors would like to thanks to the funding agencies Department of Biotechnology (DBT) grant no. BT/RLF/Re-entry/40/2017 and Science and Engineering Research Board (SERB) file no. EEQ/2020/000614, Govt. of India for providing the research grant to carry out this work. Also, special thanks to Council of Scientific & Industrial Research (CSIR), Govt. of India, for fellowship of Ms. Sonika. Authors would also like to thanks Department of Chemistry, Delhi Technological University (DTU) for the support and providing testing facilities.

Author Contribution All authors read and approved the final manuscript, also confirm the contribution to the review article as follows: Experiment design and conceptualization of article: Sonika Kag and Rashmi Kataria; laboratory experiments and analysis: Sonika Kag; manuscript writing: Sonika Kag; Critically revised the work: Rashmi Kataria, Pravir Kumar, and Sonika Kag; Editing: Rashmi Kataria; Supervision: Pravir Kumar and Rashmi Kataria.

Funding Present work is funded by DBT (SAN No. 102/IFD/SAN/1276/2019-20) and SERB (File No. EEQ/2020/000614), Govt. of India.

Data Availability Not applicable.

Declarations

Ethical Approval Not applicable.

Consent to Participate All authors are agreed to participate in the present manuscript.

Consent for Publication All authors have seen submitted version of manuscript and approved to publish.

Competing Interests The authors declare no competing interests.

References

1. Abdelraof, M., Hasanin, M. S., & El-Saied, H. (2019). Ecofriendly green conversion of potato peel wastes to high productivity bacterial cellulose. *Carbohydrate Polymers, 211*, 75–83. https://doi.org/10.1016/j.carbpol.2019.01.095
2. Abedini, A., Amiri, H., & Karimi, K. (2020). Efficient biobutanol production from potato peel wastes by separate and simultaneous inhibitors removal and pretreatment, Renewable Energy. Elsevier Ltd. https://doi.org/10.1016/j.renene.2020.06.112
3. Ahmed, F., Yan, Z., & Bao, J. (2019). Dry biodetoxification of acid pretreated wheat straw for cellulosic ethanol fermentation. *Bioresources and Bioprocessing*, 6. https://doi.org/10.1186/s40643-019-0260-x
4. Ahn, J., Jho, E. H., & Nam, K. (2015). Cupriavidus necator and its implication on the use of rice straw hydrolysates Effect of C / N ratio on polyhydroxyalkanoates (PHA) accumulation by Cupriavidus necator and its implication on the use of rice straw hydrolysates. https://doi.org/10.4491/eer.2015.055
5. Andler, R., Pino, V., Moya, F., Soto, E., Valdés, C., & Andreeßen, C. (2021). Synthesis of poly-3-hydroxybutyrate (PHB) by Bacillus cereus using grape residues as sole carbon source. *International Journal of Biobased Plastics, 3*, 98–111. https://doi.org/10.1080/24759651.2021.1882049
6. Arapoglou, D., Varzakas, T., Vlyssides, A., & Israilides, C. (2010). Ethanol production from potato peel waste (PPW). *Waste Management, 30*, 1898–1902. https://doi.org/10.1016/j.wasman.2010.04.017
7. Ben Atitallah, I., Antonopoulou, G., Ntaikou, I., Alexandropoulou, M., Nasri, M., Mechichi, T., & Lyberatos, G. (2019). On the evaluation of different saccharification schemes for enhanced bioethanol production from potato peels waste via a newly isolated yeast strain of Wickerhamomyces anomalus. *Bioresource Technology, 289*, 121614. https://doi.org/10.1016/j.biortech.2019.121614
8. BezirhanArikan, E., & Bilgen, H. D. (2019). Production of bioplastic from potato peel waste and investigation of its biodegradability. *International Advanced Researches and Engineering Journal, 03*, 93–97. https://doi.org/10.35860/iarej.420633
9. Catherine, M., Guwy, A., & Massanet-nicolau, J. (2022). Bioresource technology reports effect of acetate concentration, temperature, pH and nutrient concentration on polyhydroxyalkanoates (PHA) production by glycogen accumulating organisms. *Bioresource Technology Reports, 20*, 101226. https://doi.org/10.1016/j.biteb.2022.101226
10. Celikci, N., Dolaz, M., &Colakoglu, A. S. (2020). An environmentally friendly carton adhesive from acidic hydrolysis of waste potato starch. *International Journal of Polymer Analysis and Characterization*, 0, 1–16. https://doi.org/10.1080/1023666X.2020.1855047
11. Chanasit, W., Hodgson, B., Sudesh, K., & Umsakul, K. (2016). Efficient production of polyhydroxyalkanoates (PHAs) from Pseudomonas mendocina PSU using a biodiesel liquid waste (BLW) as the sole carbon source. *Bioscience, Biotechnology, and Biochemistry, 80*, 1440–1450. https://doi.org/10.1080/09168451.2016.1158628
12. Chen, Y. J., Huang, Y. C., & Lee, C. Y. (2014). Production and characterization of medium-chain-length polyhydroxyalkanoates by Pseudomonas mosselii TO7. *Journal of Bioscience and Bioengineering, 118*, 145–152. https://doi.org/10.1016/j.jbiosc.2014.01.012
13. Chintagunta, A. D., Jacob, S., & Banerjee, R. (2016). Integrated bioethanol and biomanure production from potato waste. *Waste Management, 49*, 320–325. https://doi.org/10.1016/j.wasman.2015.08.010
14. Choonut, A., Prasertsan, P., Klomklao, S., & Sangkharak, K. (2020). Bacillus thermoamylovorans - related strain isolated from high temperature sites as potential producers of medium - chain - length polyhydroxyalkanoate (mcl - PHA). *Current Microbiology, 77*, 3044–3056. https://doi.org/10.1007/s00284-020-02118-9

15. Ciesielska, J. M., Dabrowska, D., Palasz, A. S., & Ciesielski, S. (2017). Medium - chain - length polyhydroxyalkanoates synthesis by Pseudomonas putida KT2440 relA / spoT mutant : Bioprocess characterization and transcriptome analysis. *AMB Express*. https://doi.org/10.1186/s13568-017-0396-z
16. Colombo, B., Favini, F., Scaglia, B., Sciarria, T. P., Imporzano, G. D., Pognani, M., Alekseeva, A., Eisele, G., Cosentino, C., & Adani, F. (2017). Biotechnology for biofuels enhanced polyhydroxyalkanoate (PHA) production from the organic fraction of municipal solid waste by using mixed microbial culture. *Biotechnology for Biofuels and Bioproducts*, 1–15. https://doi.org/10.1186/s13068-017-0888-8
17. Cueva-almendras, L. C. (2022). Production of polyhydroxyalkanoate by Bacillus thuringiensis Isolated from agricultural soils of Cascas-Peru. *Brazilian Archives of Biology and Technology, 65*, e22220107. https://doi.org/10.1590/1678-4324-2022220107
18. Deshmukh, M., & Pande, A. (2022). Comparative Study for Production of biofuel from potato peel waste as feedstock by different enzymes. *11*, 1–6. https://doi.org/10.35841/2329-6674.22.11.1000175
19. Dinh, P., Minh, L., Trang, V., Minh, H., Thi, L., Phung, K., & Feng, D. (2022). Enrichment of hydrogen in product gas from a pilot-scale rice husk updraft gasification system. *Carbon Resources Conversion, 5*, 231–239. https://doi.org/10.1016/j.crcon.2022.07.003
20. Evangeline, S., & Sridharan, T. B. (2019). Biosynthesis and statistical optimization of polyhydroxyalkanoate (PHA) produced by Bacillus cereus VIT-SSR1 and fabrication of biopolymer fi lms for sustained drug release. *International Journal of Biological Macromolecules, 135*, 945–958. https://doi.org/10.1016/j.ijbiomac.2019.05.163
21. Gao, C., Qi, Q., Madzak, C., Sze, C., & Lin, K. (2015). Exploring medium - chain - length polyhydroxyalkanoates production in the engineered yeast Yarrowia lipolytica. *Journal of Industrial Microbiology and Biotechnology, 42*, 1255–1262. https://doi.org/10.1007/s10295-015-1649-y
22. Hassan, M. A., Bakhiet, E. K., Ali, S. G., & Hussien, H. R. (2016). Production and characterization of polyhydroxybutyrate (PHB) produced by Bacillus sp. isolated from Egypt. *Journal of Applied Pharmaceutical Science, 6*, 46–51. https://doi.org/10.7324/JAPS.2016.60406
23. Hiremath, L., Sura, N., & Angadi, S. (2015). Design, screening and microbial synthesis of biopolymers of Poly-Hydroxy-Butyrate (PHB) from low cost carbons. *International Journal of Advanced Research*, *3*(2), 420–425. http://www.journalijar.com
24. Hong, Z., Fen, X. U., Yu, W. U., Hong-hai, H. U., & Xiao-Feng, D. A. I. (2017). Progress of potato staple food research and industry development in China. *16*, 2924–2932.https://doi.org/10.1016/S2095-3119(17)61736-2
25. Joyline, M. (2019). Research Article Production and characterization of polyhydroxyalkanoates (PHA) by bacillus Megaterium strain JHA using inexpensive agro-industrial wastes Mascarenhas Joyline and Aruna K * *10*, 33359–33374. https://doi.org/10.24327/IJRSR
26. Khamkong, T., Penkhrue, W., Lumyong, S. (2022). Optimization of Production of Polyhydroxyalkanoates (PHAs) from Newly Isolated Ensifer sp. Strain HD34 by Response Surface Methodology. Processes 2022;10, 1632. https://doi.org/10.3390/pr10081632
27. Khawla, B. J., Sameh, M., Imen, G., Donyes, F., Dhouha, G., Raoudha, E. G., & Oumèma, N. E. (2014). Potato peel as feedstock for bioethanol production: A comparison of acidic and enzymatic hydrolysis. *Industrial Crops and Products, 52*, 144–149. https://doi.org/10.1016/j.indcrop.2013.10.025
28. Landhäusser, S. M., Chow, P. S., Turin Dickman, L., Furze, M. E., Kuhlman, I., Schmid, S., Wiesenbauer, J., Wild, B., Gleixner, G., Hartmann, H., Hoch, G., McDowell, N. G., Richardson, A. D., Richter, A., & Adams, H. D. (2018). Standardized protocols and procedures can precisely and accurately quantify non-structural carbohydrates. *Tree Physiology, 38*, 1764–1778. https://doi.org/10.1093/treephys/tpy118
29. Liang, S., McDonald, A. G., & Coats, E. R. (2014). Lactic acid production from potato peel waste by anaerobic sequencing batch fermentation using undefined mixed culture. *Waste Management, 45*, 51–56. https://doi.org/10.1016/j.wasman.2015.02.004
30. Lima, M. de A., Andreou, R., Charalampopoulos, D., & Chatzifragkou, A. (2021). Supercritical carbon dioxide extraction of phenolic compounds from potato (Solanum tuberosum) peels. *Applied Sciences*, 11. https://doi.org/10.3390/app11083410
31. Mahato, R. P., Kumar, S., & Singh, P. (2021). Optimization of growth conditions to produce sustainable polyhydroxyalkanoate bioplastic by pseudomonas aeruginosa EO1. *Frontiers in Microbiology*, 12. https://doi.org/10.3389/fmicb.2021.711588
32. Maheshwari, N., Kumar, M., Thakur, I. S., & Srivastava, S. (2018). Production, process optimization and molecular characterization of polyhydroxyalkanoate (PHA) by CO2 sequestering B. cereus SS105. *Bioresource Technology, 254*, 75–82. https://doi.org/10.1016/j.biortech.2018.01.002
33. Malakar, B., Das, D., & Mohanty, K. (2020). Optimization of glucose yield from potato and sweet lime peel waste through different pre-treatment techniques along with enzyme assisted hydrolysis

towards liquid biofuel. *Renewable Energy, 145*, 2723–2732. https://doi.org/10.1016/j.renene.2019.08.037
34. Martín, C., Christoph, J., Wei, M., Stagge, S., Xiong, S., & Jönsson, L. J. (2019). Industrial crops & products dilute-sulfuric acid pretreatment of de-starched cassava stems for enhancing the enzymatic convertibility and total glucan recovery. *Industrial Crops and Products, 132*, 301–310. https://doi.org/10.1016/j.indcrop.2019.02.037
35. Miller, G. L. (1959). Use of dinitrosalicylic acid reagent for determination of reducing sugar. *Analytical Chemistry, 31*, 426–428. https://doi.org/10.1021/ac60147a030
36. Narayankumar, S., Industries, K., & Krishnaswamy, V. G. (2020). Polyhyrdoxybutyrate production by Bacillus marcorestinctum using polyhyrdoxybutyrate production by Bacillus marcorestinctum using a cheaper substrate and its electrospinned blends with polymer. https://www.researchgate.net/publication/344167030
37. Pan, L., Li, J., Wang, R., Wang, Yu., Lin, Q., Li, C., & Wang, Y. (2021). Biosynthesis of polyhydroxyalkanoate from food waste oil by Pseudomonas alcaligenes with simultaneous energy recovery from fermentation wastewater. *Waste Management, 131*, 268–276. https://doi.org/10.1016/j.wasman.2021.06.008
38. Porras, M. A., Cubitto, M. A., & Villar, M. A. (2014). Quantitative determination of intracellular PHA in Bacillus megaterium BBST4 strain Using Mid FTIR Spectroscopy. 1–4. https://doi.org/10.13140/RG.2.1.3920.2407
39. Prakash, P., Lee, W.-H., Loo, C.-Y., Wong, H. S. J., & Parumasivam T. (2022). Advances in polyhydroxyalkanoate nanocarriers for effective drug delivery: An overview and challenges. *Nanomaterials*, *12*, 175. https://doi.org/10.3390/nano12010175
40. Raghu, M. G. H., & Divyashree, C. M. S. (2021). Statistical optimisation of polyhydroxyalkanoate production in Bacillus endophyticus using sucrose as sole source of carbon. *Archives of Microbiology, 203*, 5993–6005. https://doi.org/10.1007/s00203-021-02554-6
41. Remedios, Y., & Domingues, L. (2023). Potato peels waste as a sustainable source for biotechnological production of biofuels: Process optimization. *155*, 320–328.https://doi.org/10.1016/j.wasman.2022.11.007
42. Sampaio, S. L., Petropoulos, S. A., Alexopoulos, A., Heleno, S. A., Santos-buelga, C., Barros, L., & Ferreira, I. C. F. R. (2020). Trends in Food Science & Technology Potato peels as sources of functional compounds for the food industry : A review. *Trends in Food Science & Technology, 103*, 118–129. https://doi.org/10.1016/j.tifs.2020.07.015
43. Shah, K. R. (2012). FTIR analysis of polyhydroxyalkanoates by novel Bacillus sp. AS 3–2 from soil of Kadi region, North Gujarat, India. *Journal Of Biochemical Technology, 3*, 380–383.
44. Sluiter, J. B., Ruiz, R. O., Scarlata, C. J., Sluiter, A. D., & Templeton, D. W. (2010). Compositional analysis of lignocellulosic feedstocks. Review and description of methods. *Journal of Agricultural and Food Chemistry*, *58*(16), 9043–9053. https://doi.org/10.1021/jf1008023
45. Sriyapai, T., Chuarung, T., Kimbara, K., Samosorn, S., & Sriyapai, P. (2022). Production and optimization of polyhydroxyalkanoates (PHAs) from Paraburkholderia sp. PFN 29 under submerged fermentation. *Electronic Journal of Biotechnology, 56*, 1–11. https://doi.org/10.1016/j.ejbt.2021.12.003
46. Tesfaw, A. A., & Tizazu, B. Z. (2021). Reducingsugarproductionfromteffstrawbiomassusingdilute sulfuric acid hydrolysis: Characterization and optimization using response surface methodology. *International Journal of Biomaterials*, 2021. https://doi.org/10.1155/2021/2857764
47. Timung, R., Naik Deshavath, N., Goud, V. V., & Dasu, V. V. (2016). Effect of subsequent dilute acid and enzymatic hydrolysis on reducing sugar production from sugarcane bagasse and spent citronella biomass. *Journal of Energy, 2016*, 1–12. https://doi.org/10.1155/2016/8506214
48. Valencia, A. I. S., Rojas, U., & Fajardo, C. (2021). Effect of C / N ratio on the PHA accumulation capability of microbial mixed culture fed with leachates from the organic fraction of municipal solid waste (OFMSW). *Journal of Water Process Engineering*, 40. https://doi.org/10.1016/j.jwpe.2021.101975
49. Vu, D. H., Wainaina, S., Taherzadeh, M. J., Åkesson, D., & Ferreira, J. A. (2021). Production of polyhydroxyalkanoates (PHAs) by Bacillus megaterium using food waste acidogenic fermentation-derived volatile fatty acids. *Bioengineered, 12*, 2480–2498. https://doi.org/10.1080/21655979.2021.1935524
50. Wang, J., Liu, S., Huang, J., Cui, R., Xu, Y., & Song, Z. (2023). Environmental Technology & Innovation Genetic engineering strategies for sustainable polyhydroxyalkanoate (PHA) production from carbon-rich wastes. *Environmental Technology and Innovation, 30*, 103069. https://doi.org/10.1016/j.eti.2023.103069
51. Zhou, C., Jiang, W., Via, B. K., Fasina, O., & Han, G. (2015). Prediction of mixed hardwood lignin and carbohydrate content using. *Carbohydrate Polymers, 121*, 336–341. https://doi.org/10.1016/j.carbpol.2014.11.062

52. Zhou, W., Irene, D., Geurkink, B., Euverink, G. W., & Krooneman, J. (2022). Science of the Total Environment The impact of carbon to nitrogen ratios and pH on the microbial prevalence and polyhydroxybutyrate production levels using a mixed microbial starter culture. *Science of the Total Environment, 811*, 152341. https://doi.org/10.1016/j.scitotenv.2021.152341

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

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

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 19 September 2023, accepted 28 September 2023, date of publication 4 October 2023, date of current version 13 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3321861

RESEARCH ARTICLE

Predicting Agriculture Yields Based on Machine Learning Using Regression and Deep Learning

PRIYANKA SHARMA[1], PANKAJ DADHEECH[1], (Member, IEEE), NAGENDER ANEJA[2], AND SANDHYA ANEJA[3], (Member, IEEE)

[1]Department of Computer Science and Engineering, Swami Keshvanand Institute of Technology, Management and Gramothan, Jaipur, Rajasthan 302017, India
[2]School of Digital Science, Univerisiti Brunei Darussalam, Gadong BE1410, Brunei Darussalam
[3]School of Computer Science and Mathematics, Marist College, Poughkeepsie, NY 12601, USA

Corresponding author: Nagender Aneja (nagender.aneja@ubd.edu.bn)

ABSTRACT Agriculture contributes a significant amount to the economy of India due to the dependence on human beings for their survival. The main obstacle to food security is population expansion leading to rising demand for food. Farmers must produce more on the same land to boost the supply. Through crop yield prediction, technology can assist farmers in producing more. This paper's primary goal is to predict crop yield utilizing the variables of rainfall, crop, meteorological conditions, area, production, and yield that have posed a serious threat to the long-term viability of agriculture. Crop yield prediction is a decision-support tool that uses machine learning and deep learning that can be used to make decisions about which crops to produce and what to do in the crop's growing season. It can decide which crops to produce and what to do in the crop's growing season. Regardless of the distracting environment, machine learning and deep learning algorithms are utilized in crop selection to reduce agricultural yield output losses. To estimate the agricultural yield, machine learning techniques: decision tree, random forest, and XGBoost regression; deep learning techniques - convolutional neural network and long-short term memory network have been used. Accuracy, root mean square error, mean square error, mean absolute error, standard deviation, and losses are compared. Other machine learning and deep learning methods fall short compared to the random forest and convolutional neural network. The random forest has a maximum accuracy of 98.96%, mean absolute error of 1.97, root mean square error of 2.45, and standard deviation of 1.23. The convolutional neural network has been evaluated with a minimum loss of 0.00060. Consequently, a model is developed that, compared to other algorithms, predicts the yield quite well. The findings are then analyzed using the root mean square error metric to understand better how the model's errors compare to those of the other methods.

INDEX TERMS Agriculture, crop yield prediction, decision tree, machine learning, deep learning.

I. INTRODUCTION

Agriculture is vital to humanity, not only for food but also for employment and the economy. Crops or arable land are relatively ''new,'' even though people have been eating grains and plants for over 1,00,000 years. Around 11,000 years ago, during the Neolithic era, often referred to as the New Stone Age, people first actively managed the land and its vegetation. In India, agriculture provides a considerable portion of the country's economic support and the majority of the country's food needs. Due to India's rapid population growth and important climatic changes, the demand chain and food supply must be maintained. Several scientific approaches have been included in agriculture to preserve the harmony between the supply and demand of food. The significant climatic variance makes it difficult for farmers to choose how to be more flexible and sustainable [1]. With modern technology and innovative farming techniques, agriculture must produce more with fewer inputs. Therefore, crop production estimation is crucial in identifying issues with food security. In India, Agriculture uses 70 percent of the water worldwide [2].

More than 50% of Indian workers were employed in agriculture in 2018, which contributed 17%-18% of the

The associate editor coordinating the review of this manuscript and approving it for publication was Yiqi Liu.

© 2023 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

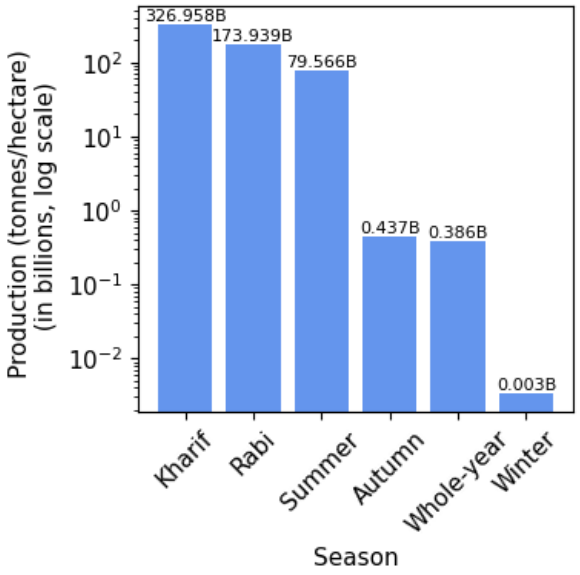


FIGURE 1. Season-wise crop production.

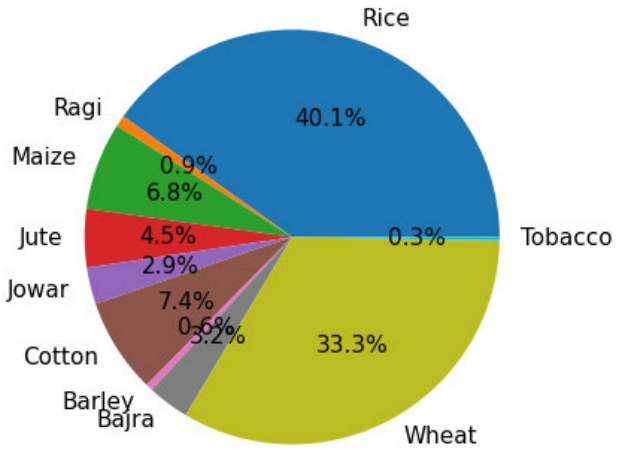


FIGURE 2. Average yearly crop production 1997-2020 [4].

country's GDP. India's national area held steady at about 3,28,726 thousand hectares between 1971 and 2020 [3]. By 2025, India's agricultural industry is expected to grow to USD 24 billion. With 70 percent of sales coming from retail, India has the world's 6th largest food and grocery market. According to preliminary estimates for the fiscal year 2022-2023, the nation will produce 149.92 million tons of food grains altogether (Kharif alone). India's rapid population growth primarily drives the sector.

Additional evidence favoring these results from increasing income levels in urban and rural locations, boosting the demand for agricultural products across the country. As a result, in addition to the introduction of numerous e-farming applications, the increased utilization of cutting-edge technologies is driving sectors such as geographic information systems (GIS), artificial intelligence (AI), blockchain and remote sensing technologies, and drones.

Figure 1 shows the season-wise analysis of crop production for Kharif, Rabi, Summer, Autumn, and Winter seasons and the year. The crop production is maximum in the Kharif season and minimum in winter.

Farmers can boost production under favorable conditions and decrease production loss under unfavorable conditions by using crop yield estimation. Numerous variables, such as farmer practices, choices, pesticides, fertilizers, weather, and market prices, impact crop yield positive forecasts. It is possible to estimate crop yield using statistical information from yields of previous years, together with climate, area-wise output, and rainfall. Recently, machine learning has advanced in various industries, including agriculture.

Numerous machine learning approaches, such as decision trees, artificial neural networks, support vector machines, and deep learning [5], have been applied to calculate agricultural production.

The crop production for India from 1997 to 2020 is shown in Figure 2. From the analysis shown, it can be observed that the majority of the land in the country was planted with wheat and rice, which produced more than 73% of the nation's staple grains.

India accounts for around 40 percent of global rice (Basmati and Non-Basmati) trade and exports to more than 150 countries. Trade ministry data showed that exports rose 11 percent to 2.16 million tons [6] in the first half of 2022-23.

The illustration shows that rice has the highest production and area allocation. The agricultural industry had strong growth in terms of exports over the past year. Exports of rice, basmati and non-basmati totaled USD 6.12 billion in FY22 (through December 2021), as represented in Figure 3.

As farming developed throughout the centuries, people ceased utilizing the stars and animal sacrifices to manage their crops in favor of ever-more-scientific techniques. Mechanization, mathematics, and scales were significant as farming expanded throughout the industrial revolution. Modern approaches were being used more and more. By the 1900s, farming was using regression analysis to gauge agricultural output. A statistical method known as regression analysis combines many statistical methods to determine the correlations between a dependent and independent variable. Regression models for crop forecasting frequently rely on inputs like weather, soil characteristics, and past crop yields. For almost a century, farming has now employed these models. Regression approaches became more popular as relevant data and processing power became more widely accessible. Classification and fruit recognition have also been emerging areas in image classification and computer vision in the agriculture sector [7].

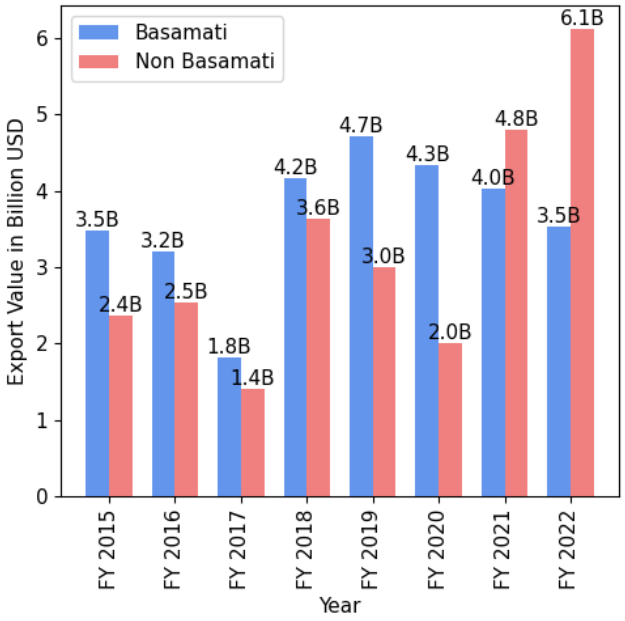


FIGURE 3. Rice exports from India in 2015-2022.

A. RESEARCH CONTRIBUTIONS

India is a predominantly agrarian nation with around 48.9% of the population is employed in agriculture in some capacity. One of the most pressing issues is farmers' suicides. Farmers are killing themselves, which is a national calamity. According to a report of the Indian National Crime Records Bureau (NCRB), more than 2 lakh Indian farmers committed suicide between 1995 and 2020. This serious issue tends to come from farmers' failure to repay loans typically obtained from banks and private landlords. This is because many losses used to occur due to inadequate knowledge about the crops, such as which crop should be sown in which season. The same insight served as the impetus for this study. By utilizing historical agricultural production statistical data for India, a robust mechanism is needed to estimate yield production at the initial stage. Early foresight can assist farmers to avoid committing suicide, suffering financial losses from their crops, and ultimately affecting the nation's GDP. Forecasting agricultural yields isn't a simple task as several factors like rainfall, wind speed, soil characteristics, climate, humidity, temperature, etc., affect agricultural production, and no single dataset is available for the same; this data needs to be collected from multiple sources. Although the same concern has been the subject of numerous studies, improved performance is still preferred. Unquestionably, the most captivating scientific frontiers that inspire academics to investigate further aspects of prediction are combined with machine learning and deep learning. The primary contributions of this research are to propose a reliable agricultural yield prediction method by analyzing the data gathered from official government websites and then applying various ML and DL models to improve accuracy metrics and test losses while offering a standard for contrast in this area of research. This study will help the researchers to understand the problem and solution domains effectively and will enhance their knowledge.

B. ARTICLE ORGANIZATION

The objective of this study is to predict agriculture yields. For a thorough understanding of the literature, Section II looks closely at the theories and related research about crop yield prediction according to the investigation's goals and the research gaps found in the literature. Section III describes the study region, data sources, technological perspective, research methodology, and approaches for agriculture yield prediction, along with the relationships among various dataset features. Section IV shows the model performance of various approaches (i.e., results). Section V concludes the entire work and also suggests additional work.

II. LITERATURE REVIEW

By using machine learning approaches, crop yield estimation can be achieved. The dataset, which contains the total area under cultivation, the canal length, the average maximum temperature, and the water sources (tanks and wells) for irrigation, were used to forecast the crop output. According to a study, the computational model developed was superior to those produced using Regression Tree, Lasso, Deep Neural Network, and Shallow Neural Network approach. RMSE is fifty percent of the standard deviation and twelve percent of the average yield for the dataset validation using projected weather data [8].

From 1998 to 2002, during the Kharif season, the accuracy was 97.5 percent using parameters: min/max/average temperatures, area, rainfall, production, and yield [9]. The study focused on estimating crop yields for the Kharif season in the Vishakhapatnam district of Andhra Pradesh. Rainfall significantly influences Kharif crops' production; hence, the scientists used modular artificial neural networks to forecast rainfall first before utilizing support vector regression to estimate crop output using rainfall and area data. To boost crop yield, these two approaches were used.

The research focused on the following four goals: investigating the Artificial Neural Network (ANN) model to forecast soybean and corn yields under unfavorable weather conditions; examining the model's capacity for regional, state, and local estimation; assessing the performance of the ANN concerning variation of the parameter; and evaluating the evolved ANN model in comparison to additional models of multivariate linear regression [10]. The study used artificial neural networks to assess rice output in various cities in Maharashtra, India. Maharashtra's 27 districts' information was gathered from the open records of the Indian government.

The study estimates superior crop yield using ML methods such as Support Vector Regression (SVR), Random Forest, Artificial Neural Network (ANN), and K-Nearest Neighbor (KNN). Seven hundred forty-five examples comprise the research's data set, with 70 percent of those cases being

randomly assigned for model training and the remaining 30 percent for testing and performance evaluation. According to the final analysis, Random Forest achieves the highest level of accuracy [11]. Utilizing Long-Short Term Memory (LSTM), satellite data in southern Brazil, the research suggests a unique model to forecast soybean yield.

The primary aim of the research is to compare the effectiveness of LSTM neural networks, random forest, and multivariate OLS linear regression [12]. Rainfall, Land surface temperature, and Vegetation indices are used as independent variables in the forecasting process for soybean data, and step two is to determine how soon the model can reasonably anticipate the yield. For all forecasts except DOY 16, Long Short Term Memory outperforms all other algorithms. Multivariate OLS linear regression outperforms all other algorithms for DOY 16 [13]. The results of using a Sequential Minimal Optimization Classifier are analyzed in this research. The experiment was conducted using the WEKA tool and data from 27 districts in Maharashtra, India.

According to the experiment's results on the same dataset, other strategies outperform Sequential minimum optimization. Sequential minimum optimization demonstrated the minimum accuracy and bad quality, whereas Multilayer Perceptron and BayesNet demonstrated the best accuracy and improved quality [14].

Crop productivity estimation is proposed using a Deep Belief Network (DBN) and Parallel Layer Regression (PLR) approach. Here, five major crops flourishing in Karnataka: pulse, ragi, and rice, are the subject of a DBN approach. The suggested methodology predicts each location in the relevant database to grow one of the five crops. Last but not least, the experimental findings demonstrate that the approach has a significant potential for accurate prediction of agricultural productivity in terms of accuracy, specificity, and sensitivity that its effectiveness has been validated using real-time data and interactions with people [15].

Approaches for machine learning to estimate agricultural yield: linear/ lasso/ridge regression and decision trees have been used. Some machine learning methods were inferior to the Decision tree [16].

A Crop Yield Prediction System (CYPS) is implemented using a KNN algorithm. For a farmer, however, yield predictions must be based on various factors that could influence crop production and quality. Season, crop type, and production area are three aspects that influence yield production; therefore, authors use specific fields like year, crop, area, region, and season to predict yield production. Making decisions related to agricultural risk management requires accurate knowledge of crop yield history [17].

KNN, Random Forest, and Decision Tree Classifier were examined by Rao et al. [18] using two distinct metrics: Gini and Entropy. Findings show that Random Forest has given the most accurate results.

Good performances of 91.35% and 91.17% have been obtained for VGG_19 and VGG_16 based on feature vectors [19]. Vanipriya et al. [20] have proposed using Hydroponics to deal with low agricultural production issues in India as it has high efficiency. Additionally, it offers soil cultivation a greener alternative. Food production depends on the economy and also on the yield of agricultural production [21].

III. PROPOSED WORK

A. STUDY AREA

India is the subject of this study because of its diverse climate, which ranges from humid to arid in the south to temperate alpine in the north. It spans 3,28,726 thousand hectares [22] from the Himalayan peaks capped in the snow to the rain forests in southern central tropical, out of which 2,97,319 thousand hectares (between 1971 and 2020) is for agricultural usage [3]. To conduct this study, ten key Indian crops (majorly grown crops)- rice, jowar, maize, bajra, tobacco, jute, barley, ragi, cotton, and wheat have been chosen.

B. DATA SOURCES

For creating the dataset, the information was gathered from the publicly available official websites https://data.gov.in and https://aps.dac.gov.in (i.e. multiple sources) [23], [24]. The dataset generated contains the details for the years 1997 to 2020: State name, district name, crop year, seasons, crop type, rainfall, wind speed, humidity, area under irrigation, area, production, and yield. Figure 2 depicts the crop's yield for the years mentioned. To forecast the crop yield for India, this study uses a Decision Tree, Random Forest, XG Boost Regression, Convolutional Neural Network, and Long-short Term Memory Network. Furthermore, accuracy, standard deviation, root mean squared error and test loss are used for validation.

The flow of the work done is demonstrated in Figure 4. Here, the dataset is partitioned into training and testing data. 70 percent of the entire data is selected for training purposes, and the remaining 30 percent is used for validation. On the training dataset, k-fold cross-validation is employed. The number of folds taken is four, and training data is randomly assigned to each fold. The trained model is then evaluated.

C. METHODS

1) K-FOLD CROSS VALIDATION

In k-FOLD Cross Validation, k-1 folds are used as training data to train the model, and rest of the data is used to validate the resulting model.

Accuracy, the performance metric, is computed at each step while this process is performed k times. k-Fold cross validation guarantees that in both the training and test sets, every observation from the original dataset has a chance to appear. If we just have a small amount of input data, this is one of the finest approaches. The flow chart for the proposed model is shown in Figure 5.

2) DECISION TREE

Classification and regression issues are resolved using a decision tree that performs two essential functions: firstly,

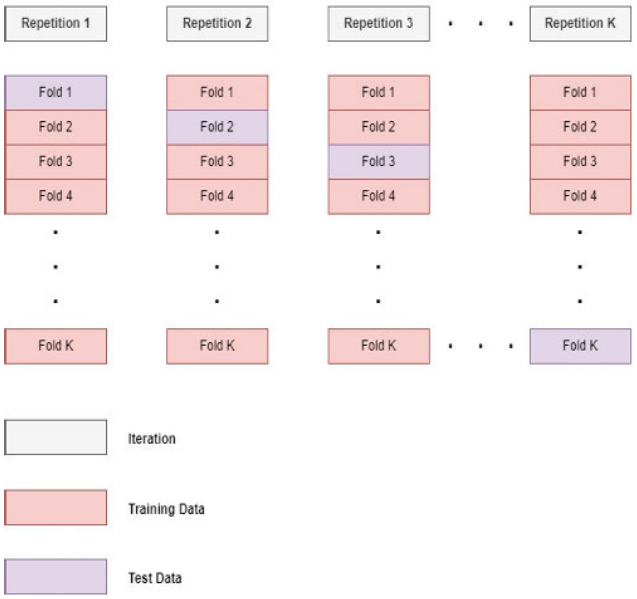


FIGURE 4. The k-fold cross-validation.

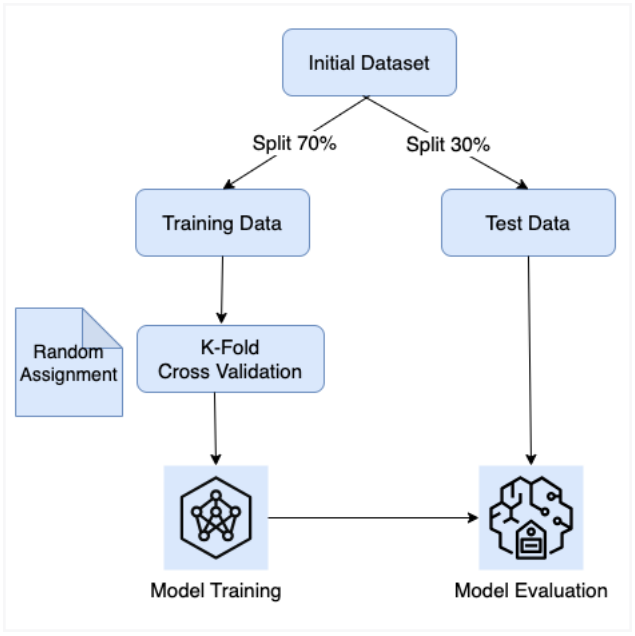


FIGURE 5. Flow chart of the approach used.

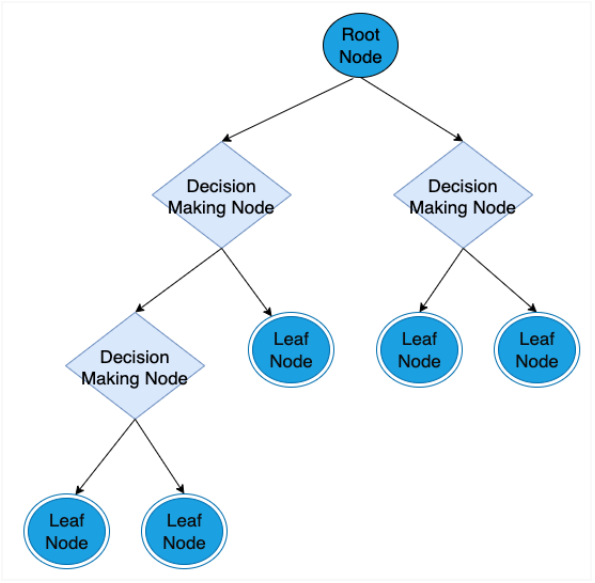


FIGURE 6. Decision tree.

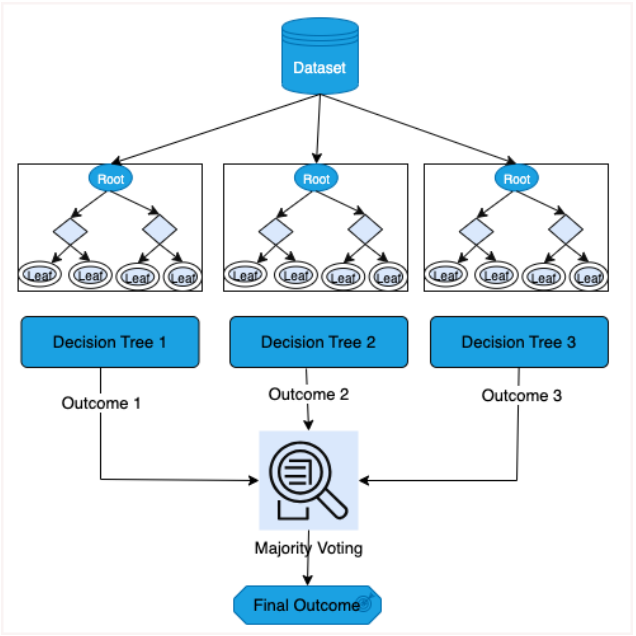


FIGURE 7. Random forest.

it categorizes the features that are pertinent to each decision, and secondly, it determines the best course of action based on the selected features. The plausible choice is given a probability distribution by the Decision Tree algorithm [25]. Here, every node represents a feature, the branch devotes to a selection, and the leaf node denotes the outcome. One characteristic should be the decision tree's root node to begin tree production. Data splitting is then necessary to finish the decision tree.

3) RANDOM FOREST

An uncorrelated forest of decision trees is built by amalgamating the bagging approach and feature randomness, i.e., the bagging approach's extension is RF [26]. Low correlation across decision trees is ensured via feature bagging or feature randomization, or the random subspace technique, which provides a random selection of features. RF only chooses a portion of the feature splits that decision trees can take into account [27].

Each decision tree in the ensemble is built using the bootstrap sample, a data sample obtained from a training set. The rest 1/3rd of the training sample is used as test data.

4) XGBoost

A popular and successful open-source alternative of the gradient-boosted trees method is called Extreme Gradient Boosting, i.e., XGBoost. Regression trees [28] act as weak

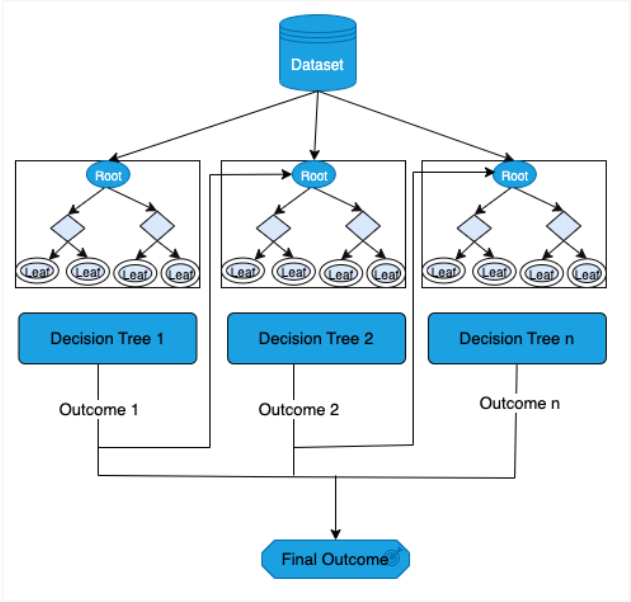


FIGURE 8. XGBoost.

learners when gradient boosting is used for regression. With the use of a continuous score, a connection is established between every input data point and a leaf. XGBoost's convex loss function takes a penalty term for model complexity to minimize a regularized (L1 and L2) objective function/ regression tree functions [29].

Repeatedly, new trees are added to the training process for predicting the errors/residuals of previous trees, which are then amalgamated with preceding trees to make the ultimate forecast. Since the technique reduces the loss while adding new models, it is frequently called ''gradient boosting'' [30].

Decision tree, random forest, and XGBoost are each represented graphically in Figures 6, 7, and 8, respectively.

5) CONVOLUTIONAL NEURAL NETWORK (CNN)

Convolution Neural Networks use some mathematical operations in between their layers called Convolution. This implementation of CNN has a total of 7 layers. The first layer is a Conv1d layer with 64 filters of the size (3*3), each with a kernel size of 3. Layer 2 (MaxPooling1D layer) has a pool size of two. Dropout is visible on layer 3. The output from layer three is flattened by layer four, which uses the ReLU activation function, and this output is then sent to layer 5. A neural network's hidden layer 6 has 330 neurons in it. The output layer, layer 7, uses the SoftMax function and has 11 neurons for 11 kinds of output. CNN architecture of the model used is shown in Figure 9.

6) LONG SHORT-TERM NETWORK (LSTM)

A type of recurrent neural network (RNN) called a Long Short-Term Memory (LSTM) network, shown in Figure 10, specifically handles sequential data, such as text, speech, and time series. Sequential data can be used to teach LSTM

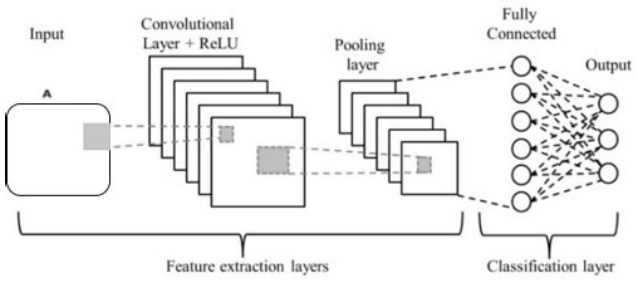


FIGURE 9. CNN architecture.

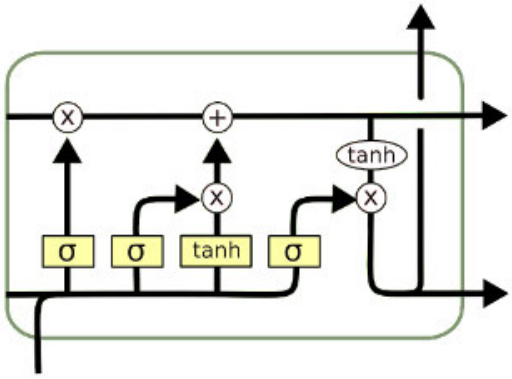


FIGURE 10. Working of LSTM.

networks long-term dependencies. For both models, i.e., CNN and LSTM, the optimizer used is ADAM optimizer that is used to change or tune the attributes of a neural network [31], [32] such as layer weights, learning rate, etc., and it is a combination of ''RMSP'' and ''gradient descent with momentum'' algorithm. The loss is calculated using MSE (i.e., mean squared error). For training the model, the number of epochs taken is 50, the batch size is 32, and the validation split is 0.2. LSTM architecture of the model used is shown in Figure 10.

IV. RESULTS

The features of the dataset are related to one another. The authors identified the crop as a significant feature and plotted the link. Figure 11 shows the production count for each of the 10 crops taken consideration in the study. Figure 12 shows the historical correlations between Area and Crop Type. It can be observed that the maximum area is allotted to wheat, followed by Rice.

Figures 11 and 12 show the historical correlations between the dataset's features, such as Crop type, Area, and Production.

Table 1 presents the forecast outcomes. The study concludes that Random Forest outperforms other machine learning algorithms regarding accuracy.

Using statistical data, Random Forest produces India's most accurate crop production estimation with accuracy = 98.96 %, MAE = 1.97, RMSE = 2.45, and SD = 1.23

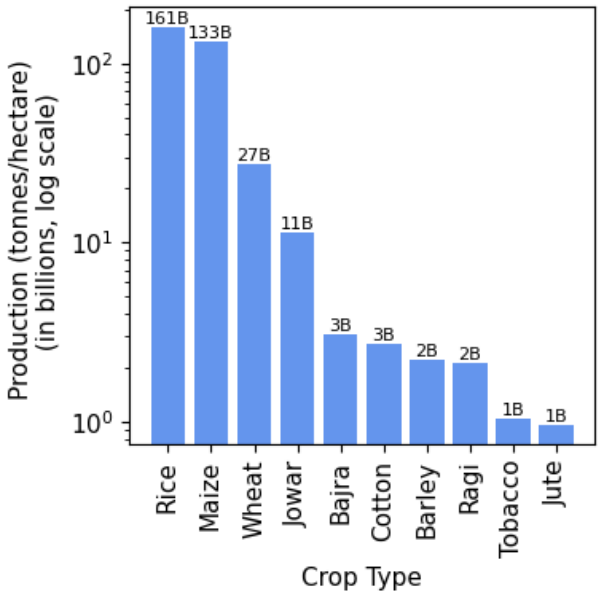


FIGURE 11. Relationship between crop and Production.

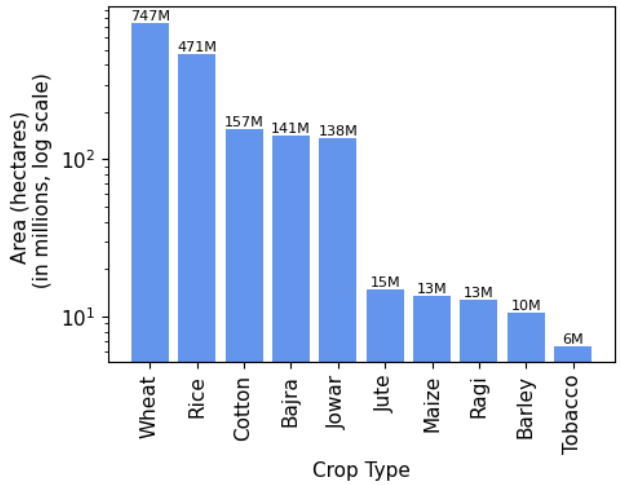


FIGURE 12. Area allotted to crops.

TABLE 1. Simulating model performance with area and production as inputs.

Model	Accuracy	Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)	Standard Deviation (SD)
Random Forest	98.96	1.97	2.45	1.23
Decision Tree	89.78	4.58	5.86	2.75
XGBoost	86.46	6.31	7.89	3.54

(Table 1). For the Decision tree and XGBoost, the accuracy, mean absolute error, root mean square error, and standard deviation are 89.78%, 4.58, 5.86, 2.75 and 86.46%, 6.31, 7.89, 3.54. Figure 13, Figure 14, and Figure 15 show the

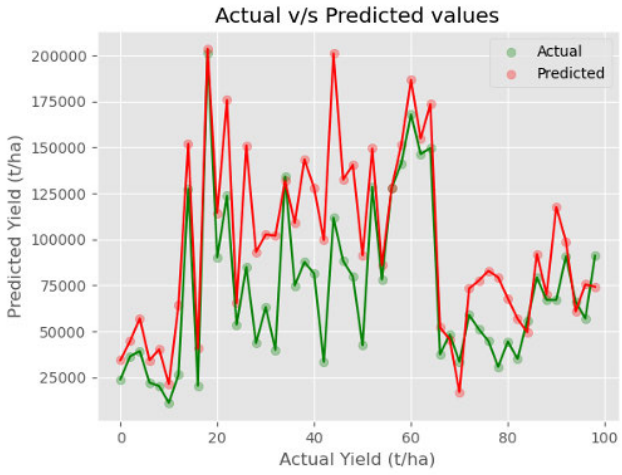


FIGURE 13. Model performance of decision tree.

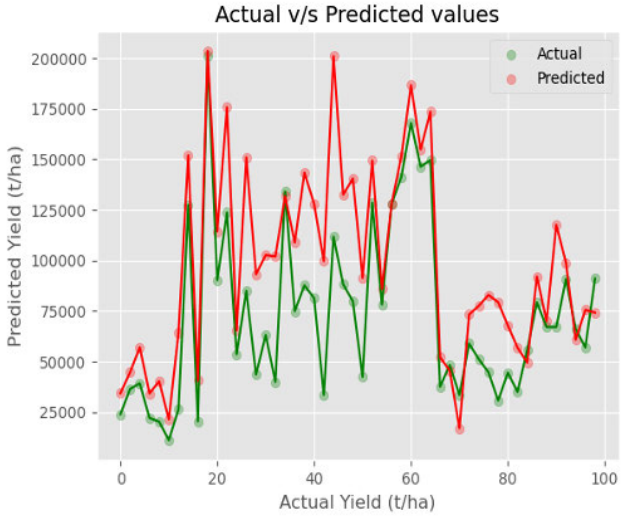


FIGURE 14. Model performance of random forest.

TABLE 2. Test results of CNN and LSTM.

Model	Test Loss
CNN	0.00060
LSTM	0.00063

model performance of Decision Tree, Random Forest, and XGBoost, respectively.

With accuracy ratings of 89.78 and 86.46, respectively, Decision Tree and XGBoost Regression performs less compared to Random Forest. Machine learning is a "black box" technique because it doesn't interpret much. Random Forest outperformed the other three regression approaches in this study's application of machine learning techniques to forecast the estimated crop output for India.

Figure 16 details the layers used in CNN and LSTM. Table 2 shows the Test loss for CNN and LSTM, evaluated as 0.00060 and 0.00063 respectively.

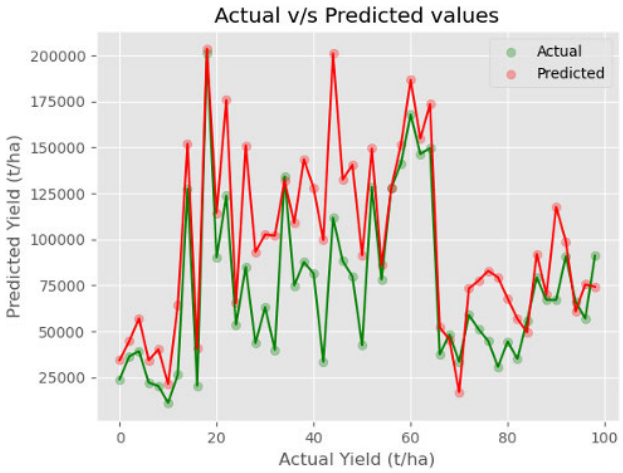


FIGURE 15. Model performance of XGBoost regression.

Layer (type)	Output Shape	Param #
conv1d_1 (Conv1D)	(None, 2, 64)	448
max_pooling1d_1 (MaxPooling 1D)	(None, 1, 64)	0
lstm_1 (LSTM)	(None, 50)	23000
dense_3 (Dense)	(None, 1)	51

Total params: 23,499
Trainable params: 23,499
Non-trainable params: 0

FIGURE 16. Layer description of CNN and LSTM.

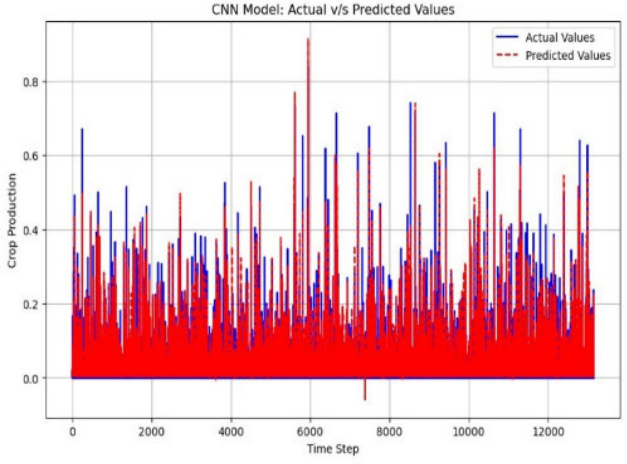


FIGURE 17. Model performance of CNN.

Figure 17 and 18 shows the model performance of CNN and LSTM respectively.

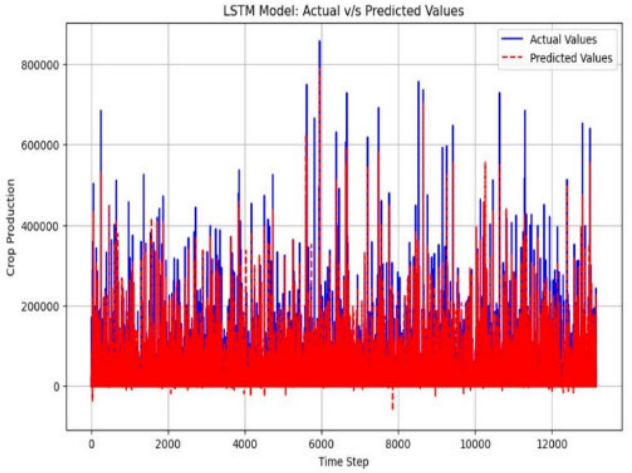


FIGURE 18. Model performance of LSTM.

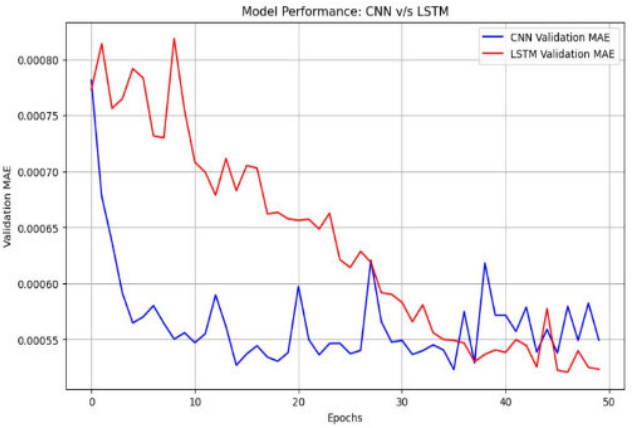


FIGURE 19. CNN v/s LSTM.

Figure 19 illustrates a comparison of CNN and LSTM. Result show that a change in the number of epochs leads to a significant change in mean absolute error. Here, it can be observed that CNN performs better than LSTM as the loss is less with CNN.

V. CONCLUSION

The demand and supply for food have grown more difficult to manage as the population grows. To assist farmers, experts have worked hard over the past few years to anticipate agricultural yield production. In order to forecast India's crop yield, this study uses various machine learning and deep learning approaches. The study underlines the advantages of cutting-edge procedures. It is beneficial for small-scale ranchers, as they may use the predictions to estimate crop production for upcoming years and plant it appropriately. Five machine learning and deep learning algorithms, Decision Tree, Random Forest, XGBoost regression, Convolutional Neural Network, and Long-Short Term Memory Networks are applied to the dataset taken into consideration. When data is analyzed at the country level, Random Forest

(with accuracy-98.96%, mean absolute error-1.97, RMSE-2.45 and standard deviation-1.23) and CNN (with minimum loss-0.00060) perform better according to the current prediction. Experimental findings demonstrate that the approach has a great potential for precise crop productivity prediction and its effectiveness has been validated using real-time data and interactions with people. More data for each crop year having more historically precise information about the climate and environment is needed. More deep learning models need to be applied to the dataset to identify the method that performs the best. To increase the model's accuracy in crop production prediction, remote sensing data could be amalgamated with statistical data of districts. The prediction can be more accurate using satellite imagery land cover or satellite image classification.

REFERENCES

[1] N. Gandhi, L. J. Armstrong, O. Petkar, and A. K. Tripathy, "Rice crop yield prediction in India using support vector machines," in *Proc. 13th Int. Joint Conf. Comput. Sci. Softw. Eng. (JCSSE)*, Jul. 2016, pp. 1–5.

[2] M. Khan and S. Noor, "Irrigation runoff volume prediction using machine learning algorithms," *Eur. Int. J. Sci. Technol.*, vol. 8, pp. 1–22, Jan. 2019.

[3] NIC IN. (2023). *Annual Report 2020–21*. Accessed: Jan. 20, 2023. [Online]. Available: https://agricoop.nic.in/Documents/annual-report-2020-21

[4] Govt India. (2023). *Crop Production Statistics Information System*. [Online]. Available: https://aps.dac.gov.in/APY/Index.htm

[5] M. Khan and S. Noor, "Performance analysis of regression-machine learning algorithms for predication of runoff time," *Agrotechnology*, vol. 8, no. 1, pp. 1–12, 2019.

[6] IBEF. (2023). *Agriculture in India: Industry Overview, Market Size, Role in Development*. Accessed: Jan. 20, 2023. India. [Online]. Available: https://www.india.gov.in/

[7] H. S. Gill, G. Murugesan, B. S. Khehra, G. S. Sajja, G. Gupta, and A. Bhatt, "Fruit recognition from images using deep learning applications," *Multimedia Tools Appl.*, vol. 81, no. 23, pp. 33269–33290, Sep. 2022.

[8] S. Khaki and L. Wang, "Crop yield prediction using deep neural networks," *Frontiers Plant Sci.*, vol. 10, p. 621, May 2019.

[9] E. Khosla, R. Dharavath, and R. Priya, "Crop yield prediction using aggregated rainfall-based modular artificial neural networks and support vector regression," *Environ., Develop. Sustainability*, vol. 22, no. 6, pp. 5687–5708, Aug. 2020.

[10] P. S. Maya Gopal and R. Bhargavi, "Optimum feature subset for optimizing crop yield prediction using filter and wrapper approaches," *Appl. Eng. Agricult.*, vol. 35, no. 1, pp. 9–14, 2019.

[11] P. S. M. Gopal and R. Bhargavi, "Performance evaluation of best feature subsets for crop yield prediction using machine learning algorithms," *Appl. Artif. Intell.*, vol. 33, no. 7, pp. 621–642, Jun. 2019.

[12] The Food and Agriculture Organization (FAO). (2023). *FAOSTAT*. Accessed: Jan. 20, 2023. [Online]. Available: https://www.fao.org/faostat/en/#rankings/countries%20_by_com%20modity_exports

[13] F. Raimundo, A. Glória, and P. Sebastião, "Prediction of weather forecast for smart agriculture supported by machine learning," in *Proc. IEEE World AI IoT Congr. (AIIoT)*, May 2021, pp. 160–164.

[14] A. Suruliandi, G. Mariammal, and S. P. Raja, "Crop prediction based on soil and environmental characteristics using feature selection techniques," *Math. Comput. Model. Dyn. Syst.*, vol. 27, no. 1, pp. 117–140, Jan. 2021.

[15] P. Mohan and K. K. Patil, "Crop production rate estimation using parallel layer regression with deep belief network," in *Proc. Int. Conf. Electr., Electron., Commun., Comput., Optim. Techn. (ICEECCOT)*, Dec. 2017, pp. 168–173.

[16] M. Kavita and P. Mathur, "Crop yield estimation in India using machine learning," in *Proc. IEEE 5th Int. Conf. Comput. Commun. Autom. (ICCCA)*, Oct. 2020, pp. 220–224.

[17] S. Kunchakuri, S. Pallerla, S. Kande, and N. R. Sirisala, "An efficient crop yield prediction system using machine learning algorithm," in *Proc. 4th Smart Cities Symp. (SCS)*, vol. 2021, Nov. 2021, pp. 120–125.

[18] M. S. Rao, A. Singh, N. V. S. Reddy, and D. U. Acharya, "Crop prediction using machine learning," *J. Phys., Conf. Ser.*, vol. 2161, no. 1, 2022, Art. no. 012033.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[20] C. H. Vanipriya, Maruyi, S. Malladi, and G. Gupta, "Artificial intelligence enabled plant emotion expresser in the development hydroponics system," *Mater. Today, Proc.*, vol. 45, pp. 5034–5040, Jan. 2021.

[21] A. Tomar, G. Gupta, W. Salehi, C. H. Vanipriya, N. Kumar, and B. Sharma, "A review on leaf-based plant disease detection systems using machine learning," in *Proc. ICRIC*, vol. 1, 2022, pp. 297–303.

[22] Govt India. (2023). *Profile*. Accessed: Jan. 20, 2023. [Online]. Available: https://www.india.gov.in/india-glance/profile

[23] Govt India. (2023). *Data*. Accessed: Jan. 20, 2023. [Online]. Available: https://data.gov.in

[24] Govt India. (2023). *Crop Production Statistics Information System*. Accessed: Jan. 20, 2023. [Online]. Available: https://aps.dac.gov.in/APY/Index.htm

[25] D. J. Reddy and M. R. Kumar, "Crop yield prediction using machine learning algorithm," in *Proc. 5th Int. Conf. Intell. Comput. Control Syst. (ICICCS)*, May 2021, pp. 1466–1470.

[26] S. Bhansali, P. Shah, J. Shah, P. Vyas, and P. Thakre, "Healthy harvest: Crop prediction and disease detection system," in *Proc. IEEE 7th Int. Conf. Converg. Technol. (I2CT)*, Apr. 2022, pp. 1–5.

[27] S. Agarwal and S. Tarar, "A hybrid approach for crop yield prediction using machine learning and deep learning algorithms," *J. Phys., Conf. Ser.*, vol. 1714, no. 1, Jan. 2021, Art. no. 012012.

[28] R. Gupta, A. K. Sharma, O. Garg, K. Modi, S. Kasim, Z. Baharum, H. Mahdin, and S. A. Mostafa, "WB-CPI: Weather based crop prediction in India using big data analytics," *IEEE Access*, vol. 9, pp. 137869–137885, 2021.

[29] N. Bali and A. Singla, "Emerging trends in machine learning to predict crop yield and study its influential factors: A survey," *Arch. Comput. Eng.*, vol. 29, pp. 1–18, Mar. 2022.

[30] S. Iniyan, V. A. Varma, and C. T. Naidu, "Crop yield prediction using machine learning techniques," *Adv. Eng. Softw.*, vol. 175, Jan. 2023, Art. no. 103326.

[31] Z. Guo, Q. Liu, and J. Wang, "A one-layer recurrent neural network for pseudoconvex optimization subject to linear equality constraints," *IEEE Trans. Neural Netw.*, vol. 22, no. 12, pp. 1892–1900, Dec. 2011.

[32] A. V. S. Kumar, P. Sharma, S. Kaur, O. S. Saleh, H. R. Chennamma, and A. Chaturvedi, "AI-equipped IoT applications in high-tech agriculture using machine learning," in *Handbook of Research on AI-Equipped IoT Applications in High-Tech Agriculture*. Hershey, PA, USA: IGI Global, 2023, pp. 38–64.

PRIYANKA SHARMA was born in Jaipur, Rajasthan, India, in 1991. She received the B.Tech. degree in information technology and the M.Tech. degree in computer science from Rajasthan Technical University, Kota, Rajasthan, in 2012 and 2016, respectively. She is currently an Assistant Professor with the Department of Computer Science and Engineering (NBA Accredited), Swami Keshvanand Institute of Technology, Management and Gramothan (SKIT), Jaipur, Rajasthan (Accredited by NAAC A++ Grade). She was also appointed as a Faculty Advisor of the Google Developer Students Club, SKIT. She has guided various M.Tech. and UG research scholars. She has published more than 15 research papers in national and international journals/conferences and book chapters. Also, she has taken part in forums hosted by Infosys, TCS, Oracle, and IBM. She has received accolades a number of times in a variety of fields. Her research interest includes investigating the advancements in machine learning and deep learning applications in various problem domains. She is a member of ACM and various professional societies. She was also given recognition by Infosys, India, for her remarkable performance in the Campus Connect Program.

PANKAJ DADHEECH (Member, IEEE) is currently a Professor and the Deputy Head of the Department of Computer Science and Engineering (NBA Accredited), Swami Keshvanand Institute of Technology, Management and Gramothan (SKIT), Jaipur, Rajasthan, India (Accredited by NAAC A++ Grade). He has more than 18 years of experience in teaching. He has been appointed as a Ph.D. Research Supervisor with the Department of Computer Science and Engineering, SKIT (Recognized Research Centre of Rajasthan Technical University, Kota). He has also guided various M.Tech. research scholars. He has published 22 patents and granted six patents at Intellectual Property India, Office of the Controller General of Patents, Design and Trade Marks, Department of Industrial Policy and Promotion, Ministry of Commerce and Industry, Government of India. He has also published and granted five Australian patents at the Commissioner of Patents, Intellectual Property Australia, Australian Government. He has also published and granted one German patent, one South African patent, and one USA patent. He has also registered and granted two research copyrights at the Registrar of Copyrights, Copyright Office, Department for Promotion of Industry and Internal Trade, Ministry of Commerce and Industry, Government of India. He has presented 60 papers at various national and international conferences. He has 67 publications in various international and national journals. He has published seven books and 32 book chapters. His research interests include high-performance computing, cloud computing, information security, big data analytics, intellectual property rights, and the Internet of Things. He is a member of many professional organizations, such as the CSI, ACM, IAENG, and ISTE. He has chaired technical sessions at various international conferences and contributed as a resource person in various FDPs, workshops, STPs, and conferences. He is also acting as the guest editor of various reputed journal publishing houses and conference proceedings and the Bentham Ambassador of Bentham Science Publisher.

NAGENDER ANEJA received the M.E. degree in computer technology and applications from the Delhi College of Engineering and the Ph.D. degree in computer engineering from the J. C. Bose University of Science and Technology, YMCA, Faridabad, India. He was an Associate IP Lead for Microsoft Patent Research Services with CPA Global India for five years. He is currently an Assistant Professor with the School of Digital Science, Universiti Brunei Darussalam, Brunei Darussalam. He is also the Founder of https://researchid.co/, a platform allowing faculty members and researchers to create profiles for their research. Several process innovations have been developed by him, including the development of automation tools for patent analysis with CPA Global, as well as the development of an expert directory for Universiti Brunei Darussalam. He has also held teaching and research positions at various universities in India and Brunei Darussalam. He is a highly experienced and accomplished researcher who has published over 35 papers in top academic journals and conferences. He has more than 20 years of experience in computer science. He has two patents from the U.S. patent and Trademark Office. His research interests include deep learning, computer vision, and natural language processing. He also received several awards, including the Brunei ICT Award, in 2016, for a mobile app that creates a social network with nearby users by matching interests in airplane mode. He also received the Research Award 2020–2021 from the J. C. Bose University of Science and Technology, YMCA, for several SCIE-indexed publications during the Ph.D. degree. His website, http://naneja.github.io, provides further information about his work and research interests.

SANDHYA ANEJA (Member, IEEE) is currently a Researcher, an Inventor, and a Computer Scientist with a strong passion for teaching. She is also an Assistant Professor with the School of Computer Science and Mathematics, Marist College, Poughkeepsie, NY, USA. She is also a Visiting Research Scholar with the Department of Computer Science, Purdue University. She has over 15 years of experience teaching computer science to undergraduate and graduate students at the University of Delhi and Universiti Brunei Darussalam. As a Researcher, she contributed to developing a mobile application to facilitate the matching of interests on available mobile devices and allow exchanging of messages and files. The application allows broadcasting names and a limited number of keywords representing users' interests without any connection in a nearby region. The broadcasting region creates a mobile wireless network limited by the Wi-Fi region that is around 200 meters. She also received a U.S. patent on this technology. As a Computer Scientist, she has received project funding from the University of Delhi as a PI and Universiti Brunei Darussalam as a co-PI. She has extensively worked on Brunei government-funded projects with IBM researchers. She is also a Contributor to Sandia and DARPA-funded projects with Purdue University.

Civil Engineering and Architecture 11(5): 2377-2391, 2023 http://www.hrpub.org
DOI: 10.13189/cea.2023.110510

Prediction of Energy Dissipation over Stepped Spillway with Baffles Using Machine Learning Techniques

Saurabh Pujari*, Vijay Kaushik, S. Anbu Kumar

Department of Civil Engineering, Delhi Technological University, India

Received February 23, 2023; Revised April 25, 2023; Accepted June 11, 2023

Cite This Paper in the Following Citation Styles

(a): *[1] Saurabh Pujari, Vijay Kaushik, S. Anbu Kumar , "Prediction of Energy Dissipation over Stepped Spillway with Baffles Using Machine Learning Techniques," Civil Engineering and Architecture, Vol. 11, No. 5, pp. 2377 - 2391, 2023. DOI: 10.13189/cea.2023.110510.*

(b): *Saurabh Pujari, Vijay Kaushik, S. Anbu Kumar (2023). Prediction of Energy Dissipation over Stepped Spillway with Baffles Using Machine Learning Techniques. Civil Engineering and Architecture, 11(5), 2377 - 2391. DOI: 10.13189/cea.2023.110510.*

Copyright©2023 by authors, all rights reserved. Authors agree that this article remains permanently open access under the terms of the Creative Commons Attribution License 4.0 International License

Abstract In river engineering, the stepped spillway of a dam is an important component that may be used in various ways. It is necessary to conduct research dealing with flood control in order to investigate the method, in which energy is lost along the tiered spillways. In the past, several research projects on stepped spillways without baffles have been carried out utilizing a range of research approaches. In the present study, machine learning techniques such as Support Vector Machine (SVM) and Regression Tree (RT) are used to analyze the energy dissipation on rectangular stepped spillways that make use of baffles in a variety of configurations and at a range of channel slopes. The results of many experiments indicate that the amount of energy that is lost increases with the number of baffles that are present in flat channels with slopes and rises. In order to evaluate the efficiency and usefulness of the suggested model, the statistical indices that were developed for the experimental research are used to validate the models that were created for the study. The findings indicate that the suggested SVM model properly predicted the amount of energy that was dissipated when contrasted with RT and the method that had been developed in the past. This study verifies the use of machine learning techniques in this industry, and it is unique in that it anticipates energy dissipation along stepped spillways utilizing baffle designs. In addition, this work validates the use of machine learning methods in this field.

Keywords Rectangular Stepped Spillways, Baffle Arrangements, Channel Slope, Support Vector Machine (SVM), Regression Tree (RT)

1. Introduction

To regulate water flows downstream of a dam, a spillway structure is employed, with stepped spillways preventing water from overflowing and causing damage to the dam. These spillways consist of a channel with built-in steps or drops. Flow patterns observed include nappe flow, transition flow, and skimming flow [1]. Numerous scholars have looked at the energy dissipation in stepped spillways [2-4]. Boes and Hager [5] looked at the benefits of stepped spillways, such as their simplicity of construction, less danger of cavitation, and smaller stilling basins at downstream dam toes owing to considerable energy loss along the chute. Hazzab and Chafic [7] conducted an experimental study on energy dissipation in stepped spillways and reported on flow configurations. Additionally, the Manksvill dam spillway was examined using a 1:25 scale physical wooden model [6]. For moderately inclined stepped channels, Stefan and Chanson [8] explored air-water flow measurements. Daniel [9] discussed how the existence of steps and step heights affect stepped spillways' ability to dissipate energy. A comparison of the smooth invert chute flow with the self-aerated stepped spillway. The energy dissipation in stepped spillways was investigated using various methods. Katourany [10] compared experimental findings to conventional USBR outcomes to examine the effects of different baffle widths, spacing between baffle rows, and step heights of baffled aprons. Salmasi et al. [11] assessed the energy dissipation of through-flow and over-flow in gabion stepped spillways, discovering that gabion spillways with pervious surfaces dissipated energy more efficiently than those with concrete walls. Other forms of

stepped spillways, such as inclined steps and steps with end sills, were also quantitatively studied for energy dissipation [12]. Saedi and Asareh [13] examined how the number of drop stairs affected energy dissipation in stepped drops and suggested using stepped drops to increase energy dissipation by providing flow path roughness. Al-Husseini [14] found that decreasing the number of steps and downstream slopes led to an increase in flow energy dissipation, and that the use of cascade spillways reduced energy dissipation compared to the original step spillway. MARS and ANN methods were used to estimate energy dissipation in flow across stepped spillways under skimming flow conditions, with both models proving reliable [15]. Frederic et al. [16] evaluated the energy dissipation effectiveness and stability of the Mekin Dam spillway by confirming that flow did not result in transitional flow and by calculating safety factors at various intervals. A numerical model was developed to validate a physical model examining the impact of geometrical parameters on the dissipation rate in flows through stepped spillways [17]. The regulation of the rates of dissipation is studied using a particular kind of fuzzy inference system (FIS). The findings are compared with a predefined numerical database to determine the predicted energy dissipation under various circumstances. The findings show that the suggested FIS may be a useful tool for the operational management of dissipator structures while taking various geometric characteristics into account. Nasralla [18] studied the four phases of the spillway and conducted eighteen runs to enhance energy dissipation through the contraction-stepped spillway. The study considered alternative baffle placements, heights, and widths. The results showed that downstream baffles on the stepped spillway of the stilling basin improve energy dissipation. Using the Flow 3D software, Ikinciogullari [19] quantitatively analyzed the energy dissipation capabilities of trapezoidal stepped spillways using four distinct models and three different discharges. The findings showed that trapezoidal stepped spillways are up to 30% more efficient in dissipating energy than traditional stepped spillways. In previous works, only a few machine learning algorithms were used to forecast energy dissipation across a rectangular stepped spillway without baffles. Therefore, this study used machine learning approaches such as Support Vector Machine (SVM) and Regression Tree (RT) to predict energy dissipation across a rectangular stepped spillway with varied rectangular-shaped baffle configurations at different channel slopes. The study compared these models using statistical analysis to assess their efficiency in predicting energy dissipation over rectangular stepped spillways with baffles.

2. Materials and Methods

2.1. Experimental Setup

The experiments were carried out at the Hydraulics laboratory of Delhi Technological University. The tests were performed in a rectangular tilting flume of 8m long, 0.30m wide and 0.40m deep which has a facility to make it horizontal and sloping as well (shown in Figure 1). The flume consists of an inlet section, an outlet section, and a collecting tank at the downstream end which is used to measure the discharge. Figure 2 depicts the model of a rectangular stepped spillway prepared using an acrylic sheet having a width of 0.30m, a height of 0.20m and a base length of 0.40m. A total of four steps were designed with a step height of 0.05m, the step length is 0.10m and rectangular-shaped baffles of length 0.10m and height of 0.05m were arranged in different manner. Figure 3 represents the different baffle arrangements used in the experimental work. At first, the experiment was conducted for no baffle condition. Thereafter the experiment was conducted for the first arrangement of three baffles, in which two baffles were placed at a distance of 0.10m from the toe of the spillway and a distance of 0.10m was maintained between the first two baffles and the third baffle was placed between the first two baffles at a distance of 0.20m from the toe of the spillway (figure 4a). After that, the experiment was conducted for the third arrangement of baffles which consists of five baffles, two more baffles were introduced at a distance of 0.30m from the toe of the spillway and a distance of 0.10m was maintained between them (figure 4b). The baffles used in the experiment were rectangular shaped which had a height of 0.05m and length of 0.10m. The experiments were conducted for five different discharges 2 l/s, 4 l/s, 6 l/s, 8 l/s and 10 l/s. For the purpose of determining the head values both upstream and downstream of the spillway model, a point gauge with a precision of 0.1mm was used. In order to determine the average velocities of the upstream and downstream portions, respectively, a pitot static tube was used in conjunction with a digital manometer.

2.2. Theoretical Background

Nasralla [18] proposed an empirical equation (1) using multivariate regression analysis to predict the relative energy over the rectangular stepped spillway in terms of nondimensional parameters like spillway relative height, relative baffle location, relative baffle height, relative baffle width and Froude number. The energy dissipation over a rectangular stepped spillway is calculated using equations 2 to 4.

$$\frac{E_T - E_L}{E_u} = -0.099 + 0.085F_1 - 0.022\left(\frac{a}{y_2}\right) + 0.047\left(\frac{h_b}{y_1}\right) - 0.051\left(\frac{w_b}{W}\right) \quad (1)$$

$$E_u = y_1 + \frac{V_1^2}{2g} \quad (2)$$

$$E_d = y_2 + \frac{V_2^2}{2g} \quad (3)$$

$$E_L = E_u - E_d \quad (4)$$

The numerous factors have an effect on the energy

dissipation that occurs via the stepped spillways, and the functional connection that exists between these parameters may be described as follows:

$$\frac{E_T-E_L}{E_u} = f(H, y_c, a, y_2, h_b, y_1, w_b, W, V, g, H_u) \quad (5)$$

After dimensional analysis, the necessary dimensionless equation may be expressed as follows:

$$\frac{E_T-E_L}{E_u} = f\left(\frac{H}{y_c}, \frac{a}{y_2}, \frac{h_b}{y_1}, \frac{w_b}{W}, F_1\right) \quad (6)$$

Figure 1. Rectangular tilting flume

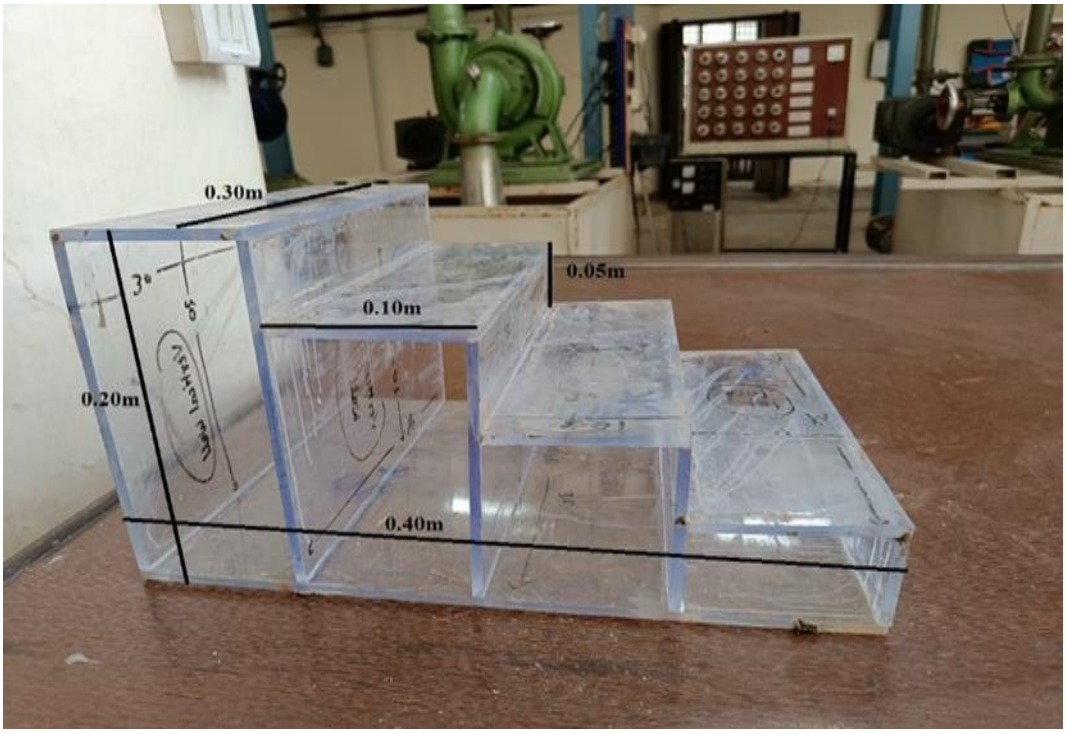


Figure 2. Dimensions of classical stepped spillway

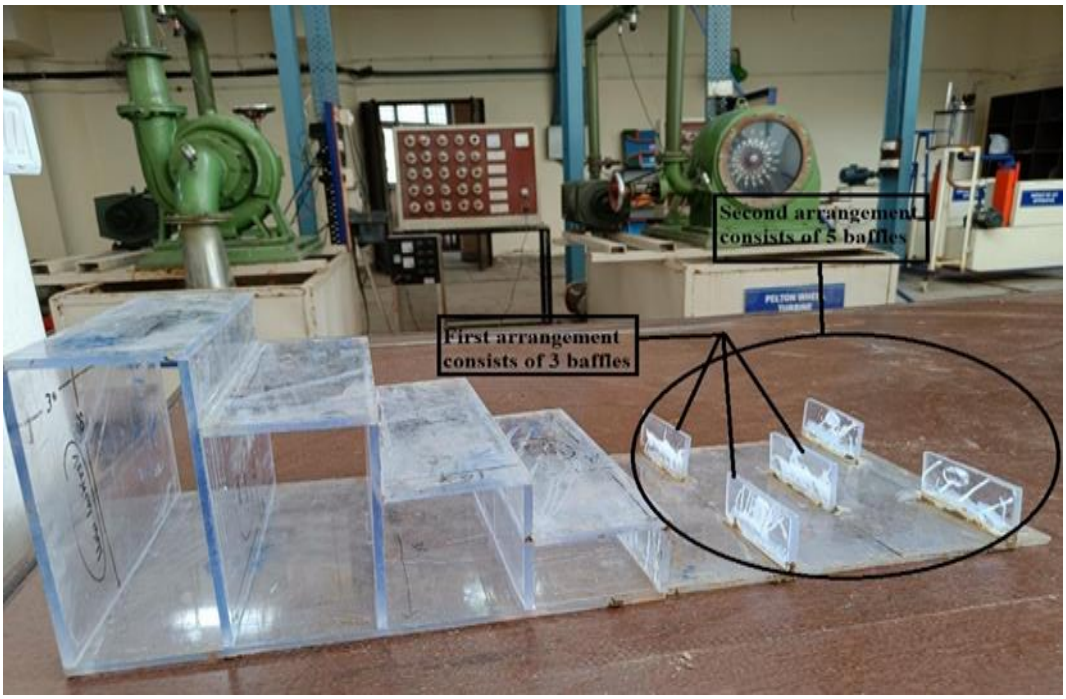


Figure 3. Arrangements of baffles in classical stepped spillway

(a) (b)

Figure 4. Rectangular stepped spillway with (a) three baffle arrangement (b) five baffle arrangement

2.3. Support Vector Machine (SVM) Model

Support vector machines, or SVMs, are a group of supervised learning algorithms that can be used for classification and regression. They offer superior accuracy in a wider range of settings, including those unique to the environment. In SVR, a fixed (nonlinear) mapping is used to map the input x into an m-dimensional feature space. This mapping does not account for any of the m-dimensional features. A linear model is then developed in this feature space using the gathered data and applied to the data [20]. To turn a linear classifier into a non-linear classifier, a non-linear function can be used to transfer the data from the input space x to a feature space F. Alternatively, a non-linear function can be used to perform the mapping. The discriminating function applicable to space F is represented by the following expression:

$$f(x) = w^T\varphi(x) + b \quad (7)$$

In the feature space, the linear model, which is indicated by the mathematical notation f (x, w), may be expressed in the following way:

$$w = \sum_{i=1}^{n} \alpha_i x_i \quad (8)$$

$$f(x,w) = \sum_{j=1}^{n} \alpha_i x_i \varphi_i(x) + b \quad (9)$$

$$f(x) = \sum_{i=1}^{n} \alpha_i x_i^T x + b \quad (10)$$

In the feature space, F, this expression takes the form:

$$f(x) = \sum_{i=1}^{n} \alpha_i \varphi(x_i)^T \varphi(x) + b \quad 0 \le \alpha_i \le C \quad (11)$$

$$k(x,x') = \varphi(x)^T \varphi(x') \quad (12)$$

$$f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i) + b \quad (13)$$

The radial basis kernel function was discovered to be the most efficient kernel function in the SVM, despite the fact that the SVM comprises a huge number of kernel functions. On the other hand, there are a variety of additional useful kernel functions that may be implemented in a variety of contexts.

Linear kernel: $k(x_i, x_j) = x_i^T x_j$

Polynomial kernel: $k(x_i, x_j) = (\gamma x_i^T x_j + r)^d,\ \gamma > 0$

RBF kernel: $k(x_i, x_j) = exp(-\gamma \|x_i - x_j\|)^2,\ \gamma > 0$

Sigmoid kernel: $k(x_i, x_j) = tanh(\gamma x_i^T x_j + r),\ \gamma > 0$

The kernel parameters C, γ, r, and d are utilized in SVM. The accuracy of the estimation, which determines the generalization performance of the SVM, is directly related to the quality of the settings for the meta-parameters C, γ, and r, as well as the kernel parameters. The difficulty associated with the prediction (regression) model is determined by the choices made for C, γ, and r. It is challenging to select the optimum parameters, as the complexity of the SVM model (and its effectiveness in terms of generalization) is dependent on all three parameters. Kernel functions are used to decrease the dimensionality of the input space for classification. The SVM model is built by exposing it to different data sets, resulting in increased accuracy. The energy dissipation over a rectangular stepped spillway was determined using high-quality data sets from the present study and Nasralla [18]. The acquired data is separated into training and test sets in MATLAB [21] prior to building the SVM model.

2.4. Regression Tree (RT) Model

Regression Trees is a machine-learning algorithm used to build prediction models of datasets, utilizing a clustering tree with post-pruning processing. The clustering tree algorithm is also known as the forecasting clustering tree or the monothetic clustering tree [22]. Regression Trees are employed for modeling dependent variables with a finite number of values that are not arranged in order. The prediction error is commonly assessed by the squared difference between predicted and observed values [23]. The clustering tree algorithm is based on the decision tree's top-down induction technique [24]. The algorithm takes a collection of data for training and creates a new internal node as well as possible. The system chooses the top test scores based on their decreased variance. The more homogeneous the cluster is, the lower the variance is, and the more accurate the forecast is. If none of the tests significantly reduce variance, the program generates a leaf and marks it as data representative [22]. A hierarchical tree-like division of the input space can be created by recursively splitting the data space and fitting a prediction model within each partition [25]. A sequence of tests and decision nodes determines the path through the tree until it approaches a terminal node, given a test data point, and a prediction is made at the terminal node based on the locally linked model. The tree consists of internal decision nodes and terminal leaves, with recursive splits used to divide the input space into local areas and denote the boundaries of these local regions.

2.5. Statistical Measures

To analyze the accuracy of the SVM and RT models, different types of error analysis were conducted using standard equations [26]. Various measures such as coefficient of determination (R2), mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean squared error (RMSE) were used. The Wilcoxon signed-rank test, a non-parametric statistical hypothesis test, was employed to evaluate the differences between two distributions using two matched samples. This test was also used to compare the accuracy of the newly constructed models with earlier methods.

3. Results and Discussion

Figure 5 represents the variation of energy loss with the flow rate for different channel slopes and baffle locations. The energy loss increases at a zero- and one-degree channel slope as the discharge increases for different baffle locations. At zero-degree channel slope, energy loss increases from 35 to 50% as the baffle increases from zero to five. At a one-degree channel slope, energy loss increases from 20 to 40% as baffles increases from zero to five. The energy loss is found to be more at a flat channel slope as compared to energy loss at a one-degree channel slope due to more energy provided by the channel slope to the flow. Figure 6 depicts the variation of relative energy loss with the relative spillway height. The relative energy loss first decreases reaching a minimum, then increases as the relative height increases. More energy loss is observed for five baffle cases. The same trend of variation is observed for a one-degree channel slope. Figure 7 shows the variation of energy loss with Froude number for different baffle locations and channel slopes. As the Froude number increases, the energy loss increases and the increment is found to be more for the five baffles

arrangement. The same trend of variation is observed at a one-degree channel slope with some fluctuations. Figure 8 represents the energy loss change with the upstream and downstream heads. As the head values increase, the energy loss increases nonlinearly. The increment is found to be steeper at a one-degree channel slope. The same trend has been observed for different baffle arrangements.

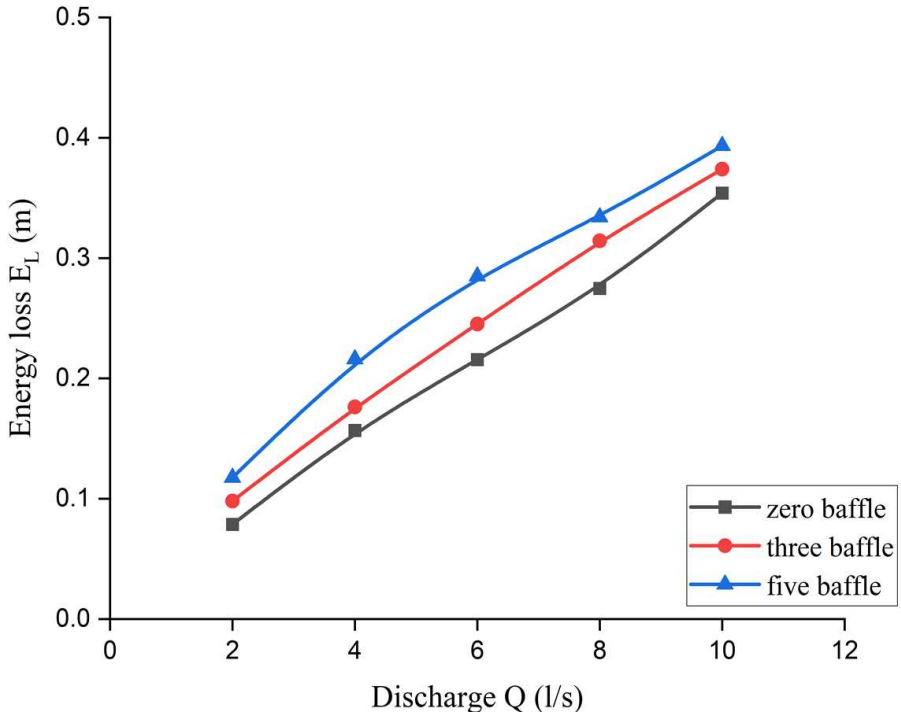
Figure 9 shows the scatter plots that compare the expected and actual relative energy loss for each of the several ML approaches. The fact that the values are so close to the line that indicates excellent agreement is a powerful sign of the capability of the SVM and RT models that were developed to make correct predictions. Comparing the two models, the SVM predicts values that are far closer to the line that provides the best fit than RT does, with scatter values that are distributed over the fitted line. As a result, the SVM model most closely corresponds with the empirical findings, as measured by high R^2 values. The various developed models for estimating relative energy loss are compared in Figure 10. When compared to other methods, it has been seen that the created SVM model comes quite near to the line that best fits the data, and it also has a significant potential for generalization. They do not display any of the symptoms that are associated with the condition of overtraining. The R^2 value, MSE value, RMSE value, and MAE value are some of the statistical characteristics that were used to evaluate the performance of the developed models with the previous approaches such as Nasralla [18] method, as shown in Table 1. It can be seen from the table that SVM (R^2 = 0.98 and RMSE = 0.032) gives the best-predicting results for the relative energy loss in comparison to other models such as RT (R^2 = 0.94 and RMSE = 0.048) and other methodologies such as Nasralla [18] method (R^2 = 0.84 and RMSE = 0.098). This is because SVM has a higher correlation and a lower relative standard error. It turns out that the SVM model with the lowest MAE is the best-suited technique for predicting energy loss across a rectangular stepped spillway. Table 2 represents the Wilcoxon two tailed analysis for the different approaches. In the test, null hypothesis was that median of the two distributions are equal, and alternate hypothesis was that median was either greater or lesser of the two distributions. It was observed that at the significance level of 0.05, null hypothesis cannot be rejected and hence the two distributions are not significantly different. A lower p-value in the Wilcoxon test reflects a value that is more significantly different from the observed values. From the Wilcoxon test analysis as shown in the table, the p-value is higher for the SVM model as compared to the RT model and Nasralla method, showing that the SVM predicted values are closer to the observed values for relative energy loss.

Table 1. Error analysis of predicted relative energy loss by machine learning approaches

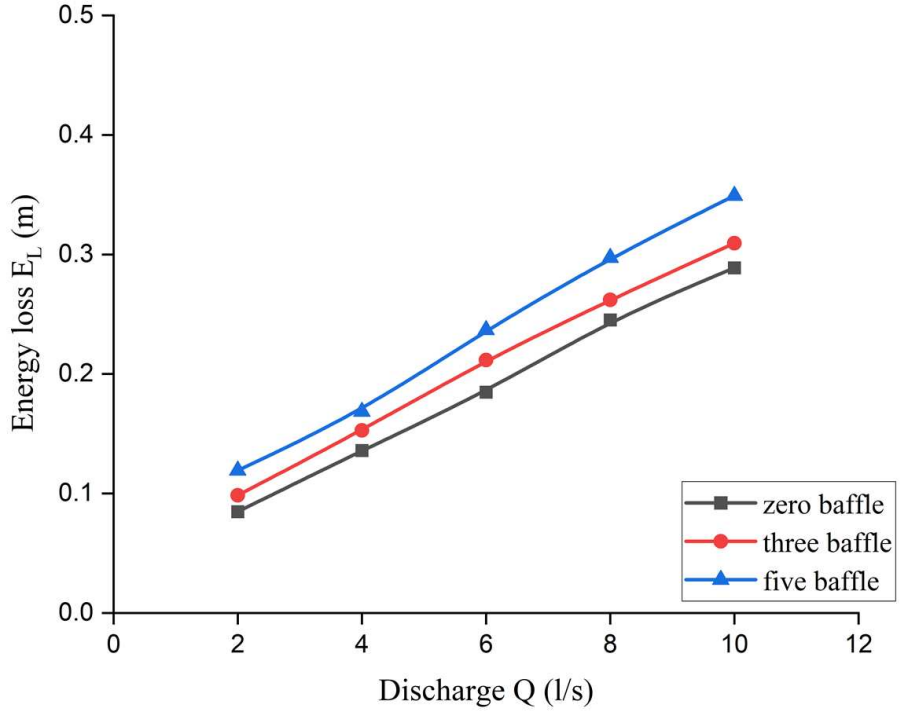
Statistical parameters	SVM model	RT model	Nasralla method
R^2	0.98	0.94	0.84
MSE	0.0009	0.0023	0.0097
RMSE	0.032	0.048	0.098
MAE	0.024	0.028	0.079

Table 2. Wilcoxon test analysis for different approaches

Wilcoxon test parameters		SVM model	RT model	Nasralla method
Descriptive statistics	Sample size	120	120	120
	Min	0.0234	0.0097	-0.6474
	Median	0.2060	0.2704	0.1235
	Max	0.6786	0.6492	0.7551
Positive ranks	Number (N)	65	64	65
	Mean rank	62.077	57.5	49.123
	Sum rank	4035	3680	3193
Negative ranks	Number (N)	55	56	55
	Mean rank	58.636	63.93	73.945
	Sum rank	3225	3580	4067
Test statistics	Test statistic (W)	3225	3580	4067
	Significance level	0.05	0.05	0.05
	Z-value	-0.1309	-1.0606	-1.1444
	p-value	0.897	0.289	0.254

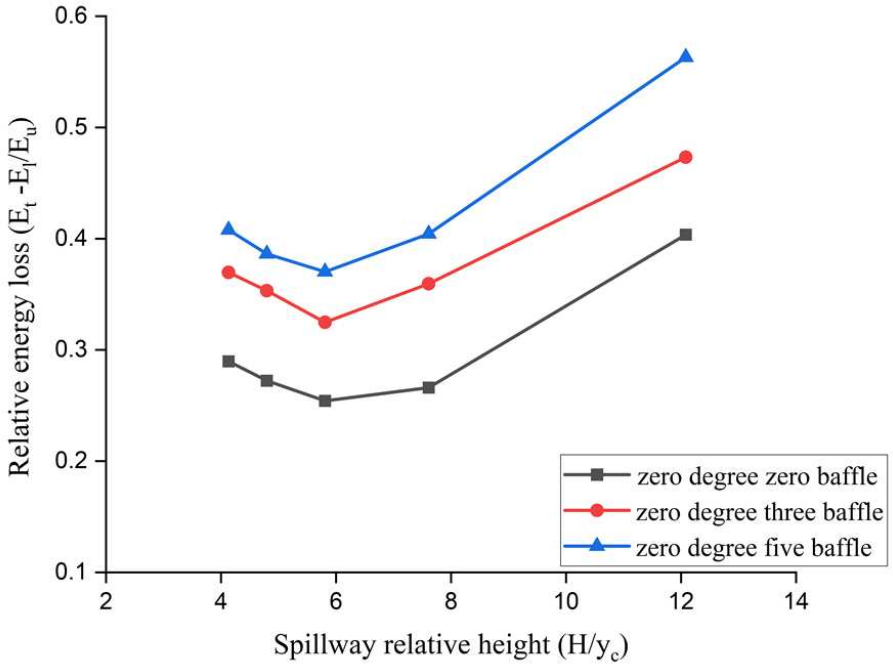


(a)

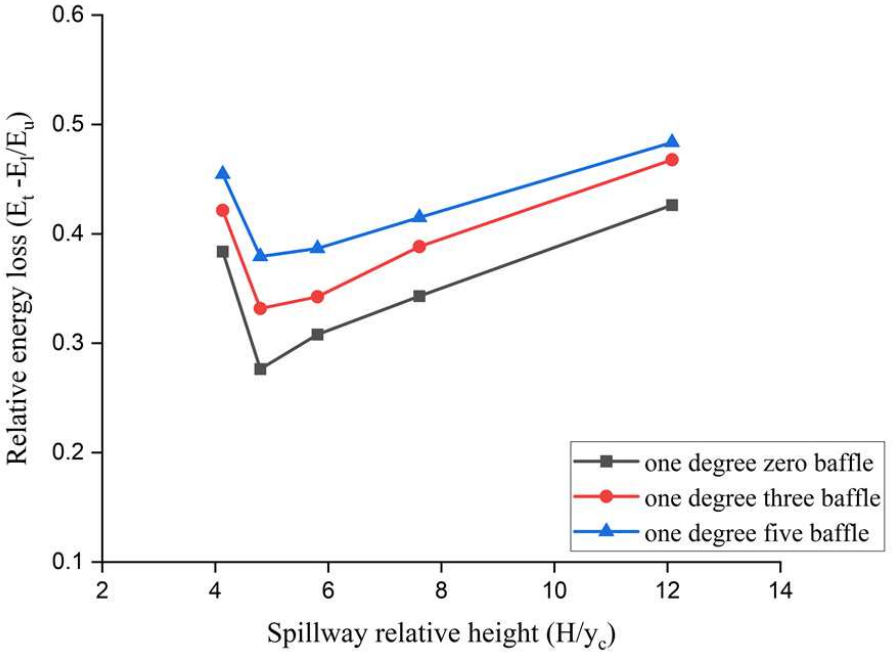


(b)

Figure 5. Variation of energy loss with discharge for (a) 0° slope (b) 1° slope



(a)



(b)

Figure 6. Variation of relative energy loss with the spillway relative height for different baffle locations and channel slopes

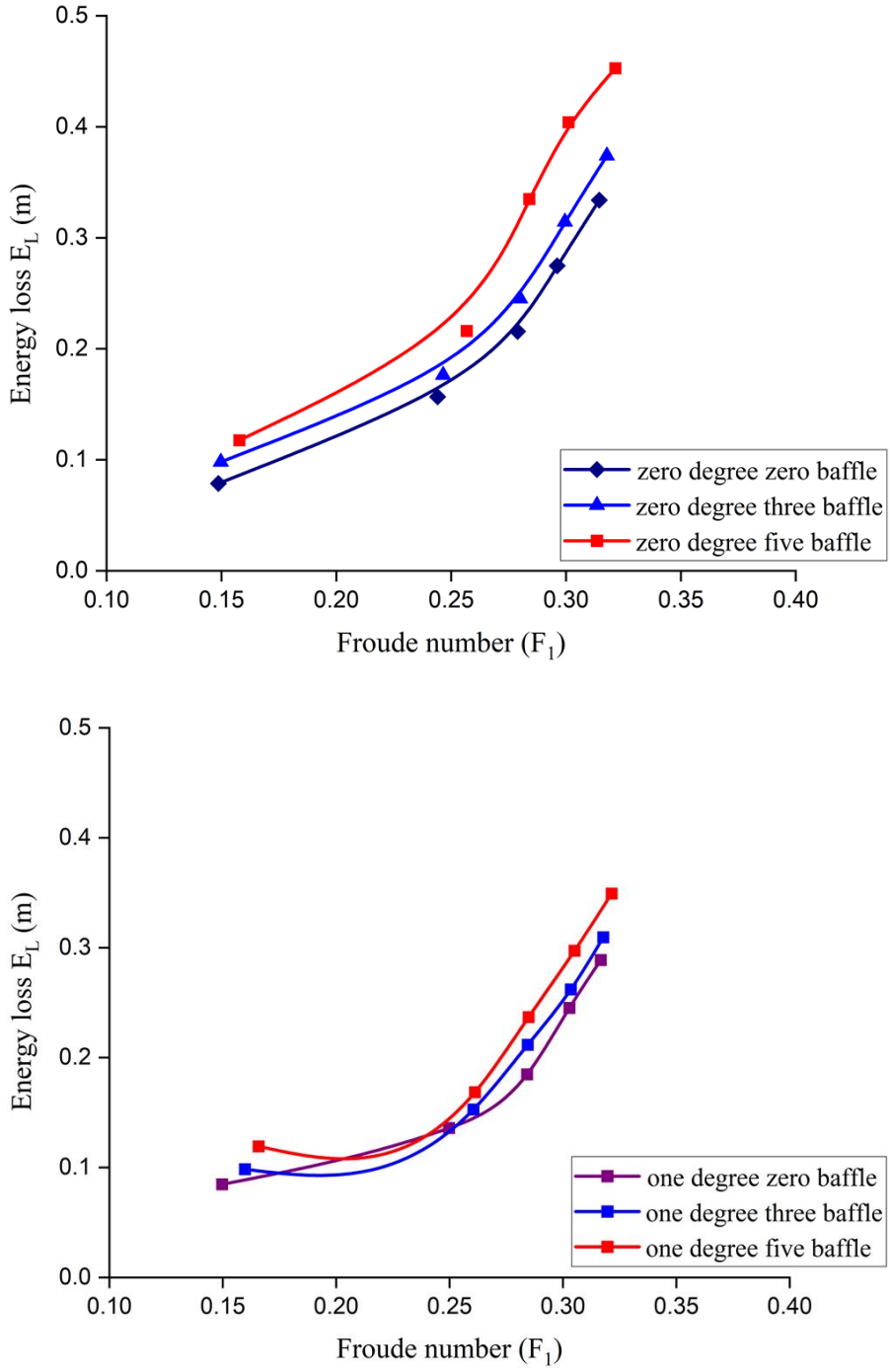
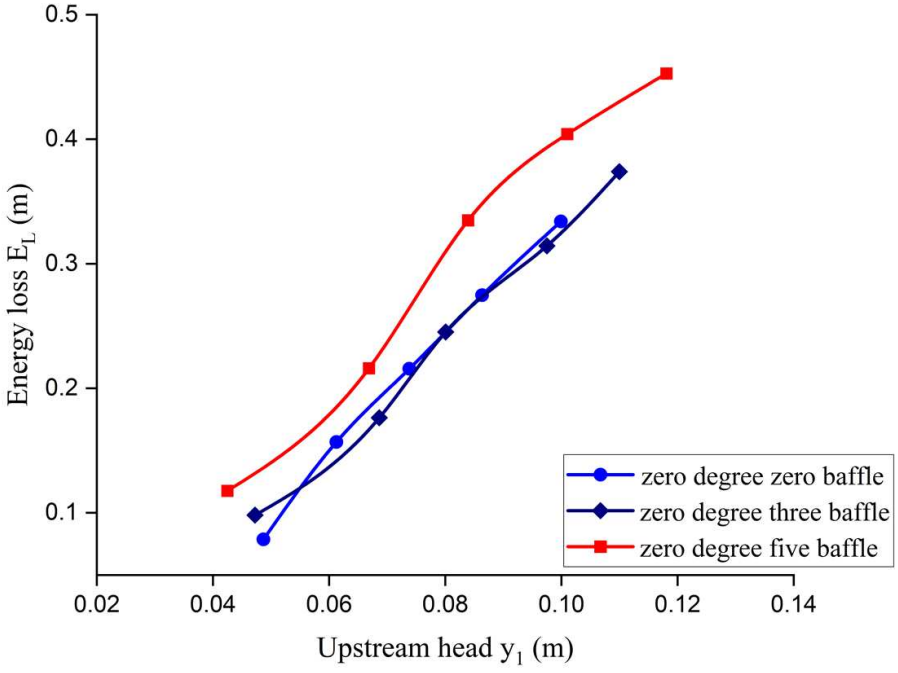
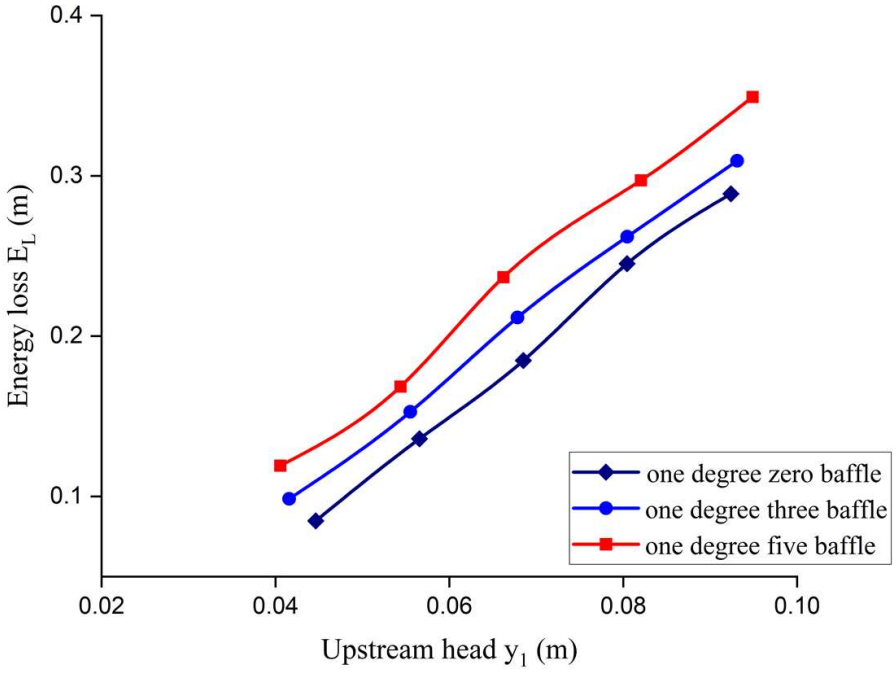


Figure 7. Variation of energy loss with Froude number for different baffle locations and channel slopes


0.5
0.4
0.3
0.2
0.1
Energy loss E_L (m)
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Upstream head y_1 (m)
zero degree zero baffle
zero degree three baffle
zero degree five baffle


0.4
0.3
0.2
0.1
Energy loss E_L (m)
0.02
0.04
0.06
0.08
0.10
Upstream head y_1 (m)
one degree zero baffle
one degree three baffle
one degree five baffle

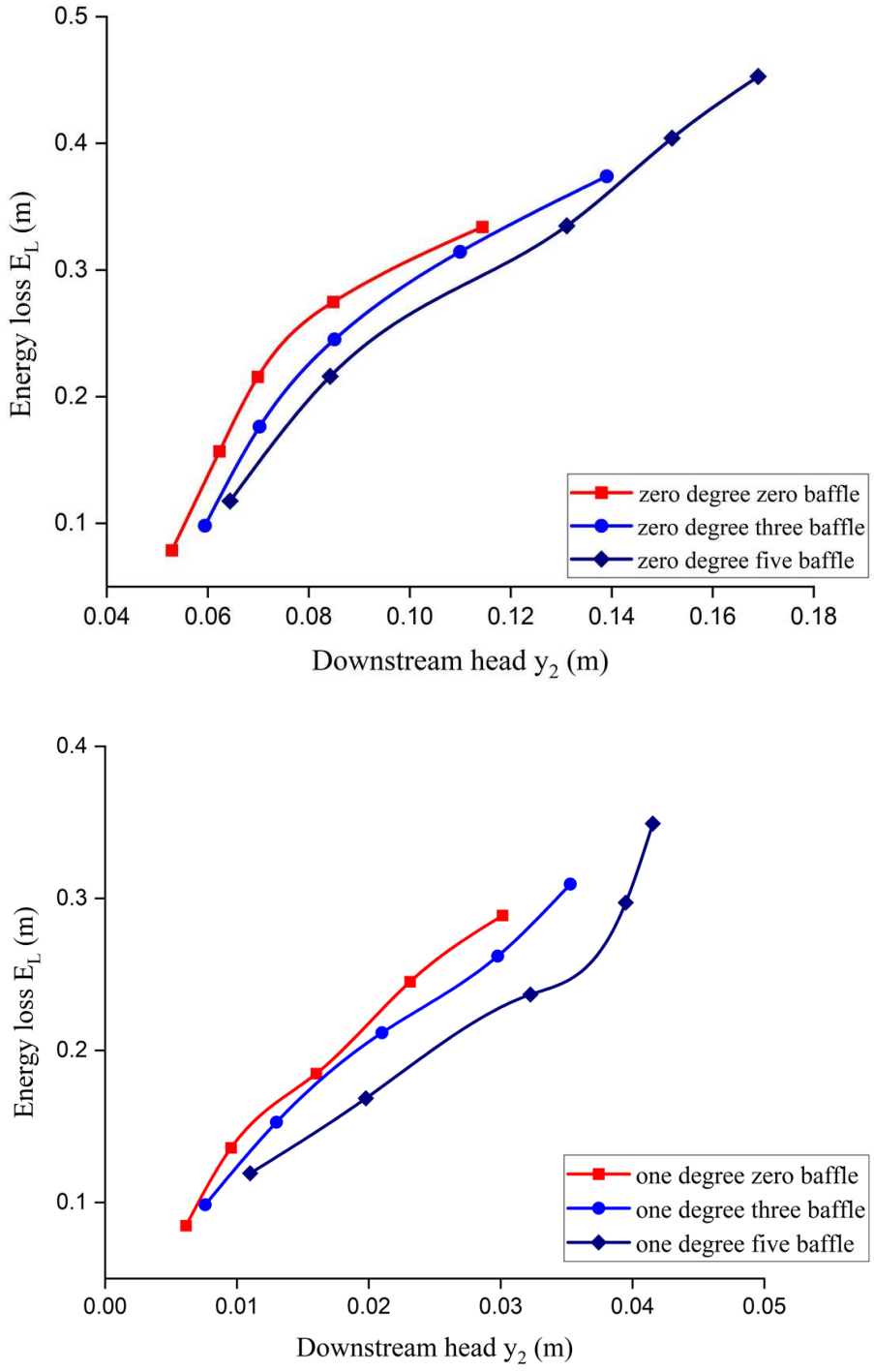
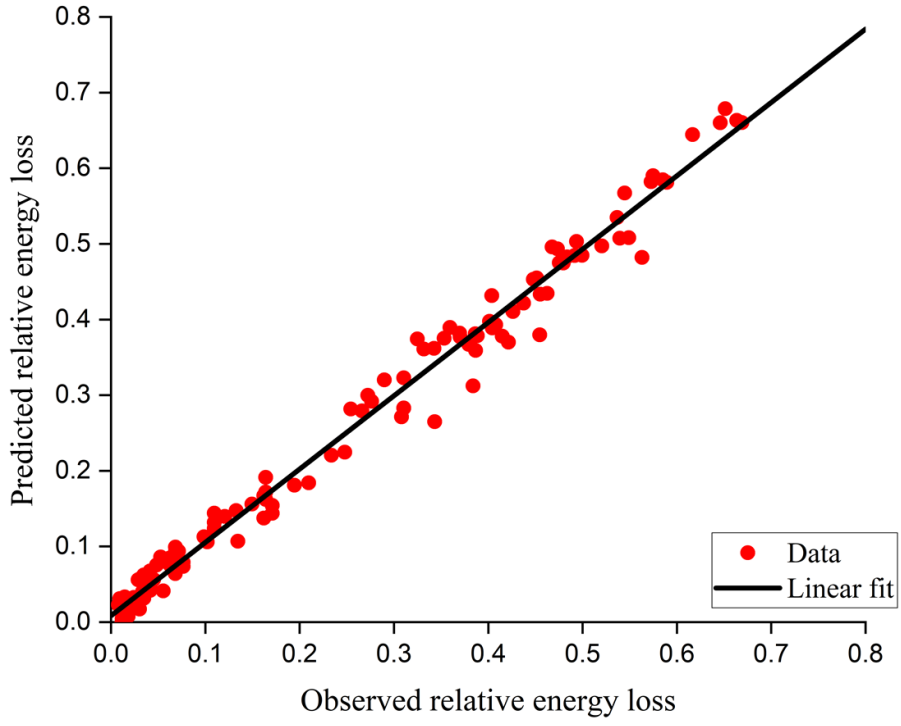
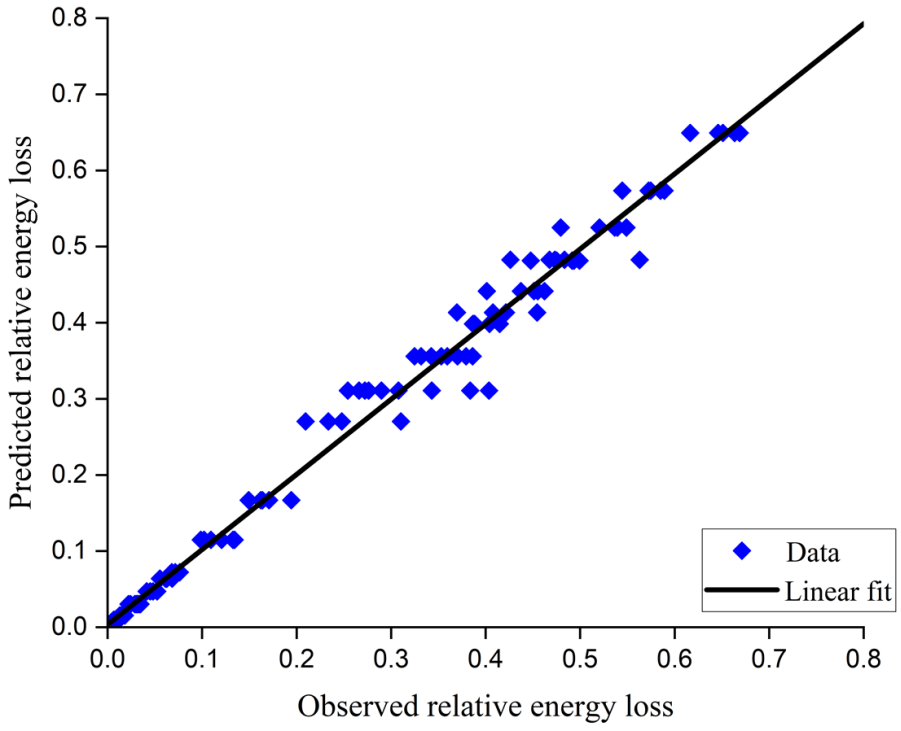


Figure 8. Variation of energy loss with upstream and downstream head for different baffle locations and channel slopes



(a)



(b)

Figure 9. Scatter plots of observed and predicted relative energy loss for various ML models (a) SVM (b) RT

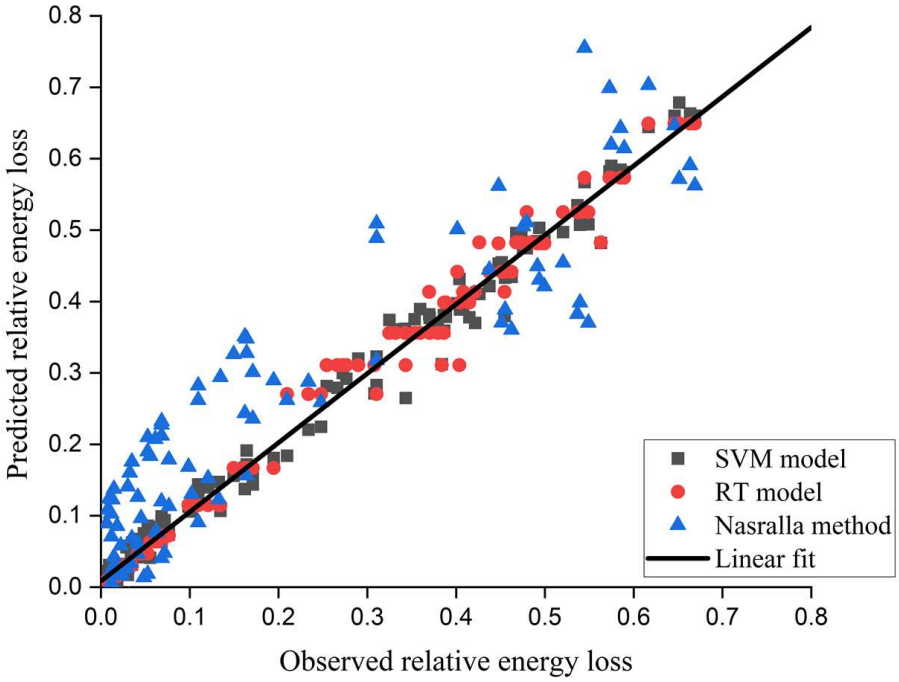


Figure 10. Comparison of predicted value of energy loss for different models

4. Conclusions

This study demonstrates the application of machine learning methods, specifically Support Vector Machine (SVM) and Regression Tree (RT), for computing the energy dissipation over a rectangular stepped spillway. The models were developed using laboratory datasets of superior quality, encompassing dimensionless geometric and flow features for different baffle configurations and channel slopes ($\theta = 0°$ and 1°). The research drew several findings, such as:

The suggested model seems to be affected by a wide variety of characteristics, such as the width and height of the spillway, the placement of the baffle, the heads upstream and downstream, the slope of the channel, and the Froude number. The flow rate at various channel slopes and baffle arrangements both have an effect on the amount of energy that is lost. It was discovered that the layout with five baffles caused a greater loss of energy when the channel slope was zero degrees. The relative height of the spillway, the heads both upstream and downstream, and the Froude number all contribute to an increase in the amount of energy that is lost. The same general pattern of variance has been seen over a wide variety of channel slopes and baffle configurations. An investigation was conducted to investigate the relationship between the relative energy loss and the nondimensional factors of a rectangular stepped spillway. It has been discovered that there is a link between each of the factors that are not linear in nature. In comparison to older methods such as Nasralla [18], the newly constructed models provide superior outcomes in terms of R^2, RMSE, MSE, MAE and Wilcoxon tests for a variety of datasets. According to the results, both SVM and RT were able to estimate with a reasonable degree of accuracy the amount of energy that would be lost across the rectangular stepped spillway. This was in compliance with the evaluation criteria. Because it had the best R^2 and the lowest RMSE, MSE, MAE and higher p-value from the Wilcoxon test, the SVM model demonstrated superior performance. The cost analysis of a stepped spillway compared to other types of spillways depends on several factors, including the design flow rate, the hydraulic head, the local construction costs, the availability of construction materials, and the environmental impact. While stepped spillways may be more expensive to construct initially, their potential benefits in terms of energy dissipation, reduced scour potential, and improved flow control may lead to long-term cost savings over other types of spillways. One limitation of the model is that it can only be used to predict how much energy would be lost via the usage of a rectangular, graded spillway that has baffles. Estimating the amount of energy lost across the various shapes of stepped spillways using novel approaches and baffle configurations should be the primary focus of research to be conducted in the future.

5. Notation

The following symbols are used in this paper:
H = height of spillway
y_c = critical depth of flow
a = distance of baffle from the toe of spillway
y_1 = upstream head
y_2 = downstream head
h_b = height of baffle

w_b = width of baffle
W = width of spillway
V = velocity of flow
g = acceleration due to gravity
H_u = hydraulic depth
E_u = Energy at upstream
E_d = Energy at downstream
E_L = Energy loss
E_T = Total energy
V_1 = Upstream velocity
V_2 = Downstream velocity
a = actual value
p = predicted value
$\bar{a}$ = mean of actual values
$\bar{p}$ = mean of predicted values
N = number of data
H/y_c = relative spillway height
a/y_2 = relative baffle distance
h_b/y_1 = relative baffle height
w_b/W = relative baffle width
F_1 = Froude number at upstream
$(E_T-E_L)/E_u$ = relative energy loss

Acknowledgements

The authors would like to express their gratitude to the Department of Civil Engineering at Delhi Technological University in Delhi, India, for their assistance.

REFERENCES

[1] Chanson H., "The hydraulics of stepped chutes and spillways," Can. J. Civ. Eng., 29(4), 634, 2002. doi: 10.1139/l02-061

[2] Rajaratnam N., "Skimming flow in stepped spillway," J. Hydraul. Eng., 116(4), 587–691, 1990. doi: 10.1061/(ASCE)0733-9429-(1990)116:4(587)

[3] Chanson H., "Comparison of energy dissipation between nappe and skimming flow regimes on stepped chutes," J. Hydraul. Res., 32(2), 213–218, 1994. doi:10.1080/00221686.1994.10750036

[4] Pegram G. S., Officer A. K., Mottram S. A., "Hydraulic of skimming flow on modeled stepped spillways," J. Hydraul. Eng., 125(5), 500–510, 1999. doi:10.1061/(ASCE)0733-9429(1999)125:5(500)

[5] Boes M., Hager H. F., "Hydraulic design of stepped spillways," J. Hydraul. Eng., 129(9), 671, 2003. doi:10.1061/(ASCE)0733-9429(2003) 129:9(671)

[6] Barani G., Rahnama M., Sohrabipoor N., "Investigation of flow energy dissipation over different stepped spillways," J. Am. Appl. Sci., 6, 1101–1105, 2005.

[7] Hazzab A., Chafic C., "Experimental investigation of flow and energy dissipation in stepped spillways," Larhyss J., 2006(5), 91–104, 2006. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.403.5207&rep=rep1&type=pdf

[8] Stefan F., Chanson H., "Energy dissipation, flow resistance and gas-liquid interfacial area in skimming flows on moderate-slope stepped spillways," Environ. Fluid. Mech., 9(4), 427–441, 2009.

[9] Daniel B., "A comparative study of self-aerated stepped spillway and smooth invert chute flow: The effect of step-induced macro-roughness," Chinese-German joint symposium on hydraulic and ocean engineering (CG JOINT 2010), Y. Wang, ed., Tianjin University Press, Tianjin, 451–456, 2010.

[10] Katourany S., "Experimental investigation the effect of space and size of baffles on energy dissipation in baffled apron drop," Arch. Sci., 65(6), 278–287, 2012.

[11] Salmasi F., Chamani M. R., Zadeh D. F., "Experimental study of energy dissipation over gabion spillways with low heights," IJST Trans. Civ. Eng., 36(C2), 253–264, 2012.

[12] Rad I. N., "An investigation of energy dissipation in various types of stepped spillways including inclined steps and steps with end sills by numerical model," Commun. Adv. Comput. Sci. Appl., 1–12, 2014. DOI: 10.5899/2014/cacsa-00030

[13] Saedi A., Asareh A., "Impact of the number of drop stairs on energy dissipation in stepped drops," Bull. Env. Pharmacol. Life Sci., 4(1), 29–35, 2014.

[14] Al-Husseini T. R., "Experimental study of increasing energy dissipation on stepped spillway," J. Kerbala Univ., 13(3), 87–100, 2015.

[15] Parsaie A., Haghiabi A. H., Saneie M., Torabi H., et al., "Prediction of energy dissipation on the stepped spillway using the multivariate adaptive regression splines," ISH J. Hydraul. Eng., 22(3), 281–292, 2016. doi: 10.1080/09715010.2016.1201782

[16] Frederic B. M., Roger T., Benoit N., Austin A.D.N., "Energy dissipation regimes and stability of the overflow dam (Spillway) for the Mekin dam in Cameroon," Int. Refereed J. Eng. Sci. (IRJES), 6(6), 28–34, 2017.

[17] Mojtahedi A., Soori N., Mohammadian M., "Energy dissipation evaluation for stepped spillway using a fuzzy inference system," SN Applied Sciences, 2(8), 2020. https://doi.org/10.1007/s42452-020-03258-0

[18] Nasralla T. H., "Energy dissipation in stepped spillways using baffled stilling basins," ISH Journal of Hydraulic Engineering, 28(3), 243–250, 2021.https://doi.org/10.1080/09715010.2021.1871786

[19] Ikinciogullari E., "Energy dissipation performance of the trapezoidal stepped spillway," Journal of Engineering Research, 2021. https://doi.org/10.36909/jer.13649

[20] Parsaie A., Yonesi H. A., Najafian S., "Predictive modeling of discharge in compound open channel by support vector machine technique," Model. Earth Syst. Environ., 1 (1–2), 2015. doi: 10.1007/s40808-015-0002-9

[21] MATLAB R., [Computer Software], MathWorks, Natick, MA, 2019a.

[22] Vens C. et al., "Inductive Databases and Constraint-Based Data Mining," 365–387, 2010.

[23] Loh W. Y., "Classification and regression trees," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., 1(1), 14–23, 2011.

[24] Quinlan J., "Induction of decision trees," Mach. Learn., 1, 1–81, 1986.

[25] Breiman L., "Classification and Regression Trees," Routledge, London, 2017.

[26] Naik B., Kaushik V., Kumar M., "Water surface profile in converging compound channel using gene expression programming," Water Supply, 22(5), 5221–5236, 2022. https://doi.org/10.2166/ws.2022.172

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 19 January 2023, accepted 5 February 2023, date of publication 13 February 2023, date of current version 17 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3244393

RESEARCH ARTICLE

Pronunciation Scoring With Goodness of Pronunciation and Dynamic Time Warping

KAVITA SHEORAN[1], ARPIT BAJGOTI[1], (Student Member, IEEE), RISHIK GUPTA[1], NISHTHA JATANA[1], GEETIKA DHAND[1], CHARU GUPTA[2], PANKAJ DADHEECH[3], UMAR YAHYA[4], (Member, IEEE), AND NAGENDER ANEJA[5]

[1]Department of Computer Science and Engineering, Maharaja Surajmal Institute of Technology, New Delhi 110058, India
[2]Department of Computer Science and Engineering, Bhagwan Parhsuram Institute of Technology, New Delhi 110089, India
[3]Department of Computer Science and Engineering, Swami Keshvanand Institute of Technology, Management and Gramothan, Jaipur, Rajasthan 302017, India
[4]Department of Computer Science and Information Technology, Faculty of Science, Islamic University in Uganda, Kampala, Uganda
[5]School of Digital Science, Univerisiti Brunei Darussalam, Gadong BE1410, Brunei Darussalam

Corresponding author: Umar Yahya (umar.yahya@iuiu.ac.ug)

ABSTRACT The current pronunciation scoring based on Goodness of Pronunciation (GOP) uses posterior probabilities of the Acoustic Models. Such algorithms suffer from generalization since they are utilized to determine a score metric for each phoneme rather than on the completeness or comparison with the ideal utterance of the words. This paper proposes a novel method to overcome such limitations by using combined scores of prosodic, fluency, completeness, and accuracy. This is achieved using context-aware GOP in conjugation with dynamic time warping (DTW) matching of the pitch contours of a weighted average of the context tokens found in the audio file that is rich in mispronounced phonemes. The proposed work gives flexibility in tuning the results according to different speech aspects based on a single hyperparameter. The results achieved are encouraging and have been validated on the speechocean762 dataset, where Automatic Speech Recognition (ASR) model has been trained on the Librispeech dataset. The resultant mean error of the proposed approach is 3.38% and the value of the correlation coefficient achieved is 0.652.

INDEX TERMS Pronunciation scoring, goodness of pronunciation, hidden Markov model-deep neural network, dynamic time warping, kaldi, speechocean762.

I. INTRODUCTION

English is a global language with complex grammatical literature. Since this language is used in so many documentaries and information exchanges, a lot of companies and institutions need their employees to be able to speak it. This language's acquisition and understanding have become a common issue, mainly focusing on the utterance of the words, which doesn't impair communication from the perspective of the speakers or the listeners in any way [1]. Teaching this language has many factors that make it difficult for teachers to efficiently teach the learners due to the following reasons:

- Teachers concentrate on improving the grammatical and vocabulary of learners so much that they become proficient in language reading and understanding skills rather than pronunciation [2].
- Learners are often discouraged as their teacher is much more proficient in the language and pronounces it differently [3].
- Continuous monitoring of the learner's pronunciation growth is complicated to achieve.

Computer Aided Pronunciation Teaching (CAPT) overcomes such disadvantages by developing a versatile model that concentrates more on the fluency and prosodic aspects of the speech, allows for easy progress monitoring, and provides automatic pronunciation scoring at any time. To help non-native English speakers, algorithms have been built for quick self-assessment of their performance in phoneme pronunciation, stress, and fluency without relying on a second person. A deep learning approach to generating a result that can be evaluated based on the posterior probabilities of acoustic

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

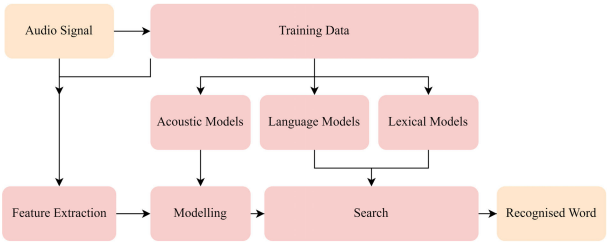


FIGURE 1. Block diagram of ASR System.

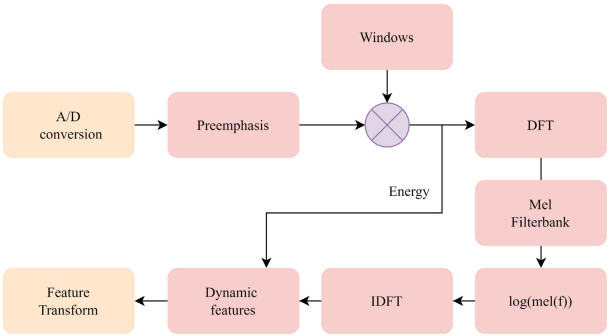


FIGURE 2. Steps involved in MFCC Feature Extraction.

models in ASR systems and the modified probabilities of the result describing the score for each phone of utterance will benefit the learner.

A. GOODNESS OF PRONUNCIATION

A measure of the effectiveness of speech recognition systems, specifically Automatic Speech Recognition (ASR) systems, is called Goodness of Pronunciation (GOP). By putting a number on how well an ASR system can detect and record speech, it can be used to gauge how well it performs. The GOP value is a numeric number between 0 and 1, where 1 denotes complete recognition and 0 denotes no recognition at all. This is also known as pronunciation accuracy, and it is often calculated as the ratio of the number of correctly detected phonemes to the total number of phonemes in the test speech signal. A detailed overview of GOP and its improved versions is described in related works [13].

B. AUTOMATIC SPEECH RECOGNITION SYSTEM

Speech recognition is when a computer or software can take phrases and words from spoken language and turn them into a format that a machine can understand. It is also known as computer voice recognition, speech-to-text, and automatic speech recognition.

Modules for automatic speech recognition consist of five stages.

- Audio Signal Detection
- Feature Extraction
- Acoustic Modelling
- Language and Lexical Modelling
- Training and Recognition

Fig. 1 shows the Automatic Speech Recognition System block diagram. The various blocks used in the ASR model are discussed in detail in the subsections below.

1) FEATURE EXTRACTION

The most crucial step in voice recognition is feature extraction since it separates one speech from another.

For speech recognition, the Mel-frequency cepstral coefficient (MFCC) is the most obvious and often used feature for the extraction method. The logarithmic placement of the frequency bands here resembles the human system reaction more than any other system. They are the result of a Mel-frequency spectral analysis in which the frequency bands of the Mel scale are evenly spaced. The MFCC computation method, which figures out the MFCC vector from each frame, is based on short-term analysis. Using the given formula in eq. 1, it is possible to calculate the MFCC.

$$Mel(f) = 1127 \log(1 + \frac{f}{700}) \tag{1}$$

Fig. 2 lists the steps in the feature extraction process. Firstly, the audio signal is converted from analog to digital format. Preemphasis increases the magnitude of energy in the higher frequency which improves phone detection. Then the signal is broken into multiple pieces(windows) to extract individual phones. The signal is then converted from the time domain to the frequency domain by applying Discrete Fourier Transform. The received signal is mapped according to human perceived levels using the mapping formula given in eq. 1 and logarithm function is applied to the Mel frequency output. The Inverse discrete Fourier transform (IDFT) block performs an inverse transform of the output from the previous step. The MFCC model takes in the first 12 coefficients of the signal after applying the IDFT operations, also taking the energy of the signal as a feature. Along with these 13 features, the MFCC technique will consider the first-order derivative and second-order derivatives of the features thus constituting a total of 39 features [4].

I-Vectors or identity-vectors are a more simplistic speaker recognition model that were presented in [5], they do away with the dichotomy between speaker and channel variability subspaces and describe both in the "total variability space," a common limited low-dimensional space. Particularly in ASR models, they are used to get the uniform dimension of every datapoint of audio feature [6].

2) ACOUSTIC MODELING

The foundational component of the ASR system is acoustic modeling [7]. In acoustic modeling, the relationship between the acoustic data and phonetics is formed. The acoustic model is vital to system performance and oversees the computational burden [8]. The relationship between the fundamental speech units and the acoustic observations is established by training. The system must be trained by using one or more patterns that match speech sounds from the same class to make a pattern

that represents the features of the class. Several models may be used for acoustic modeling. The Hidden Markov Model (HMM) is a popular technique since it is effective for both training and recognition [9].

a: HIDDEN MARKOV ACOUSTIC MODELLING

Hidden Markov Acoustic Model is a finite state Markov model where several output distributions serve as signals for a hidden Markov model. The output distribution model parameters are spectral variability whereas the alteration parameter in the Markov chain models is temporal variability. The capacity to recognize speech depends on these two forms of variability. Compared to a template-based method, hidden Markov modeling is more versatile and has a strong mathematical foundation. HMM makes it simpler to incorporate knowledge sources into structured architecture than the knowledge base method. A drawback of HMM is that it does not offer much insight into the recognition process. Analyzing system failure is done to enhance the performance of the HMM system, but it is a challenging task. However, careful knowledge inclusion has greatly enhanced the HMM-based system [11].

b: HMM-DNN BASED ACOUSTIC MODELLING

A Hidden Markov Model (HMM) and a Deep Neural Network (DNN) are combined to form an HMM-DNN model. The DNN is used to model the intricate patterns in the data, whereas the HMM is used to model the temporal relationships in the data. Due to its ability to combine the benefits of the HMM with the DNN, this model is highly suited for situations involving sequential data with intricate patterns [10] where the HMM-DNN model uses the HMM as prior knowledge to direct the training of a DNN and enhance voice recognition performance [16].

3) LANGUAGE MODELS

The likelihood of a word sequence is calculated by the Language Model (LM) component. By adding linguistic knowledge from big text corpora, LMs help acoustic models be more accurate. Implicitly learned syntactic and semantic criteria are employed by LMs to re-score the acoustic model hypotheses. Using a pronunciation dictionary, a series of phonemes is translated into words to match the phonetic transcriptions produced by the acoustic modeling with the raw text utilized in language models [12].

4) FORCED ALIGNMENT

Automatic speech recognition (ASR) uses a forced alignment technique to line up the words in transcription with the associated audio. A huge dataset of speech and transcripts is used to build an acoustic model as the first step in forced alignment. The acoustic model gains the ability to forecast the possibility that certain sounds will appear at various locations during the spoken stream. Once the model has been trained, it may be used to match the words in a transcription with the audio by identifying the sound sequence that, given the words in the transcription is most likely to have produced the audio.

In this paper, inspired by the posterior calculation of GOP and objective comparison of signals with DTW, we propose a new pronunciation scoring equation that combines both features using a hyparameter. In our approach, context-aware log posterior probabilities extracted by any Acoustic model of an ASR system are averaged together for each utterance of a speaker. Thus, the posterior features only represent the prosodic features of a speech. The DTW then extracts the fluency, and word level accuracy of the speech separately, The scores of both the extracted scores are added by a weighted average to get a total score. Compared with other acoustic model-based scoring, our model gave better results without even biasing or fine-tuning the acoustic model on the pronunciation dataset.

II. RELATED WORK

An algorithm developed by Witt et al. [13] used native acoustics modeling which relied on the Gaussian mixture model-hidden Markov model (GMM-HMM) to define the Goodness of Pronunciation and thus calculated a score from the formulated GOP. Improvements were made to this method by most of the works released afterward either by suggesting modifications to the GOP-based formulation or by enhancing the parameters of the original acoustic models. In the prior work of Zhang et al. [5], the score was calculated using scaled log posteriors instead of posteriors. Using the chosen state sequence acquired from the forward-backward technique, Luo et al. [14] formulated the GOP. The GOP concept put forward by Witt et al. [13] was employed by Wang et al. [15] with erroneous pattern detectors in the process of diagnosing phoneme mispronunciation, but these works only focused on the detection of the phoneme in a frame by forced alignment of phonemes and only took the prosodic score into consideration. Sudhakara et al. [16] introduced context-aware GOP which takes both senone and transition state probabilities into consideration. Ryu et al. [17] inferred that pronunciation scoring must combine phone level as well as articulatory-level diagnoses such as voicing, place of articulation, and manner of articulation on consonants. Lin et al. [18] used the acoustic model and replaced the forced alignment layer with a self-attention layer to get an utterance score based on transfer learning, but the results greatly depend on fine-tuning the scores of datasets. Cheng et al. [19] suggested the problems of GOP-based scoring and proposed an ASR-free scoring method based on I-vector, Normalization Flow (NF), and Discriminative Normalization Flow (DNF) for improvements in GOP and the better performance of the combination of both the approaches (GOP and ASR-free score). Bugdol et al. [20] stated that rather than checking the probability of finding the presence of phonemes in sequential order in an audio file, another approach is to use the Dynamic Time Warping (DTW) algorithm on the vectors of the amplitude or Mel-Frequency Cepstral Coefficients (MFCC) of the utterance in the frame between the ideal file to compare with and

TABLE 1. Comparing the related works.

Reference	Prosodic	Fluency	Completeness	Accuracy	Public Dataset
Witt et al. [13]	✓		✓		✓
Wang et al. [15]	✓		✓		✓
Sudhakara et al. [16]	✓		✓		✓
Ryu et al. [17]	✓	✓		✓	✓
Lin et al. [18]	✓	✓		✓	✓
Cheng et al. [19]		✓	✓	✓	✓
Bugdol et al. [20]	✓	✓		✓	✓
Permanasari et al. [22]		✓		✓	
Miodonska et al. [23]			✓	✓	
Karhila et al. [24]			✓	✓	✓
Chao et al. [35]	✓	✓	✓	✓	✓
Kim et al. [36]	✓	✓			✓

the file to test. The problem arises when the voice of these two speakers differs a lot in pitch and frequency, also the algorithm is quite slow for large audio files. An optimized version of this algorithm was proposed by Salvador et al. [21] called FastDTW which constraints the matching paths calculated by constraining the calculations on a band rather than the whole mesh grid. Permanasari et al. [22] suggested a method to reduce the Euclidian distance in DTW between audio files using the volume and different pronunciation times. Miodonska et al. [23] proposed a combined approach that uses phone-level mispronunciation detection using DTW, but the comparison approach always demands an ideal file to get a metric for distinction. Karhila et al. [24] used the Connectionist temporal classification (CTC) layer model to predict the phone sequence in the audio file and compared it to the ideal base file using Levenstein edit distance to calculate the difference between the sequence and predict the pronunciation score.

The former approaches discussed are based only on the pronunciation scoring based on the likelihood of individual phones in sequential order, hence is limited to phone-level features extraction of the audio file to test [13], [14], [15], [16], [17], [18], [19]. The latter checks for a minimum distance for comparing different length time series or MFCC or LPC for audio comparison [20], [21], [22]. Both approaches compute independent tasks, but the combined effect can be informative for the scoring of the overall quality of speech of a person's audio [23], [24]. A high-level score representing the percentage score is preferred by many speakers i.e., the overall grade they scored in terms of probability or percentage, therefore we integrated both these methods to get a score representing the presence of phones, accent, and prosodic scores and calculating the output as only a probability score. An overview of the methods discussed above is tabulated in table 1.

A. BASIC GOP FORMULATION

The GOP score is defined as the posterior probability of a target phone p, given a sequence of acoustic features O, from among a set of phones Q. The GOP score is heavily based on Bayesian statistics and was developed by Witt et al. [13].

$$GOP(p) = P(p|O) = \frac{P(O|p)P(p)}{\sum_{q\epsilon Q} P(O|q)P(q)} \tag{2}$$

B. REVISED GOP FORMULATION

Eight years after the introduction of the first GOP formulation by Witt et al. [13], in 2008 Zhang et al. [5] came up with the revised approach for GOP formulation which incorporated logarithmic scaling, using the absolute score value and normalization done with a number of frames (NF) for a specific phone.

$$GOP_R(p) = \left| \frac{P(O|p)P(p)}{\max_{q\epsilon Q} P(O^p|p)} \right| \tag{3}$$

C. DEEP NEURAL NETWORK GOP

The GOP devised by Hu et al. [25] replaced the Gaussian mixture model (GMM) with Deep Neural Network (DNN). This was achieved by replacing the use of likelihood probability $P(p|O)$, which was done in the face of GMM. Using a DNN in place of a GMM resulted in the enhanced capability for discrimination in the phoneme models.

However, using the DNN model to get acoustic features for GOP formulation has significant shortcomings, by using Multilayered Perceptron (MLP), structural locality arises since feature distance is not captured because of the fully connected characteristic of the DNN. Long short-term memory (LSTM) networks work better at modeling long-term dependencies of speech signals but due to dependencies between the time-frames being processed in an RNN, parallelization cannot easily be exploited to the same extent as in feedforward networks [38], [40], which greatly increases the training as well as inference time which is not preferred for real-time CAPT. Time Delay Neural Networks(TDNN) have been shown to effectively learn the temporal dynamics of the signal even from short-term feature representations and work better in GOP computing [16].

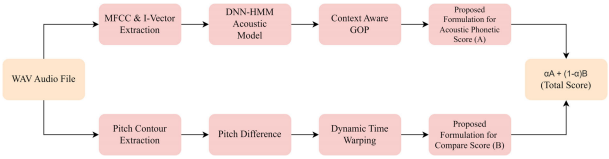


FIGURE 3. Pronunciation Scoring Framework.

III. PROPOSED APPROACH

This section explains the proposed workflow using context-aware GOP and DTW of pitch contours as well as the combined score for CAPT.

A. MODEL AND WORKFLOW

As shown in Fig. 3. The experimental setup is divided into two sections, section I contains the following components: the wav file, MFCC extraction, passing through a trained DNN-HMM Acoustic model, calculating the modified GOP scores, and calculating the posterior score from the proposed formula 7 and 8. Section II contains the same wav file as input, calculating the adjacent difference in pitch contour vectors of the same spoken words from the wav file as well as a comparison file from google text to speech API, then the vectors are passed through optimized DTW from which the compare scores are calculated, finally, the output is calculated as the weighted average of the 2 section scores. A detailed explanation of the two sections is discussed in section C.

Posterior Probability workflow is shown in Fig 4 [16], the DNN-HMM model posterior probabilities and senone probabilities are used to generate an equation for context-aware GOP, and the details are discussed in sections B and C.

B. POSTERIOR PROBABILITIES CALCULATION

The definition of GOP-NN is a bit different from the GOP-GMM. GOP-NN was defined as the log phone posterior ratio between the canonical phone and the one with the highest score [25]. Firstly, we define Log Phone Posterior (LPP):

$$LPP = \log P(p|O; t_s, t_e) \tag{4}$$

Then we define the GOP-NN using LPP:

$$GOP(p) = LPP(p) - \max_{q \epsilon Q} LPP(q) \tag{5}$$

LPP could be calculated as:

$$LPP(p) = \frac{1}{t_e + t_s + 1} \sum_{t=t_s}^{t_e} \log p(p|O_t) \tag{6}$$

$$p(p|O_t) = \sum_{s \epsilon p} p(s|O_t) \tag{7}$$

where s is the senone label, $s|s \in p$ are the states belonging to those triphones whose current phone is p. Normally the classifier-based approach achieves better performance than the GOP-based approach. Being different from other GOP-based methods, an extra supervised training process is needed. The input features for supervised training are phone-level and segmental features. The phone-level feature is defined as:

$$[LPP(p_1), \ldots, LPP(p_m), LPP(p_1|p_i), \ldots, LPP(p_j|p_i)]^T \tag{8}$$

where the Log Posterior Ratio (LPR) between phone p_j and p_i is defined as:

$$LPR(p_j|p_i) = \log \frac{p(p_j|O; t_s, t_e)}{p(p_i|O; t_s, t_e)} \tag{9}$$

This formula calculates the ratio of two phonemes and returns a value less than 0, where a high negative value indicates the audio segment of this phone should be a mispronunciation.

C. PROPOSED FORMULATION

The theory discussed in equations 4 to 9 set up the basis of the accumulated utterances' phone-level score, followed by the compare score and combined total score of the pronunciation.

1) PHONE-LEVEL FEATURE

The log phone posterior ratio accumulated for a specific test audio file is taken and let p be the vector representing the phone set in series and $LPR(p_j|p_{\max})$ defines the score that is the ratio of canonical phone and the one with the highest score in sequence with vector p as defined in eq. 9. Let $f(p_i) = LPR(p_i|p_{\max})$ then the overall phone score is formulated as:

$$S = \sum_{i=1}^{k} \frac{\sum_{j=1}^{m} f(p) 1\{p_j = p_i\}}{m} \frac{\sum_{j=1}^{m} f(p) 1\{p_j = p_i\}}{\max_{j \epsilon m}(f(p)) \sum_{j=1}^{m} 1\{p_j = p_i\}} \tag{10}$$

where k is the number of phonemes in the language, m is the size of the vector p and $1\{p_i{=}p_j\}$ is an indicator function that returns a value of 1 when the respective phones in the condition are matched. The value S defines the probability of bad pronunciation given the set of LPR scores for an audio file. This probability score is passed through a nonlinear function defined as:

$$p(O) = \frac{2e^{-3S}}{1 + e^{-3S}} \tag{11}$$

where s is the cumulative probability from eq. 10, eq. 11 and gives a probability score of goodness in the overall phone matching, this function doesn't allow the scores to be at the margins.

2) COMPARE SCORE

The audio file is passed through an ASR model to get the words spoken corresponding to the timestamps of each spoken word. Words that contribute to a significant amount regarding the context of the speech are randomly sampled, the sampled signal is represented as:

$$x = [x_1, x_2, \ldots, x_{A-1}, x_A] \tag{12}$$

where A is the sample count of the signal. The vector in eq. 12 is further processed as follows:

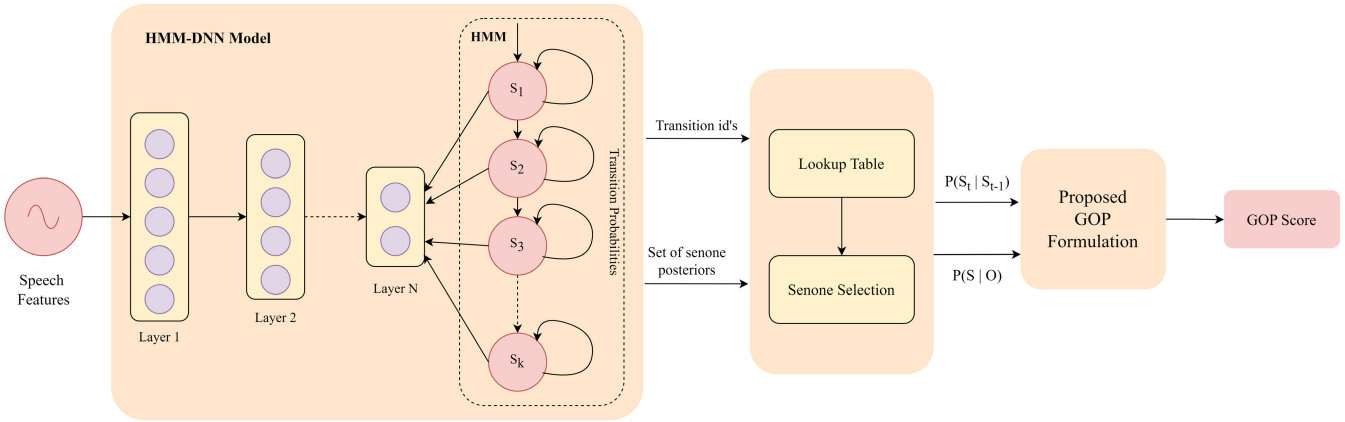


FIGURE 4. Posterior Score: Flow Diagram of GOP calculation using pre-trained weights of HMM-DNN acoustic model [16].

a: PITCH EXTRACTION

The modified Auto-Correlation function-based pitch detection is defined as:

$$r(\tau) = x \otimes x(\tau) = \sum_{i=q}^{q+N-1} x(i)x(i+\tau) \tag{13}$$

where q is the current frame, N is the frame length, τ is the lag index, and $x(i)$ is the current sample of the audio signal.

The main peak in the auto-correlation function is at the zero-lag location ($\tau = 0$). The location of the next peak gives an estimate of the period, and the height gives an indication of the periodicity of the signal, this estimate is given by:

$$\hat{f}_0 = \frac{1}{\tau_{max}}, r(\tau_{max}) = \max_{\tau} r(\tau) \tag{14}$$

where $\tau > 0$.

Applying equations 13 and 14 on the audio signal gives a frame-wise pitch array given by:

$$x = [\hat{f}_1, \hat{f}_2, \ldots, \hat{f_{n-1}}, \hat{f}_n] \tag{15}$$

where n is the number of frames in the ideal speech signal. The auto-Correlation method is used for pitch detection instead of other methods because of its better performance on audio speech signals elaborated in the discussion section. In order to compare the frames of the signals it is often more informative to analyze the overall transition of the pitch track than the absolute value. Therefore delta of the obtained frame-wise fundamental frequencies is taken as follows to obtain a new signal vector as follows:

$$\Delta_k = f_k - f_{k-1} \tag{16}$$

where f_k is any arbitrary frequency in the new array in eq. 14, the new array after applying eq. 15 gives:

$$\Delta x = [\Delta_0, \Delta_1, \ldots, \Delta_{n-1}, \Delta_n] \tag{17}$$

An ideally pronounced speech of the sampled audio Y is then generated using Google Text to Speech (GTTS) library and pitch is extracted similarly to the sampled audio in eq. 12 represented as:

$$\Delta y = [\Delta_0, \Delta_1, \ldots, \Delta_{m-1}, \Delta_m] \tag{18}$$

where m is the no. of frames in the ideal speech signal.

b: DTW MATCHING

Let W be the mesh grid of each point in Δx and Δy given by (i, j), then the FastDTW algorithm by Stan et al. [21] is used to define a warping path and corresponding DTW distance as:

$$D_{\min}(i_k, j_k) = \min_{i_{k-1},\, j_{k-1}} D_{\min}(i_{k-1}, j_{k-1}) + d(k, k-1) \tag{19}$$

where k is an arbitary point in the meshgrid W and d is the Euclidean distance given by:

$$d(i,\ j) = ||f(i) - f(j)||_2 \tag{20}$$

overall path cost:

$$D = \sum_{k}^{d} (i_k,\ j_k) \tag{21}$$

The path cost D is changed to a probability score using a threshold value for comparison i.e., 750, and passed through a smoothing function to change it in the scale of zero to one as:

$$p(d) = e^{\frac{t-d}{t}} \tag{22}$$

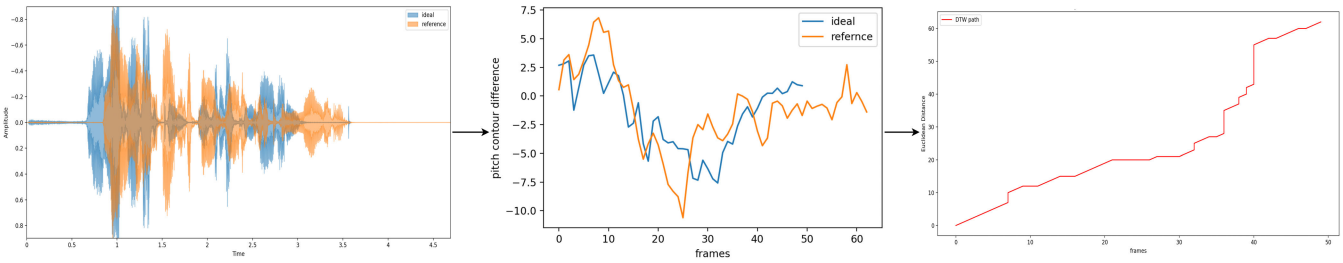
where $t = 750$ (threshold value) The flow diagram of the extraction of compare scores is shown in fig. 5.

3) TOTAL SCORE

Finally, the total score is calculated as a combination of phone feature score and audio pitch difference comparison score as a weighted average where the weights are hyperparameters that can be tuned by the examiner or learner according to the preference of one score over another as:

$$p(t) = \alpha p(o) + (1 - \alpha)\, p(d) \tag{23}$$

Here α is a hyperparameter and $0 < \alpha < 1$.



Waveforms of audio files | Adjacent pitch difference | L2 norm path of DTW

FIGURE 5. Compare Score: DTW of difference of pitch contours of ideal and test waveforms a) Waveform b) Pitch Contour Difference c) DTW with L2 norm metric.

IV. EXPERIMENTS

This section discusses the datasets used for the model and also discusses how the hyperparameter tuning was carried out.

A. DATASET

The speechocean762 speech corpus is designed for pronunciation assessment use, consisting of 5000 English utterances from 250 non-native speakers, where half of the speakers are children [26]. Five experts annotated each of the utterances at the sentence, word, and phoneme levels. The sentence-level scores by experts are used as the proposed equations calculate utterance level mispronunciation whereas high sentence-level scores rely on a consistently "good" word and phoneme pronunciation. The sentence-level scores of the dataset are further divided into Accuracy, Prosody, Fluency, and Completeness. The mean score of completeness and prosody is measured by eq. 11 and sentence-level Accuracy and Fluency are measured by eq. 15. For the Acoustic model, the Librispeech dataset was used which consists of approximately 1000 hours of English speech derived from audiobooks [32].

B. TRAINING SETUP

The Kaldi ASR toolkit contains recipes to aid ASR researchers, WSL2 Ubuntu Terminal Environment is used to set up the dependencies of Kaldi which mostly runs on bash script files. The Kaldi nnet3-chain recipe has some specific properties:

- Fixed transition probabilities are used in the HMM.
- 3 times lower frame rate is used at the output of the neural network, which dramatically reduces the amount of processing necessary during testing and makes real-time decoding considerably simpler.
- TDNN is used as neural nets as it is faster in short sequences and easier to tune.

The ASR system is about 3 times faster at decoding and training time is also reduced, the workflow is depicted in Fig. 5. The recipe is pre-trained on the Librispeech dataset [32] and is used to get the log posterior ratio (LPR) from the utterances of the speechocean762 dataset [26]. The GOP score is calculated using equations 10 and 11 for the extracted LPR from Kaldi.

The LPR score is used to differentiate the words that have incorrect phonemes, such words are extracted at random from the dataset utterances and compared with the same word uttered ideally by google text-to-speech (GTTS), both the files are then converted into their pitch contours and subtracted with their adjacent values, using Praat Parselmouth library in python with the following the pitch settings:

- Audio Sampling Rate: 22050 samples per second.
- Pitch Floor: 75 Hz - below this frequency will not be recruited.
- Pitch Ceiling: 600 Hz - above this frequency will be ignored.
- Silence threshold: 0.03 - frames that do not contain amplitudes above this threshold are probably silent.

The DTW scores with L2 norm are calculated and fed into eq. 15 to obtain compare score. Both results are combined using eq. 16 to get the total score.

V. RESULTS

This section details the evaluation process of the proposed approach on the LibriSpeech and the Speechocean762 dataset. The proposed approach is evaluated on different settings and is further compared with other pronunciation scoring methods.

Fig. 6 shows the different values of Mean Absolute Error (MAE) and Pearson Correlation Coefficient (PCC) metrics, it can be noticed that as the value of hyperparameter α changes from 0 to 1 the error in the total score decreases and then increases linearly. Similarly, the PCC graph first increases then decreases drastically, this implies that the Posterior Score alone is not effective in conveying the overall Pronunciation score, conversely, the compare score alone cannot determine the same, thus a combination of both these aspects, when taken into consideration, provides a much better result.

Table 2 shows corresponding values of MAE and PCC for different values of hyperparameter α. The maximum value of the total score is obtained at $\alpha = 0.5$, further shifting its value affects the total score.

The graph of deviation for the first 90 speakers with 20 utterances each in comparison with the proposed approach with the general weighted average with hyperparameter as it

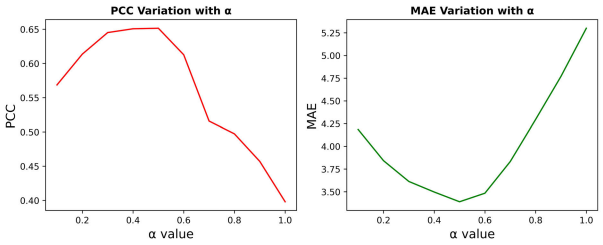


FIGURE 6. Plot of PCC vs α and MAE vs α in range 0 to 1.

TABLE 2. Mean absolute difference of the total score, posterior scores, and compare scores for different values of α.

Hyperparameter(α)	MAE of Total Score	PCC of Total Score
0.4	3.49	0.650
0.5	**3.38**	**0.652**
0.6	3.53	0.612

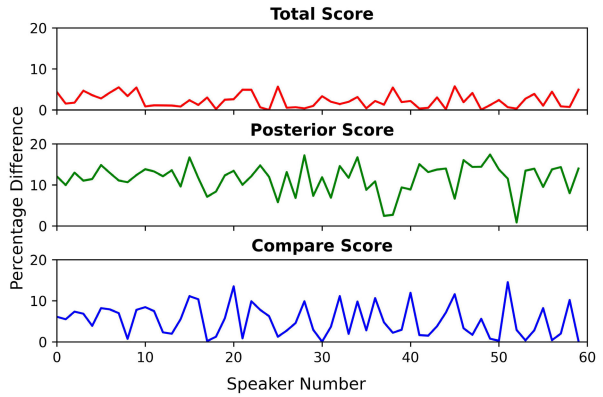


FIGURE 7. Graph of the absolute difference between the Total Score, Posterior Score and Compare Score for $\alpha = 0.5$.

approaches 0.5 is shown in fig. 7. The percentage difference in total score between the proposed method predictions and the dataset is plotted, it can be seen that although the total score gives less difference, the posterior score and compare score individually give a high difference error. To check the individual scores in comparison with the dataset score on the utterance level pronunciation, we sampled 10 different speakers' recordings with their respective total scores from 5 experts from the dataset, their average score, and their corresponding scores generated by our system taking the value of $\alpha = 0.5$ (for general results), as tabulated in table 3.

Table 3 compares the scores given by experts to the predicted scores. It can be clearly seen that the proposed approach matches the average score given by the 5 experts. Expert1 and expert2 scores are lenient whereas expert5 scores are strict but taking an unbiased estimate of pronunciation scores gives us an expected score for a speaker's pronunciation.

The mean difference error for the total score of the ideal expert scores and the proposed approach is shown in fig. 8. The horizontal axis shows the speaker number from the dataset and the vertical axis shows the total score of the speaker in percentage.

TABLE 3. All expert scores and their average score compared with the predicted score.

S no.	E1	E2	E3	E4	E5	Avg	Prediction
1	94.45	84.25	96.26	76.20	85.20	**87.27**	**86.13**
2	95.30	81.55	97.00	71.75	79.90	**85.10**	**78.63**
3	89.55	81.55	84.66	77.85	77.85	**82.92**	**81.24**
4	93.10	85.15	95.95	72.01	77.95	**84.85**	**81.62**
5	90.25	83.55	94.60	70.65	76.88	**83.18**	**84.88**
6	96.15	84.15	98.55	69.30	77.95	**85.16**	**86.90**
7	87.25	78.20	86.85	85.65	78.37	**83.26**	**82.29**
8	95.50	84.20	86.8	86.55	80.21	**86.65**	**81.49**
9	79.80	85.00	88.05	72.80	75.65	**80.26**	**80.29**
10	92.40	83.95	93.22	72.95	76.29	**83.76**	**83.47**

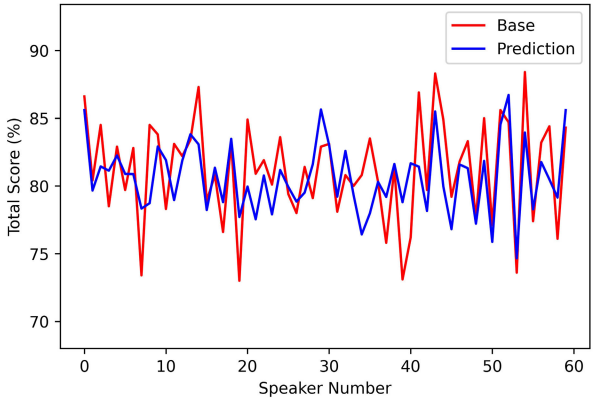


FIGURE 8. Graph of Total Score for expert average scores and model predictions.

TABLE 4. Pearson Correlation Coefficient scores compared with different methods.

Method	Dataset	PCC	Fine Tuned
DNN-HMM [16]	LS, FE Corpus	0.637	Yes
GOP + DNF [19]	WSJ, ERJ	0.667	Yes
PWLD-SVM [24]	WSJ, Spraakbanken	0.610	Yes
HMM-TDNN [28]	LS, TIMIT	0.392	Yes
HuBERT + GOP [35]	KESL	0.650	Yes
Wav2Vec2 Base [36]	SpeechOcean762	0.680	Yes
Proposed Approach	**LS, Speechocean762**	**0.652**	**No**

Table 4 compares the proposed approach with various other approaches using different datasets such as Speechocean762, TIMIT, LibriSpeech, and more. Latest models such as HuBERT [35] and Wav2Vec2 [36] were also compared. It should be noted that our model was not fine-tuned whereas all other models were fine-tuned before getting these results.

VI. DISCUSSION

The results show that there is a deviation of 3.38% from the actual mean total score of the professionals from the speechocean762 dataset. This result is relatively close to the actual labels as leniency is not taken into consideration, but changing the value of α in eq. 4 will fine-tune the predictions

to minimize the mean difference error between the two data vectors, which can be seen from the second row where biasing the total score to compare the score by 10% decreases the error by 0.75%.

A. AUTO-CORRELATION FOR PITCH EXTRACTION

Pitch Extraction is done with the help of the Auto-Correlation function, there are many improved methods for the extraction of the pitch but this method is adaptable to noise as the local maxima generated on the auto-correlation of the speech signal in the time domain is unaffected by the addition of slight noise. The main limitation of using this method is the presence of auto-correlation peaks that exceed the peaks corresponding to the pitch period. As a result, we get 'picking' of peaks, and consecutively incorrect pitch evaluation can occur, but as speech signals don't follow a pattern in a frame unlike musical signals, the limitation is not effective in this case [37].

B. TOTAL SCORE DEPENDENCY ON THE TYPE OF ACOUSTIC MODEL

A large variety of Acoustic models are used in related works, but the proposed method does not focus on a specific type of acoustic model as the model is not fine-tuned on the pronunciation dataset, any trained Acoustic model of an ASR system can be used to calculate the Phone-level features, and although it will decrease the percentage error of the Posterior Score in proposed eq. 11, the total score percentage error will not decrease as shown in Fig. 7. Here DNN-HMM model with TDNN layers is specifically used as its low inference time gives real-time scores in CAPT, although it has low phone score accuracy than other LSTM and transformer-based Acoustic models, it performs better on short utterances.

C. PROBLEM WITH FINE-TUNING MODEL ON DATASET

The graphs plotted in Fig 7 depict that without fine-tuning the Posterior Scores, which are purely dependent on the output of the Acoustic model show a percentage difference of up to 15%, this indicates that the models used in related works [16], [19], [24] produced results which were biased due to some extra feature extraction in the model to fit the scores of the specific dataset used, and are not generalized. Adding the compare score solves the high prosodic score dependency as the mispronounced words are compared with their ideal utterance to generate a different score which deals with the stress, Accuracy, and Fluency of the spoken word. The methods used in this approach use a single hyper-parameter α to tune the total score of the dataset which resolves the issue of biasing. The compare score percentage error can also be improved by changing the threshold value in eq. 22 from a constant scalar value to another hyper-parameter that is proportional to the number of phonemes in the word to maximize the compare score's contribution.

VII. CONCLUSION

Considering the basic GOP approach and distinct length vector comparison using DTW, we derive two formulas to calculate the pronunciation score of an audio file. The first calculates the accumulated posterior probabilities of the acoustic model, and the second randomly clips some of the words spoken in it to get vectors to compare with the ideal pronunciation of those specific words using DTW. These two scores are combined together using a weighted average. The results are evaluated with the total scores of the speachocean762 dataset, each scored by 5 professionals. The suggested method significantly resembles the average expert score from the dataset and offers results that can be tailored for both beginners and experienced learners. A limitation of the proposed approach is that the result of the prosodic score cannot be tuned according to the level of the learner and is entirely dependent on the accuracy of the acoustic model used. Future plans include providing more flexibility to the prosodic score and concentrating more on constructing a corpus for gender and region-based score calculation.

REFERENCES

[1] A. P. Gilakjani, "English pronunciation instruction: A literature review," *Int. J. Res. English Educ.*, vol. 1, no. 1, pp. 1–6, 2016.

[2] J. Harmer, *The Practice of English Language Teaching*. Harlow, U.K.: Pearson Longman, 2007.

[3] A. P. Gilakjani and R. Rahimy, "Using computer-assisted pronunciation teaching (CAPT) in English pronunciation instruction: A study on the impact and the teacher's role," *Educ. Inf. Technol.*, vol. 25, no. 2, pp. 1129–1159, Mar. 2020.

[4] A. Winursito, R. Hidayat, and A. Bejo, "Improvement of MFCC feature extraction accuracy using PCA in Indonesian speech recognition," in *Proc. Int. Conf. Inf. Commun. Technol. (ICOIACT)*, Mar. 2018, pp. 379–383.

[5] F. Zhang, C. Huang, F. K. Soong, M. Chu, and R. Wang, "Automatic mispronunciation detection for Mandarin," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2008, pp. 5077–5080.

[6] N. S. Ibrahim and D. A. Ramli, "I-vector extraction for speaker recognition based on dimensionality reduction," *Proc. Comput. Sci.*, vol. 126, pp. 1534–1540, Jan. 2018.

[7] R. K. Aggarwal, "Improving Hindi speech recognition using filter bank optimization and acoustic model refinement," Doctor Philosophy, Dept. Comput. Eng., Nat. Inst. Technol. Kurukshetra, Kurukshetra, India, 2012.

[8] A. Kumar, M. Dua, and T. Choudhary, "Continuous Hindi speech recognition using monophone based acoustic modeling," *Int. J. Comput. Appl.*, vol. 24, pp. 1–5, Jan. 2014.

[9] G. Gaurav, D. Deiv, G. Sharma, and M. Bhattacharya, "Development of application specific continuous speech recognition system in Hindi," *J. Signal Inf. Process.*, vol. 3 no. 3, pp. 394–401, 2012, doi: 10.4236/jsip.2012.33052.

[10] M. Afify, F. Liu, H. Jiang, and O. Siohan, "A new verification-based fast-match for large vocabulary continuous speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 546–553, Jul. 2005.

[11] H. Wang, J. Xu, H. Ge, and Y. Wang, "Design and implementation of an English pronunciation scoring system for pupils based on DNN-HMM," in *Proc. 10th Int. Conf. Inf. Technol. Med. Educ. (ITME)*, Aug. 2019, pp. 348–352.

[12] Y.-C. Tam, Y. Lei, J. Zheng, and W. Wang, "ASR error detection using recurrent neural network language model and complementary ASR," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 2312–2316.

[13] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Commun.*, vol. 30, nos. 2–3, pp. 95–108, Feb. 2000.

[14] D. Luo, Y. Qiao, N. Minematsu, Y. Yamauchi, and K. Hirose, "Analysis and utilization of MLLR speaker adaptation technique for learners' pronunciation evaluation," in *Proc. Interspeech*, Sep. 2009, pp. 1–4.
[15] Y.-B. Wang and L.-S. Lee, "Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 5049–5052.
[16] S. Sudhakara, M. K. Ramanathi, C. Yarra, and P. K. Ghosh, "An improved goodness of pronunciation (GoP) measure for pronunciation evaluation with DNN-HMM system considering HMM transition probabilities," in *Proc. Interspeech*, Sep. 2019, pp. 954–958.
[17] H. Ryu and M. Chung, "Mispronunciation diagnosis of L2 english at articulatory level using articulatory goodness-of-pronunciation features," in *Proc. 7th ISCA Workshop Speech Lang. Technol. Educ. (SLaTE)*, Aug. 2017, pp. 65–70.
[18] B. Lin and L. Wang, "Deep feature transfer learning for automatic pronunciation assessment," in *Proc. Interspeech*, Aug. 2021, pp. 4438–4442.
[19] S. Cheng, Z. Liu, L. Li, Z. Tang, D. Wang, and T. F. Zheng, "ASR-free pronunciation assessment," 2020, *arXiv:2005.11902*.
[20] M. Bugdol, Z. Segiet, and M. Krcichwost, "Pronunciation error detection using dynamic time warping algorithm," in *Information Technologies in Biomedicine*, vol. 4. Cham, Switzerland: Springer, 2014, pp. 345–354.
[21] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intell. Data Anal.*, vol. 11, no. 5, pp. 561–580, Oct. 2007.
[22] Y. Permanasari, E. H. Harahap, and E. P. Ali, "Speech recognition using dynamic time warping (DTW)," *J. Phys., Conf. Ser.*, vol. 1366, no. 1, Nov. 2019, Art. no. 012091.
[23] Z. Miodonska, M. D. Bugdol, and M. Krecichwost, "Dynamic time warping in phoneme modeling for fast pronunciation error detection," *Comput. Biol. Med.*, vol. 69, pp. 277–285, Feb. 2016.
[24] R. Karhila, A.-R. Smolander, S. Ylinen, and M. Kurimo, "Transparent pronunciation scoring using articulatorily weighted phoneme edit distance," 2019, *arXiv:1905.02639*.
[25] W. Hu, Y. Qian, and F. K. Soong, "An improved DNN-based approach to mispronunciation detection and diagnosis of L2 learners' speech," in *Proc. SLaTE*, 2015, pp. 71–76.
[26] J. Zhang, Z. Zhang, Y. Wang, Z. Yan, Q. Song, Y. Huang, K. Li, D. Povey, and Y. Wang, "Speechocean762: An open-source non-native English speech corpus for pronunciation assessment," 2021, *arXiv:2104.01378*.
[27] A.-L. Georgescu, H. Cucu, and C. Burileanu, "Kaldi-based DNN architectures for speech recognition in Romanian," in *Proc. Int. Conf. Speech Technol. Hum.-Comput. Dialogue (SpeD)*, Oct. 2019, pp. 1–6.
[28] J. Shi, N. Huo, and Q. Jin, "Context-aware goodness of pronunciation for computer-assisted pronunciation training," 2020, *arXiv:2008.08647*.
[29] Y. Gong, Z. Chen, I.-H. Chu, P. Chang, and J. Glass, "Transformer-based multi-aspect multi-granularity non-native english speaker pronunciation assessment," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 7262–7266.
[30] M. Sancinetti, J. Vidal, C. Bonomi, and L. Ferrer, "A transfer learning based approach for pronunciation scoring," 2021, *arXiv:2111.00976*.
[31] S. Kim, M. L. Seltzer, J. Li, and R. Zhao, "Improved training for online end-to-end speech recognition systems," 2017, *arXiv:1711.02212*.
[32] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210.
[33] M. Tu, A. Grabek, J. Liss, and V. Berisha, "Investigating the role of L1 in automatic pronunciation evaluation of L2 speech," 2018, *arXiv:1807.01738*.
[34] X. Han and T. Huwan, "The modular design of an english pronunciation level evaluation system based on machine learning," *Secur. Commun. Netw.*, vol. 2022, pp. 1–11, Jun. 2022.
[35] F.-A. Chao, T.-H. Lo, T.-I. Wu, Y.-T. Sung, and B. Chen, "3M: An effective multi-view, multi-granularity, and multi-aspect modeling approach to english pronunciation assessment," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2022, pp. 575–582.
[36] E. Kim, J.-J. Jeon, H. Seo, and H. Kim, "Automatic pronunciation assessment using self-supervised speech representation learning," 2022, *arXiv:2204.03863*.
[37] L. Sukhostat and Y. Imamverdiyev, "A comparative analysis of pitch detection methods under the influence of different noise conditions," *J. Voice*, vol. 29, no. 4, pp. 410–417, Jul. 2015.
[38] F. Wu, L. P. García-Perera, D. Povey, and S. Khudanpur, "Advances in automatic speech recognition for child speech using factored time delay neural network," in *Proc. Interspeech*, Sep. 2019, pp. 1–5.
[39] W. Ying, "English pronunciation recognition and detection based on HMM-DNN," in *Proc. 11th Int. Conf. Measuring Technol. Mechatronics Autom. (ICMTMA)*, Apr. 2019, pp. 648–652.
[40] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. Interspeech*, Sep. 2015, pp. 1–5.

KAVITA SHEORAN has been working at the Maharaja Surajmal Institute of Technology, since 2004, where she is currently an Associate Professor at the Department of Computer Science and Engineering. She has published several papers in SCI, Scopus, and Elsevier-indexed journals. Her current interests include software engineering and soft computing. She has chaired several sessions of international conferences and published several patents. She is a Life Member of the Indian Society for Technical Education (ISTE) and the Computer Society of India (CSI).

ARPIT BAJGOTI (Student Member, IEEE) was born in Delhi, India, in 2002. He is currently pursuing the B.Tech. degree in computer science and engineering with the Maharaja Surajmal Institute of Technology. His research interests include audio processing, ASR systems, and computer vision.

RISHIK GUPTA was born in Delhi, India, in 2002. He is currently pursuing the B.Tech. degree in computer science and engineering with the Maharaja Surajmal Institute of Technology. His research interests include speech detection and computer vision.

NISHTHA JATANA received the B.Tech. degree from the Computer Science and Engineering Department, the M.Tech. degree in computer technology and applications, and the Ph.D. degree in software testing from Guru Gobind Singh Indraprastha University, in 2021. She is currently an Assistant Professor at the Computer Science and Engineering Department, Maharaja Surajmal Institute of Technology. She has 12 years of teaching and research experience. She has 19 research publications, including eight journals, out of which three are SCIE-indexed journal articles, one book, and seven conference papers, published in various ESCI/Scopus publications. Her research interests include software engineering, meta-heuristic approaches, and network security.

GEETIKA DHAND has been associated with the Maharaja Surajmal Institute, since 2004, where she is currently an Associate Professor at the Department of Computer Science and Engineering. She has published several articles in SCI, Scopus, and Elsevier indexed journals. She has chaired several sessions of international conferences. She has published several patents. Her current research interests include wireless sensor networks, computer networks, and database management systems. She is a Life Member of the Indian Society for Technical Education (ISTE).

CHARU GUPTA graduated the B.E. degree in computer science and engineering. She received the M.Tech. degree (Hons.) in computer science and engineering from the JSS Academy of Technical Education, Noida, and the Doctoral degree from the Department of Computer Science, Banasthali Vidyapith, Rajasthan, India. She is currently serving as an Associate Professor at the Bhagwan Parshuram Institute of Technology (Affiliated to GGSIPU, Dwarka), Delhi, with a teaching experience of more than ten years. She is a Research Associate with Nokia in the ''Invent with Nokia'' venture. She has to her credit various sessions at national and international conferences. She is a Faculty Coordinator (Delhi Section) of Free and Open Source Cell (FOSS cell) from the International Centre for Free and Open Source Software (ICFOSS), Goverment of Kerala, India. She is also the Faculty Coordinator of the e-Yantra Laboratory setup initiative (eLSI) in collaboration with IIT Bombay. She has also served as a Coordinator for Software Engineering for GGS Indraprastha University (IPU) B.Tech. syllabus designing committee. She has a research publications in various national and international journals/conferences of repute (SCIE/ESCI/SCOPUS). She is also a member of the *Encyclopedia of Neutrosophic Researchers* (Vol. 3, Book edited by Prof. Florentin Smarandache, University of New Mexico). She has been the co-convenor of various webinar series and a Faculty Coordinator of the National and International Conferences at BPIT. Her research interests include natural language processing, neutrosophic logic and its applications, time series analysis and forecasting, and evolutionary computation. She is a reviewer, a section editor, and an editorial member of many international journals/conferences. She is the lead editor and the co-editor in various book series by Wiley, De Gruyter, Nova publishing, and many more.

PANKAJ DADHEECH is currently working as an Associate Professor and the Deputy Head at the Department of Computer Science and Engineering (NBA Accredited), Swami Keshvanand Institute of Technology, Management and Gramothan (SKIT), Jaipur, Rajasthan, India (Accredited by NAAC A++ Grade). He has more than 17 years of experience in teaching. He has been appointed as a Ph.D. Research Supervisor at the Department of Computer Science and Engineering, SKIT (Recognized Research Centre of Rajasthan Technical University, Kota). He has also guided various M.Tech. Research Scholars. He has published 22 patents and granted five patents at Intellectual Property India, Office of the Controller General of Patents, Design and Trade Marks, Department of Industrial Policy and Promotion, Ministry of Commerce and Industry, Government of India. He has also published and granted five Australian patents at Commissioner of Patents, Intellectual Property Australia, Australian Government. He has also published and granted one German patent, one South African patent, and one USA patent. He has also registered and granted two research copyrights at Registrar of Copyrights, Copyright Office, Department for Promotion of Industry and Internal Trade, Ministry of Commerce and Industry, Government of India. He has presented 60 papers in various national and international conferences. He has 67 publications in various international and national journals. He has published seven books and 18 book chapters. His research interests include high performance computing, cloud computing, information security, big data analytics, intellectual property right, and the Internet of Things. He is a member of many professional organizations, such as the IEEE Computer Society, CSI, ACM, IAENG, and ISTE. He has chaired technical sessions in various international conferences and contributed as a resource person in various FDP's, workshops, STTP's, and conferences. He is also acting as a guest editor of the various reputed journal publishing houses and conference proceedings and the Bentham Ambassador of Bentham Science Publisher.

UMAR YAHYA (Member, IEEE) received the B.Sc. degree from the Islamic University of Technology (IUT), Bangladesh, in 2011, and the M.Sc. and Ph.D. degrees in computer science from Universiti Brunei Darussalam (UBD), Brunei, in 2014 and 2019, respectively. He is currently a Senior Lecturer of computer science at the Islamic University in Uganda (IUIU), Uganda. His research interests include intelligent systems, computational biomechanics, applied machine learning, the Internet of Things, and wearable computing.

NAGENDER ANEJA received the M.E. degree in computer technology and applications from the Delhi College of Engineering, and the Ph.D. degree in computer engineering from the J. C. Bose University of Science and Technology, YMCA. He is currently working as an Assistant Professor at the School of Digital Science, Universiti Brunei Darussalam. He is also working in deep learning, computer vision, and natural language processing. He is also the Founder of ResearchID Company.

Advances in Natural Sciences:
Nanoscience and Nanotechnology

REVIEW

Recent developments in phycosynthesis of zinc oxide nanoparticles for biomedicine and environmental applications

To cite this article: Shrutika Chaudhary and Navneeta Bharadvaja 2023 *Adv. Nat. Sci: Nanosci. Nanotechnol.* **14** 043001

View the article online for updates and enhancements.

You may also like

- Influence of different sized nanoparticles combined with ultrasound on the optical properties of *in vitro* knee joint and knee osteoarthritis cartilage tissue was studied by OCT and diffuse reflectance spectroscopy
Ran An, Huajiang Wei, Ze Zhuang et al.
- Waste to wealth: a solution to textile dyes related pollution
Boya Palajonnala Narasaiah and Badal Kumar Mandal
- Zinc oxide nanoparticles decorated fluorescent and antibacterial glass fiber pre-filter paper
Anadi Gupta, Vishal Mishra and Rohit Srivastava

This content was downloaded from IP address 14.139.251.98 on 03/11/2023 at 05:36

IOP Publishing | Vietnam Academy of Science and Technology Advances in Natural Sciences: Nanoscience and Nanotechnology
Adv. Nat. Sci.: Nanosci. Nanotechnol. **14** (2023) 043001 (15pp) https://doi.org/10.1088/2043-6262/acf2ef

Review

Recent developments in phycosynthesis of zinc oxide nanoparticles for biomedicine and environmental applications

Shrutika Chaudhary and Navneeta Bharadvaja

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

E-mail: navneeta@dtu.ac.in

Received 10 April 2023
Accepted for publication 29 May 2023
Published 12 October 2023

Abstract

Nanoparticles are becoming integral components in every sector considering their unmatched properties with their counter bulk material. However, in the last couple of decades, several reports suggested metal nanoparticles are toxic to biological systems either directly or indirectly. Zinc oxide nanoparticles (ZnO NPs) possess excellent optical, electrical, food packaging properties, etc. Although, the use of chemically toxic reducing agents for the synthesis of ZnO NPs induces toxicity. Therefore, biogenic synthesis of ZnO NPs has been exploited using different sources such as plant leaves, stems, fungi algae, etc NPs synthesised via these methods are biodegradable, biocompatible, low toxic, and highly effective in different applications. Algae being widely available and easy to harvest becomes a suitable candidate for ZnO NPs synthesis. Algae mediated/phycosynthesis is a technique where algae accumulate zinc oxides and reduce them to zinc ions to form ZnO NPs. The ease of synthesis of ZnO NPs using this method produces NPs in abundant quantity and of variable sizes. Intracellular and extracellular syntheses are two mechanisms of phycosynthesis of ZnO NPs. These ZnO NPs have several beneficial properties like antimicrobial, anti-cancerous, antioxidant, larvicidal, antidiabetic, etc. Additionally, it has low scale-up cost, and low energy input. This article provides detailed information about the routes of phycosynthesis of ZnO NPs using different algal strains, properties, and their potential application in the biomedical field and environmental remediation.

Keywords: phycosynthesis, ZnO NPs, anti-cancerous agents, anti-microbial agent, environmental remediation
Classification numbers: 2.04, 4.02, 5.08

1. Introduction

Nanoparticles (NPs) have emerged as a boon in the field of science because of their extraordinary properties. However, alarming toxicity of NPs synthesised from chemical methods have restricted their application on various fields. Comparatively, biological methods (biogenic) of metal NPs synthesis have superior properties in terms of biocompatibility and toxicity. Therefore, biogenic synthesis of metal NPs has proven to be more advantageous as they are more eco-friendly, and contaminant-free [1]. The use of these materials has the added benefit of ensuring that no harmful chemicals or methods are employed, both of which can diminish the NPs' biocompatibility [2, 3]. The biogenic sources include plant parts, fungi, algae, bacteria, etc. Metal salts are easily converted into NPs through biological synthesis using plants [4], bacteria [5], algae [6], and fungi [7]. In recent decades, algae have become a key participant in the environmentally friendly synthesis of many NPs. This subfield of nanotechnology is known as phyco-nanotechnology [8]. Algae can exist in both multicellular and unicellular forms and can produce oxygen through photosynthesis. Although they lack the fundamental

 © 2023 Vietnam Academy of Science & Technology

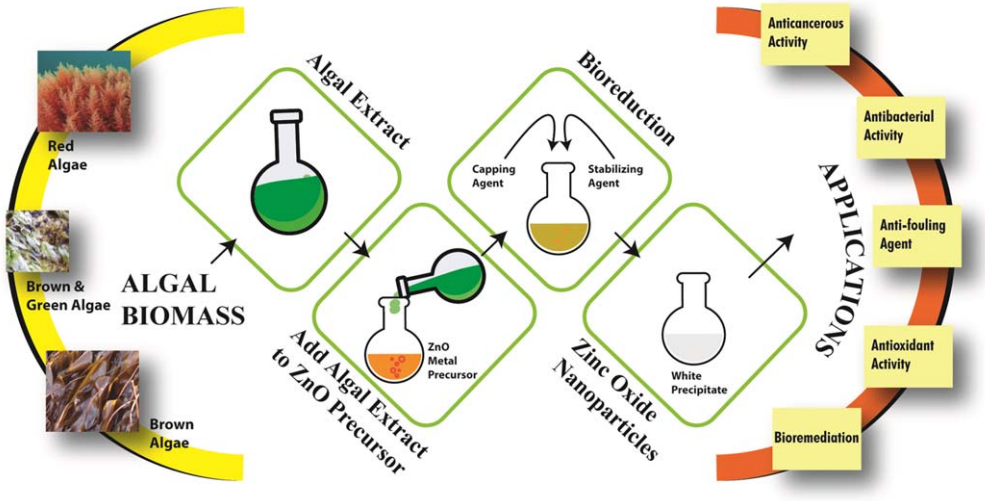


Figure 1. Various types of algae involved in phycosynthesis of ZnO NPs and their potential applications.

structure of plants, such as possessing roots and leaves [9]. Algae-mediated synthesis of ZnO NPs can be done either intracellularly or extracellularly [10]. It is feasible as the enzymes, biochemicals, and other reducing compounds that are already present in the algae can work against the metal precursor that has been supplied [11]. In addition, the microalgae and macroalgae extract-based NPs are the easiest to produce because they do not need any specific culture conditions to be kept alive.

Many metal/metal oxide NPs have been synthesised so far such as CaO NPs [12, 13], SnO_2 NPs [14], CeO_2 NPs [15], BaO NPs [16], MnO_2 NPs [17], MgO NPs [18], ZnO NPs [58], etc but the additional properties of ZnO NPs have attracted many researchers to employ it in different fields of science. One of the benefits of utilising ZnO NPs is that they progressively disintegrate in acidic conditions and release zinc ions (Zn^{2+}) in basic conditions, both of which contribute to apoptosis [19]. The unique features of ZnO NPs, such as their broad band gap, high electron mobility, and high visible transparency, help facilitate the use of ZnO NPs in the production of fabrics that are resistant to the damaging effects of ultraviolet (UV) rays [9]. Other benefits of choosing ZnO NPs that have recently been revealed are their optical properties, high stability, catalytic activity, high surface energy, optoelectronics property, detection of biomolecules, etc [20, 21]. Among all the metal oxide NPs synthesised so far, ZnO NPs possess excellent benefits and can be applied in different fields of science. In addition, they can degrade potentially harmful compounds present in the water bodies and have potential use in biomedical sciences (figure 1). Because they are utilised in such a diverse range of applications, such as fertilisers, feed, bioenergy, and the treatment of wastewater, etc, algal cells have a very high level of sustainability [22]. Recent studies have revealed that the ZnO NPs are not only involved in the removal of harmful pollutants from the environment, but also possess antimicrobial activity, used in cosmetics as a UV-protector, used in target-based delivery of medicines for cancer treatment, and few studies have also revealed its potential as an anti-diabetic agent, wound healing property, etc (figure 1) [19, 23]. The antimicrobial property of ZnO NPs is now used in the manufacturing of textiles which are resistant to microbes and may have better mechanical strength and longer life as compared to the other textiles [24]. In addition, numerous macroalgae play important roles in the food, pharmaceutical, and manufacturing industries [1, 2]. It has been observed that these NPs exhibit excellent adsorption characteristics [25]. When compared to bulk materials, ZnO NPs have a significantly larger surface area. Additionally, the number of active sites that are present on the surface is significantly higher, making them an exceptional material for adsorption applications. Because of this property, they may also adsorb pollutants on their surface, which might be beneficial in the process of removing water pollutants [26]. However, the utilisation of algae for the synthesis of NPs is still in its early stages [27]. Considering the possibilities presented by ZnO NPs, there was a need for the NPs synthesis method that was eco-friendly and biologically compatible. Therefore, the primary objective of this review is to shed light on the recent advancements in the phycosynthesis of ZnO NPs with detailed information about their synthesis and their significance in the biomedicine and environmental remediation.

2. Various methods for ZnO NPs synthesis

There are several sources that have been employed so far for the successful synthesis of NPs. These sources include plants (parts like roots, leaves, stem, etc), bacteria, and fungi (figure 2) [28]. Fungi have good bioaccumulation ability and tolerance and have also been used for NPs synthesis. *C. albicans, A. niger, A. terrus*, etc are some of the fungi used for the synthesis of ZnO NPs (table 1) [29]. Cyanobacteria consists of potential metabolites which aid in the formation of NPs. *Streptomyces* sp., *Bacillus* sp., *Staphylococcus* sp., etc are some of the most widely used bacteria for the

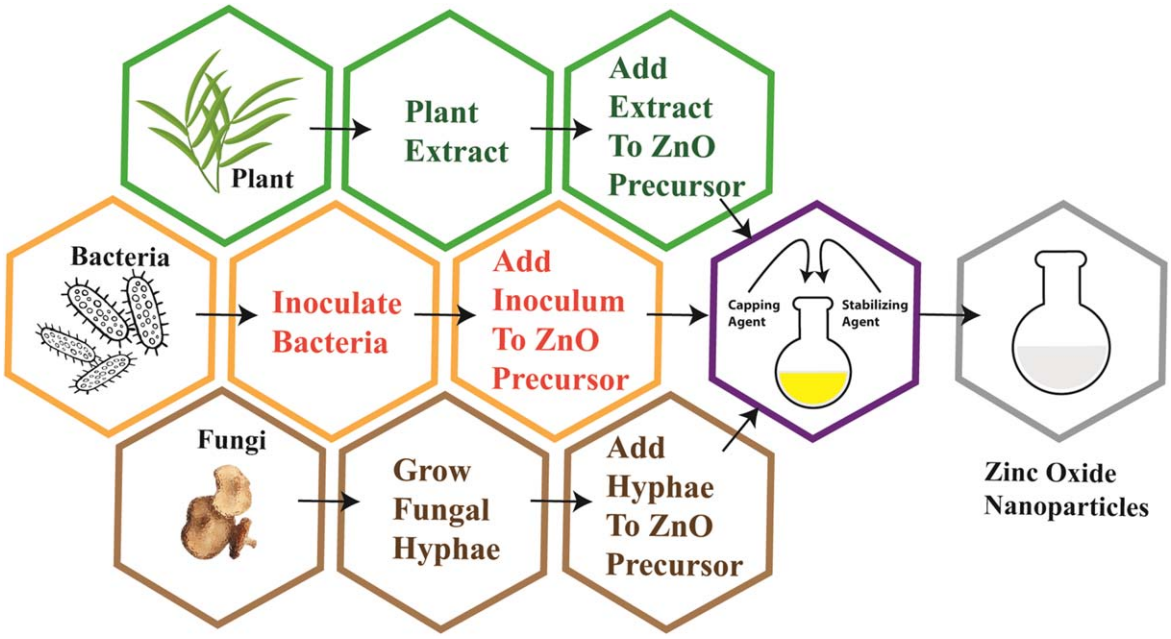


Figure 2. Other methods for biogenic synthesis of ZnO NPs.

development of NPs [30]. Plants have huge potential for NPs synthesis. The important phytochemicals and bioactive compounds such as saponins, alkaloids, flavonoids, terpenes, quinic acid, etc present in the plants play a major role in the formation of NPs [20, 31, 32]. For the NPs formation, it requires extract preparation which is further mixed with the metal precursor for bioreduction to form NPs (figure 2). A detailed information about the sources (plants, fungi, and bacteria) used for the ZnO NPs synthesis with the size of the ZnO NPs formed and their potential applications has been mentioned in table 1. These biogenic ZnO NPs have proven that it has advanced properties like piezoelectricity, photoluminescence, and is more biocompatible as compared to other metal oxide NPs for biomedical applications [33].

3. Phycosynthesis of ZnO NPs synthesis

Phyco-nanotechnology represents a novel and innovative approach for synthesising NPs using algae. Although several chemical and physical synthesis methods were adopted in the past for NPs synthesis, those methods were relatively none too biocompatible and eco-friendly when compared to the biogenic NPs synthesis. Algae have been gaining popularity as a nano-bio-factory due to their high production of secondary metabolites, proteins, and pigments [55]. Several research has pointed that seaweeds are rich in bioactive compounds and can be a good source of carbohydrates, proteins, metabolites, and many more [56]. They have also been considered one of the important sources for the green synthesis of NPs [57] due to their ease of collection, low cost for scaling up, and very quick growth rate [69]. Algae are simple organisms, but they are among the most abundant and widespread photosynthesis-carrying life forms on Earth [59, 69]. As they can reduce metal salts into metal ions to form metal NPs, algae are seen as a better option for the environmentally friendly synthesis of NPs [60]. Algae have branched into numerous classes, each of which is capable of generating metal and metal oxide NPs [61, 62]. In recent years, commercial and industrial exploitation of algae as fertilisers, bioremediation agents, and many more has been observed. The reasons behind choosing algae can be justified on the grounds that it facilitates simple reactions, has minimal energy requirements, eliminates the requirement for reducing or capping agents, and is cost-effective. The two fundamental processes obeyed for the synthesis of NPs involve (I) intracellular, and (II) extracellular synthesis. Intracellular synthesis occurs inside the cell wall or membrane, while extracellular production occurs on the algal cell's surface [63]. These bioactive components exhibit important pharmacological qualities, which can be advantageous for the treatment of numerous diseases and can have therapeutic benefits [56]. The inhibiting and anti-inflammatory actions that algae possess can be exploited in the treatment of a variety of medical conditions, including certain forms of cancer [64]. In addition, seaweeds also exhibit anti-fouling and heavy metal accumulation characteristics [65]. In addition, recent research has shown that seaweed-based aqueous solutions can biologically reduce the levels of metal ions present in the solution [66, 67]. The zinc oxide NPs are synthesised via a combination of cellular and non-cellular mechanisms. The primary benefit of employing algae as a substrate of NPs production is that both living and non-living algae can synthesise NPs when given the right conditions [68]. Synthesis of ZnO NPs involves the following steps: (a) dissolving algal extract in water or some organic solvent; (b) heating the extract mixture at a specific temperature for a specific time-period; (c) preparation of metal ion precursor; (d) addition of as-synthesised metal ion precursor to algal extract solution and stirring continuously for a desired time-period; and (e) observing the creation of

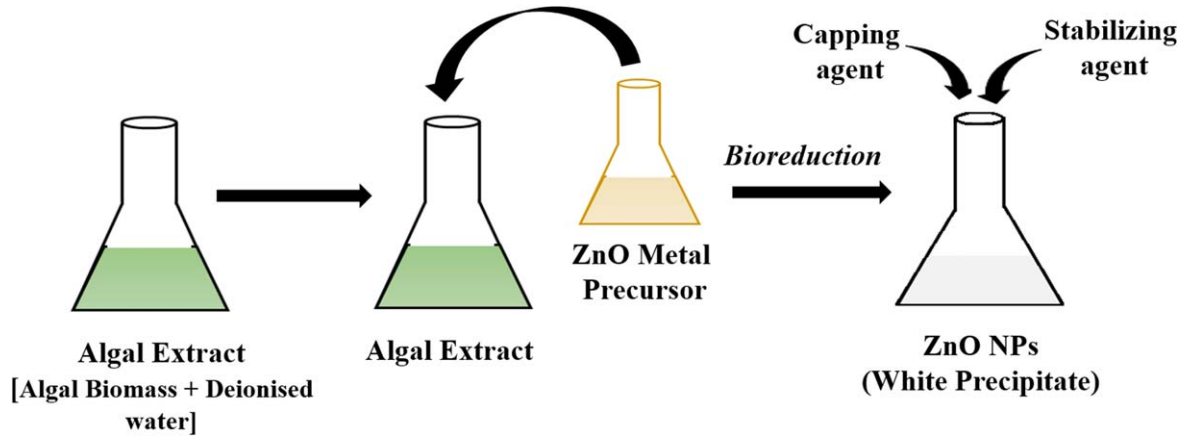


Figure 3. Flowchart of the general procedure followed for the ZnO NPs synthesis from algal biomass.

NPs by analysing a transformation in the colour of the mixture (figure 3). Methods for the synthesis of these NPs vary, as described in the cited studies. Furthermore, if the size of the NPs is small, it will have a larger surface area. Thus, it will exhibit better antimicrobial ability against drug-resistant microbes and may penetrate cells easily [69]. NPs of smaller dimensions will eventually clump together and contact other nucleons to produce a thermostable structure [68]. Thirumoorthy *et al* used zinc acetate as the precursor for the production of ZnO NPs. The effective production of ZnO NPs was confirmed by the colour change from yellow to light brown (figure 3) [70]. Priyadharshini *et al* utilised *Gracilaria edulis*, as the starting material for their ZnO NPs synthesis. *Gracilaria edulis* harvest was combined with zinc nitrate, and the mixture was then left to stand at an ambient temperature in the dark for 12 h, leading to the formation of ZnO NPs in the later stages [71]. Abdulwahid *et al*, in a commendable effort, used *Cladophora glomerata* as a starting material for making ZnO NPs [72]. In the demonstration of ZnO NPs synthesis in 2018, Sanaeimehr *et al* employed *S. muticum* algae for the green-synthesis ZnO NPs [23]. *Pterocladia capillacea*, a kind of red algae, was also employed by Mansour *et al* (2022) as a substrate for ZnO NPs synthesis. NaOH was added drop by drop to a combination of $Zn(Ac)2H_2O$ and *Pterocladia capillacea* extract [22]. El-Belely *et al* took *Arthrospira platensis* (under log phase condition) to prepare its biomass by using Zarrouk's medium. Precipitate formation indicated successful synthesis of ZnO NPs [74]. Recently, *Ulva fasciata* Delile was used as the source for the ZnO NPs synthesis (UFD-ZnO NPs). The size of the NPs formed was around 77 nm. The antibacterial properties of UFD-ZnO NPs revealed its application as a preservative for food items (table 3) [75].

4. Methodology for the phycosynthesis of ZnO NPs

The biological diminution of metal particles into their nanostructure is the primary mechanism that underlies the production of NPs by algae. This mechanism involves the chlorophyll content, phycobilin's, minerals, and other components that all perform a crucial task in the process. For bioreduction, the metal precursor of zinc oxides, such as zinc sulphate or zinc acetate, can be taken. Then, the biological components of the algae must transform the zinc precursors into their zero-valent states [27]. There are three stages to this process: (a) initiation, (b) development, and (c) completion. During the activation stage, the algae secrete enzymes that aid in reducing metal ions, later leading to their nucleation. The entire procedure is analysed by monitoring the solution's colour change. Consolidation between NPs is the following step, which results in NPs with thermodynamically stable shapes and sizes. The third phase is where the NPs take on their definitive form [27]. When it comes to creating NPs of the right size, physical factors play a crucial role. The variables include things like pH, stirring conditions, time, precursor concentration, temperature, and so on. Synthesising NPs from algae can be done in two different methods (figure 4): Intracellular and extracellular synthesis.

4.1. Intracellular synthesis

The term 'intracellular' describes the location of the reaction, which is within the algae's cells. The entire process depends on the intracellular pathways and enzymes taking part in procedures like nitrogen fixation, respiration, photosynthesis, etc in the algal cells. The NADPH or NADPH-dependent reductase enzymes produced at the time of photosynthesis serve as the reductants (figure 4). In addition, it can be produced in the chloroplast thylakoid membrane electron transport system or in the algal cell wall [27, 76].

4.2. Extracellular synthesis

In extracellular biosynthesis, the word 'extracellular' refers to the actions that must be carried out on the exterior of the algal cell wall [76]. The entirety of the operation is governed by the constituents like enzymes/coenzymes, pigments, ions, metabolites, etc, as well as a few non-proteinaceous materials like nucleic acids, hormones, fats, and many more (figure 4) [27]. On the other hand, in the case of extracellular synthesis, pretreatments such as washing and blending are necessary [76].

Table 1. Detailed information about the sources (plants, fungi, and bacteria) used for the ZnO NPs synthesis with the size of the ZnO NPs and their potential applications.

S. no.	Biological sources	Species used for ZnO NPs synthesis	Size	Applications	References
1.	Plants	*Aloe vera* (Leaves)	8–18 nm	Nanoantibiotics	[4]
2.		*Coffee* (Powder)	4.6 nm	Photocatalytic activity, and photo-oxidation of methylene blue dye	[34]
3.		*Plectoranthus amboinicus* (Leaves)	20–50 nm	Mosquito larvicidal activity, anti-biofilm, and antibacterial activity	[35]
4.		*Carica papaya* (Milk latex)	11–26 nm	Red-S dye degradation, and anti-bacterial activity	[36]
5.		*Polygala tenuifolia* (Roots)	33.03–73.48 nm	Antioxidant, and Anti-inflammatory activity	[37]
6.		*Cassia fistula* (Leaves)	5–15 nm	Photo-degradation, antioxidant, and antibacterial activity	[38]
7.		*Anchusa italica* (Flower)	8–14 nm	Antimicrobial activity	[39]
8.		*Trifolium pretense* (Flower)	100–190 nm	Antibacterial activity	[40]
9.		*Solanum nigrum* (Leaves)	20–30 nm	Antibacterial activity	[41]
10.		*Black tea* (Dried solid waste)	19.3 nm	Ibuprofen, Ephedrine and Propranolol removal from urine	[31]
11.		*Phyllanthus niruri* (Leaves)	25.61 nm	Photocatalytic activity	[42]
12.		*Rosa canina* (Fruit)	<50 nm	Antibacterial activity	[43]
13.		*Parthenium hysterophorous* (Leaves)	6–108.5 nm	Antimicrobial activity	[44]
14.	Bacteria	*Sphingobacterium thalpophilum*	40 nm	Antimicrobial activity	[5]
15.		*Nostoc sp.* EA03	50–80 nm	Bactericidal activity	[30]
16.		*Staphylococcus aureus*	10–50 nm	Antimicrobial activity	[45]
17.		*Halomonas elongate* IBRC-M 10,214 strain	18.11 nm	Antibacterial activity	[46]
18.		*Bacillus megaterium* NCIM2326	45–95 nm	Antimicrobial activity	[47]
19.		*Streptomyces* sp.	20–50 nm	Anticancer, and antimicrobial activity	[48]
20.	Fungi	*Alternaria tenuissima*	15.45 nm	Anticancer, antioxidant, photocatalytic, and antibacterial activity	[49]
21.		*Candida albicans*	25 nm	Steroidal pyrazoline synthesis	[50]
22.		*Pichia kudriavzevii*	10–61 nm	Antioxidant activity	[51]
23.		*Fusarium keratoplasticum* A1–3	10 to 42 nm	Cytotoxicity activity	[52]
24.		*Periconium* sp.	40 nm.	Antioxidant activity	[53]
25.		*Aspergillus niger*	53–69 nm	Dye degradation	[29]
26.		*Aspergillus terreus*	10–45 nm	Medical textile, and UV protection	[54]

To achieve a more successful synthesis of NPs, it was required to take into account certain physiological variables such as temperature, pH, as well as the nature and concentration of the precursor [27]. For preparing dried algal biomass, any dead algae are first washed, then blended, and then placed in the shade for a predetermined amount of time. Then the biomass is either sonicated or incubated to get a screened product [68].

5. Factors influencing the phycosynthesis of NPs

For the efficient synthesis of NPs from algae, many key factors need to be considered. These factors can be biological or physical including temperature, pH, the concentration of reactants and metal ions, choice of algae, stirring conditions, etc. All these factors perform major character in the synthesis of NPs possessing desired size and amount. Optimising these factors may also help in synthesising more stable NPs and may also produce better results during application.

5.1. pH

While synthesising NPs from algae, pH can influence the shape of NPs formed (figure 5) [68]. Nagarajan *et al* mentioned the synthesis of ZnO NPs from seaweeds (*Sargassum myriocystum*). The absorption peak observed at pH 8 indicated that the metal precursor (zinc nitrate) was completely reduced to ZnO NPs. The absorption peak formed at 372 nm showed a blue shift indicating the bulk NPs synthesis but at an absorption peak of 380 nm, spherical NPs were formed [77]. pH also has an impact on the repulsion and attraction of metals present in the solution where NPs have to be formed. It can also affect the charge present in the functional groups and the conformation of the cells as well [66]. Khanna *et al* demonstrated the synthesis of NPs from *C. serrulata* extract. Smaller NPs were formed when the basic pH was maintained

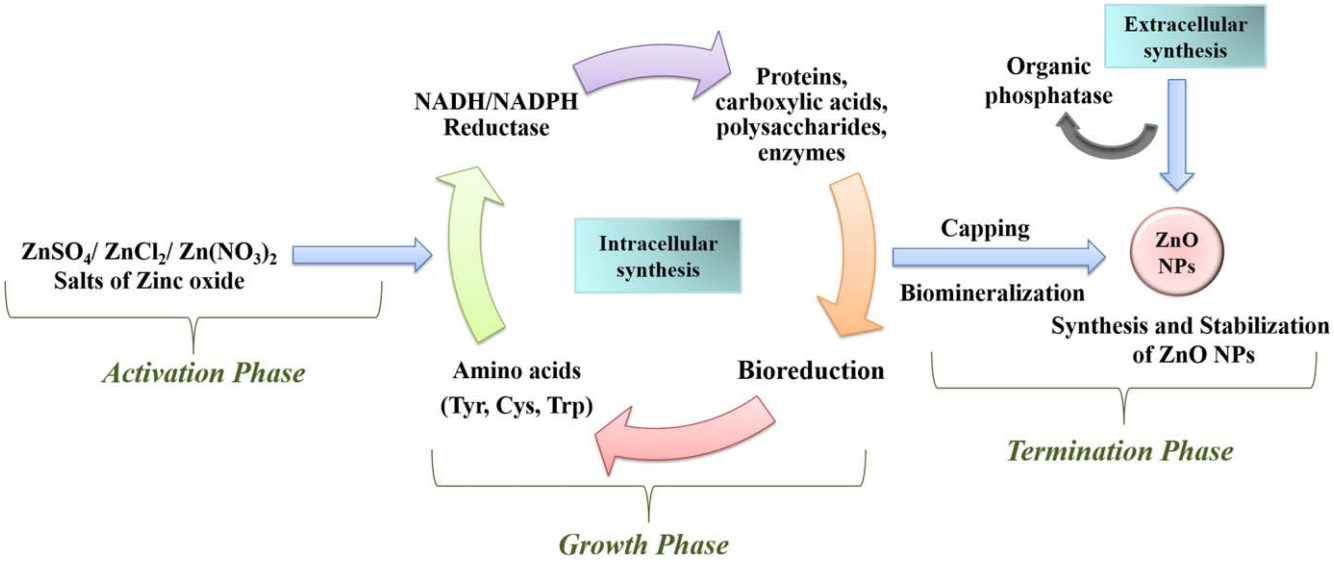


Figure 4. The mechanism of intracellular and extracellular ZnO NPs phycosynthesis.

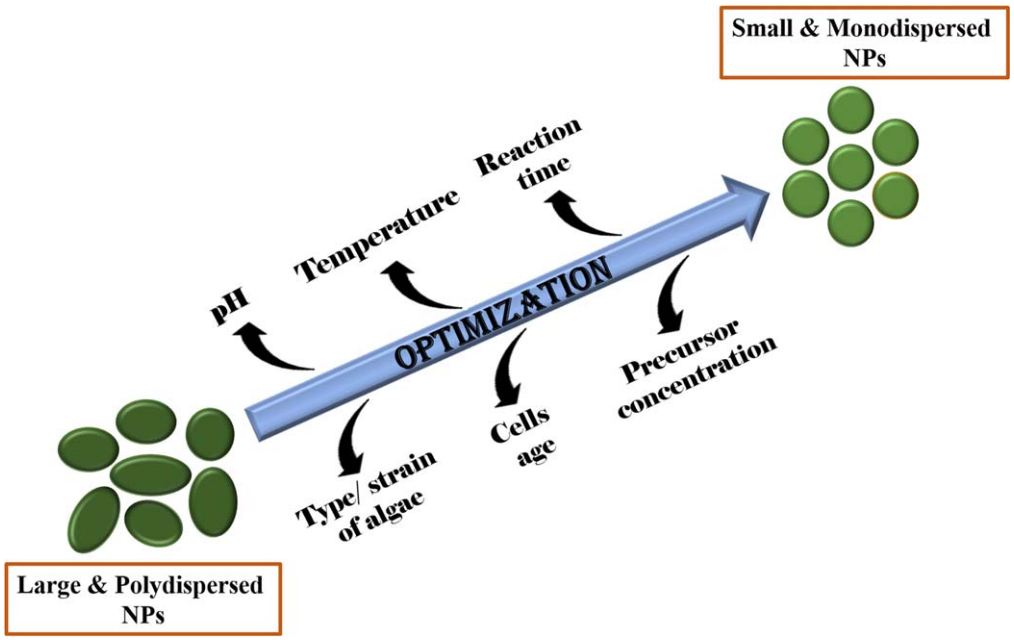


Figure 5. Optimisation parameters required to obtain small, and monodispersed NPs.

between 6.65 and 9.95. Higher pH is favourable because the reducing and stabilising power of the NPs increases. In the case of lower pH, aggregation of the NPs was observed due to their large size [76]. At wide surface plasmon resonance (SPR), the particles formed are large and of different shapes (can be rectangular, triangular, cylindrical, etc). This condition happens at low pH but at higher pH, algae form spherical NPs of small sizes [68]. For ZnO NPs synthesis, high pH (alkaline) is favourable as maximum productivity is obtained and the particles are more stable and reactive. The possible reasons can be the release of biomolecules from the algal cells providing stability and acting as a capping agent for NPs [78].

5.2. Temperature

Another important factor governing the synthesis of NPs is temperature. It not only affects the size but also the monodispersity of the NPs (figure 5). During phycosynthesis, size of NPs is directly proportional to the temperature. Increase in temperature increases the rate of reaction, and thus enhances NPs synthesis [76]. Maggy *et al* demonstrated the synthesis of NPs from cyanobacteria. The NPs' size increased as the temperature was increased from 25 to 100℃. At lower temperatures, crystalline spherical NPs were produced [11]. According to Nagarajan *et al*, an absorption peak was observed at 80℃ when 5 ml of *Sargassum myriocystum* extract was mixed with $ZnNO_3$ to produce ZnO NPs [77].

Thus, it can be concluded that temperature influences the synthesis of NPs.

5.3. Reactant and metal ion concentration

Concentration of extract added to the solution also governs the NPs synthesis. When the concentration of *C. serrulata* was between 5%–20%, the SPR band intensity was increased and the size of the NPs was reduced. But, when the concentration was further increased to 25%, the band intensity was reduced, leading to an increase in the size [68, 76]. Apart from the concentration of extract, concentration of metal precursor also contributes to the morphology of the NPs. The concentration of zinc acetate for the ZnO NPs synthesis was varied between 0.5–2.5 mM. The wavelength of absorption was increased, depicting the active biomolecules present in the cells. Further increase in concentration up to 2.5 mM showed a significant decrease in absorption, leading to the agglomeration of ZnO NPs [78]. Similarly, production of ZnO NPs was observed at different concentrations of seaweed *Sargassum myriocystum*. However, concentration above 25 ml leads to the particle agglomeration. Now, considering the concentration of zinc nitrate from 0.25–0.75, no absorption peak was noted. But, when the concentration was increased to 1mM, a significant peak was observed at 372 nm [77]. Therefore, it can be concluded that the reactant or the metal ion concentration can have their impact on the activity of the NPs (figure 5).

5.4. Reaction time

Time is also an important factor as it can also produce quality results when optimised. Changing reaction time varies the amount and the size of the NPs produced. An example of the effect of reaction time is demonstrated by Aboelfetoh *et al* where 10% of *C. serrulata* was mixed with the metal precursor for 8 days to observe the production of NPs. An increase in the peak intensity without any shift leads to the conformation that reaction time also has its impact on the formation of NPs [68, 76].

5.5. Algae strains

Prior to selection of algal strain for NPs synthesis, the intrinsic properties of the cell/organism should be considered. Some of the intrinsic properties include the enzymatic activity of the algae, pathways followed by the organism and the growth rate of the alga. Apart from the above mentioned properties, size of the inoculums and the selection of biocatalyst are also important enough to be considered [68]. This is an important factor because there are few algae whose functions are very slow and may take few weeks to produce NPs. Thus, species should be chosen such that the synthesis should be rapid.

6. Characterisation techniques for ZnO NPs

Identification of successful synthesis of ZnO NPs is very important. It is necessary to identify the size, morphology, functional groups, surface topology, etc. For the identification of the purity and crystallinity of the ZnO NPs, the x-ray diffraction (XRD) technique is followed [70]. Phycosynthesis of ZnO NPs was performed recently in 2021 where characterisation using the XRD technique revealed the crystalline spherical structure of ZnO NPs. For functional group identification, Fourier transform infrared (FTIR) spectroscopy was used. The results revealed the presence of O–H, C–H, and C=O groups [70]. To identify the structure and size of the ZnO NPs, scanning electron microscopy (SEM) was used by Mansour *et al* recently in 2022. It was found that the ZnO NPs possessed rod-like structures with an average size of 1 μm [22]. Apart from SEM, a transmission electron microscope (TEM) was used to observe detailed morphology of ZnO NPs. Govindasamy *et al* in 2021 used TEM for the determination of their size. ZnO NPs formed at calcination temperature 400℃ had size ranging from 8–10 nm [79]. Thermal properties of these ZnO NPs can be analysed by thermogravimetric (TG) and differential thermal analysis (DTA). Thermal properties of histidine capped-ZnO NPs were determined by two-step decomposition method [80]. The data obtained from the analysis revealed the presence of capped histidine, stable ZnO NPs, and water crystallisation [80]. To determine the optical properties and absorbance of NPs, UV-visible spectroscopy is performed. ZnO NPs formation from algae (chlorella extract) had shown an absorption peak at 362 nm [81]. To identify the chemical state and elements existing inside a material, x-ray photoelectron spectroscopy (XPS) is used. El-Belely *et al* used XPS and confirmed the presence of carbon, nitrogen, and oxygen [74]. Brunauer–Emmett–Teller (BET) is another technique for the determination of morphological and optical characteristics of ZnO NPs. Mansour *et al* used BET to observe the surface area and pore size of ZnO NPs. The BET surface area and average pore size were 113.751 $m^2\ g^{-1}$ and 2.5 nm respectively [22]. For qualitative elemental analysis, energy dispersive x-ray spectroscopy (EDX) was applied. Chouke *et al* used EDX to determine the composition of ZnO NPs. The strong signals obtained for Zn and O confirmed the purity of composition [86].

7. Applications of phycosynthesised ZnO NPs

7.1. Biomedicine application

ZnO NPs in biomedicine have several potential benefits. It can be effective against wide range of applications such as frequently occurring microbial infection to neutralising cancer cells. Additionally, it can show antioxidant, and antidiabetic properties [82]. Some of the applications are described below.

7.2. Antioxidant activity

Many synthetic antioxidants have been synthesised earlier but their toxic effects at higher concentrations have prohibited its use. Antioxidants act as a free radical scavenger and help in the prevention of serious illnesses like neurodegenerative disorders, heart diseases, diabetes, etc [83]. Zinc oxide nanomaterials' ability to scavenge free radicals was evaluated using the α,α-diphenyl-β-picrylhydrazyl (DPPH) test. The wavelength at which the colour transformed to yellow from purple was measured to be 517 nm. ZnO NPs were shown to be effective hydroxyl radical scavengers, with a reduction of around 87.3 percent [70]. In a praiseworthy effort, antioxidant properties were evaluated using four macroalgae (*H macroloba, H tuna, A acetabulum,* and *Enteromorpha* sp.). Among these, *Enteromorpha sp.* possessed maximum antioxidant activity [84]. *S fusiforme* (SFPS) is a brown alga used to analyse its antioxidant activity. Two different polymeric fractions were prepared (SFPS-1 and SFPS-2). According to the assays performed, it was confirmed that SFPS-2 had higher antioxidant properties as compared to SFPS-1 [85]. Ananthi *et al* in 2009 mentioned the usage of *Turbinaria ornata* (TCP) as an antioxidant. The antioxidant activity of TCP was evaluated by following DPPH, 2,2′-azino-bis(3-ethylbenzothiazoline-6-sulfonic acid (ABTS), nitric oxide (NO) assays and has proven to be a good antioxidant and anti-inflammatory agent [83]. In addition, the antioxidant activity was determined with the ABTS scavenging assay. Different concentrations of ZnO NPs from 175 to 2800 μg ml^{-1} were added to the ABTS radicals. While the EC50 was calculated to be 600 μg ml^{-1}, maximum radical scavenging was seen at a concentration of 2800 μg ml^{-1} [23]. In recent studies, it has been observed that the green synthesis of ZnO nanospheres (NSs) via *Aspidopterys cordata* also possessed antioxidant properties when DPPH assay was carried out [86]. The formula used to determine the degree of inhibition is as follows (equation (1)):

$$\text{Percentage of Inhibition (IP\%)} = \frac{A_C - A_S}{A_C} \times 100. \quad (1)$$

7.2.1. Antimicrobial activity. The antimicrobial activity of ZnO NPs was tested against various microbes. For example, the gram-positive bacteria *S. mutans, M. luteus, S. aureus,* and the gram-negative bacterium *P. aeruginosa* were all killed by ZnO NPs, as reported by Nagarajan *et al.* The disc assay and well diffusion processes were followed to examine the antibacterial activity. The application of NPs resulted in the formation of an 'inhibition zone' (table 2) [19, 87]. In order to kill bacteria, ZnO NPs generate reactive oxygen species. The interaction of H_2O_2 with biomolecules of the membrane or the lipid membrane present at the external surface is another possible cause of bactericidal activity [77]. Fungi, in addition to bacteria, have also been seen to engage in these behaviours. Activity against *C. albicans* was found to be quite high, while activity against *A. niger* was found to be just modest [77]. The agar diffusion method was obeyed for the analysis of the antibacterial agent's efficacy against *S. aureus* and *E. coli*. In this method, the antibacterial activity of ZnO NPs was demonstrated by the observation of an inhibition zone against two different bacterial strains [70, 88]. Suspension of ZnO NPs has also proven its activity against *E. coli*. Preparation of ZnO NPs as nanofluids was done using two dispersants: (a) Polyethylene glycol (PEG), and (b) polyvinylpyrolidone (PVP). The antimicrobial effects of the NPs rise as their size decreases [89, 90]. This is due to the damage caused by the NPs when they come into contact with the cell membrane, which results in the generation of reactive oxygen species (ROS) [19]. Another approach related to the preparation of ZnO nanofluids was mentioned, where NPs were combined with sodium pyrophosphate. This nanofluid showed promising results by working against microorganisms that degrade food quality and act as food spoilers [91]. In recent studies, it has been reported that the biogenic ZnO NPs have antiviral effects against herpes simplex virus (HSV) as well as human papillomavirus (HPV). ZnO NPs inhibit viral activity by disrupting their membranes. Due to their high permeability, it also disrupts the life cycle of viruses and also generates ROS [92]. The antimicrobial property of these NPs was used against *F. oxysporum* and *R. solani*. The following equation was used for calculating the percentage of mycelial growth that has been inhibited (equation (2)):

$$\%\text{Inhibition} = \frac{C - T}{C} \times 100, \quad (2)$$

where '*C*' represents the ratio of fungal mycelial biomass to the dry weight in control; '*T*' represents the ratio of fungal mycelial biomass to the dry weight in control in various test concentrations [72].

These NPs inhibit cell division, increase membrane rupture, and interfere with the development of the cell wall. Fungi have ergosterol in their cell walls, and ZnO NPs block this sterol, making the fungi walls more permeable and eventually killing the fungi [72]. Due to their positive charge, these NPs are drawn to and bind with -COOH groups, which halts budding and ceases transcription [72]. Antifungal activity was also tested against *Erythricium salmonicolor*, a microorganism behind the pink disease. ZnO NPs acted against the growth of *Erythricium salmonicolor* by damaging their cell walls. Thus, we can say that ZnO NPs can be an asset as an antifungal agent [93]. The antibacterial function of ZnO NPs is due, in part, to their propensity to produce ROS and to cause alterations in the permeability of cell walls. However, the release of Zn^{2+} ions has a deleterious effect on biological events, metabolism, the fundamental function of microorganisms, etc Zn^{2+} is amphoteric, meaning it reacts both with acids and alkalis [74]. Formation of a nanostructured microsphere by functionalising ZnO NPs with silver (Ag@ZnO) revealed that it can act effectively as the disruptor of bacterial biofilm by releasing ROS and rupturing their cell surface [94]. In 2021, Brindha *et al* demonstrated the ZnO nanofiber formation absorbed with *Andrographis paniculata* (AP-ZnO) to test its antimicrobial property. These AP-ZnO possessed larvicidal activity as well against *Aedes aegypti* [95]. Recent studies have also revealed that when ZnO NPs

Table 2. Table of algae that are utilised in the manufacturing of ZnO NPs, as well as the microbes against whom the antimicrobial effect is assessed.

S. no.	Algae used for ZnO NPs	Nanoparticles size (nm)	Antimicrobial activity of as-synthesised ZnO NPs	Microbial growth inhibition screening (inhibition zone in mm)	References
1.	*Tetraselmis indica*	20 to 40	*E. coli, S. aureus*	12.3 ± 0.3, 18.4 ± 0.5	[70]
2.	*Anabaena cylindrica*	40 to 60	*S. aureus, Pseudomonas aeruginosa*	9–12, 10–22	[19]
3.	*Caulerpa peltata, Red Hypnea Valencia, Brown Sargasum myriocystum*	36	*Neisseria gonorrohea, Candida albicans, Aspergillus niger, Klebsiella pneumonia, Vibrio cholera, S. aureus, Streptococcus mutans*	5–10	[77]
4.	*Arthrospira platensis*	30 to 55	*E. coli, P. aeruginosa, B. subtilis*	19.9 ± 0.1, 19.1 ± 0.3, 24.1 ± 0.3	[74]
5.	*Cladophora glomerata*	14.39 to 37.85	*Rhizoctonia solani, Fusarium oxysporum*	—	[72]

were capped with histidine, it possessed effective photoluminescence (PL emission) properties. Histidine capped-ZnO NPs had good antibacterial properties against *Klebsiella pneumonia, E. coli,* and *Enterobacter sp.* at 10 μg ml^{-1} concentration [80]. Similarly, when these ZnO NPs were synthesised from *Bauhinia racemose*, it possessed both antibacterial and antioxidant properties [96]. A list of algae that contribute to ZnO NPs synthesis and antimicrobial properties against different microorganisms is provided in table 2.

7.3. Anticancerous activity

ZnO NPs have been found to have anticancer characteristics, which have led to their discovery. In order to examine the ZnO NPs' potential anti-cancer action on cancer cells, a number of different procedures and assay have been carried out. A (3-(4, 5-dimethylthiazolyl-2)−2, 5-diphenyltetrazolium bromide) MTT assay was performed in 2018 on three different cells: (i) *Caco-2 cells*; (ii) *normal Vero cells*; and (iii) *clone-9 cells*, which were taken from colorectal adenocarcinoma cells of humans, from the kidney of African green monkey, and epithelial cells isolated from the liver of a normal rat [78]. A comparison was made between the viability of the cells and that of a control group. The following equation shows how the percentage of viability and toxicity was determined (equations (3), (4)) [78, 97]

$$\text{Viability (\%)} = \frac{\text{Absorbance of sample}}{\text{Absorbance of control}} \times 100, \quad (3)$$

$$\text{Tocixity (\%)} = 100 - \frac{\text{sample absorbance}}{\text{control absorbance}} \times 100. \quad (4)$$

Cancer cells had an IC-50 of 79.57 μg ml^{-1}, while in normal cells it was observed between 26 and 27 μg ml^{-1} [78]. The preparation of GE-ZnO NPs involved the use of extract from *Gracilaria edulis*, and the MTT assay was performed on various concentrations of ZnO NPs ranging from 10–100 μg ml^{-1} [71]. The percentage of cell viability could be computed by first measuring the optical density of the cell and then using the formula that is presented in the next line (equation (5)):

$$\text{Percentage of Cell Viability} = \frac{\text{OD value of experimental NPs treated sample}}{\text{OD value of experimental untreated sample}} \times 100. \quad (5)$$

The GE-ZnO NPs were found to have stronger anticancer effects on the *PC-3 cells* [71]. Generally, ZnO NPs get activated as soon as it enters the tumour cells and produces ROS, free radical that usually hinders the basic mechanism of cells. Herein, ROS can either bind with nucleic acids or interrupt standard functioning of cell organelles such as mitochondria, and ribosomes. This ROS generation leads to the apoptosis of tumour cells (figure 6) [19, 23]. El-Belely *et al*, in an admirable effort, detailed the anticancer effects of ZnO NPs using 2 cell lines: (a) normal lung *WI38 cell line* and *Caco-2 cell line*. They measured anticancer efficacy using an MTT test. After 48 h, the IC50 values for cell viability were 53.34 & 9.95 ppm respectively [74]. Performance of ZnO NPs as anticancer agent is also dependent on the concentration of NPs used. In praiseworthy effort, Sanaeimehr *et al* reported that around 97% reduction in cancer cell line of human liver (*HepG2*) was seen when they were exposed to 2800 μg ml^{-1} of ZnO NPs for 72 h. But, when the NPs' concentration was 175 μg ml^{-1}, only 50% reduction in cell line was found [23]. In addition, the cell viability assay was performed on the skin melanoma cell line (*B16F10*). The micro plate reader was used to determine if the cell line was

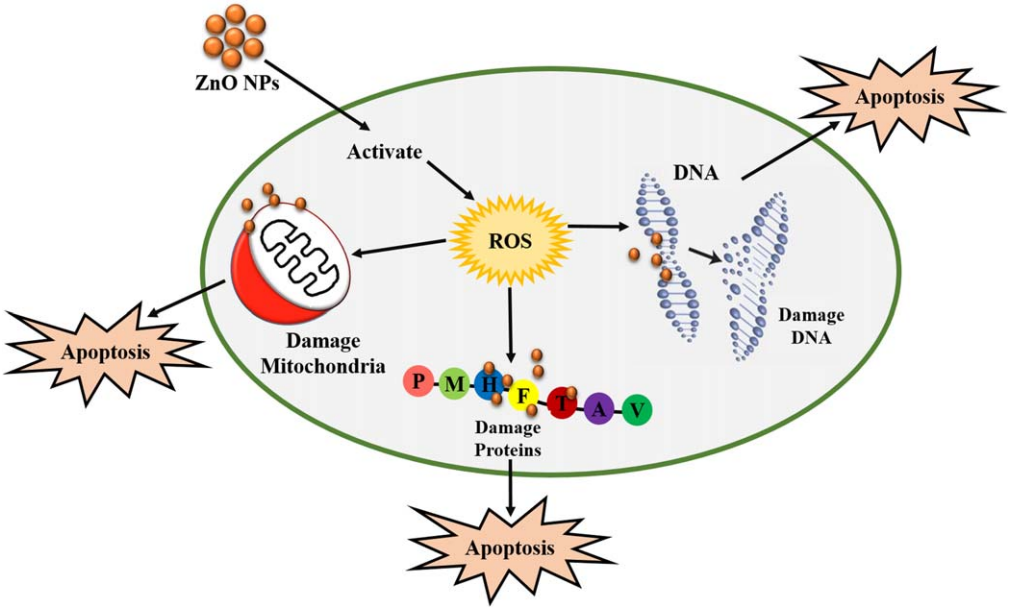


Figure 6. Schematic representation of anticancer activity of ZnO NPs.

still alive at 595 nm. Researchers compared ZnO particles to ZnO NPs. 50% reduction in viability of cells was observed when ZnO NPs (3%) was used whereas regular particles exhibited 80% reduction [19]. Therefore, it is safe to say that the toxicities of nanomaterials are directly proportional to the concentration of the particles.

7.3.1. Antidiabetic activity. Diabetes is a well-known disease nowadays caused by the dysfunction of insulin in the human body. It has been mentioned in different research that ZnO NPs can act as an antidiabetic agent. In an approach, ZnO NPs were used against alloxan diabetic rats. ZnO NPs had great impact in reducing the fast blood glucose (FBS) and can be utilised in the production of nano-antidiabetic drugs [99]. Similarly, usage of ZnO NPs against diabetes was mentioned in a review where alloxan-induced rats were used as model to test the antidiabetic effects of green-synthesised ZnO NPs and they confirmed that ZnO extracts along with ZnO can be utilised for the treatment of diabetes mellitus [100]. In 2008 Hiromura *et al* illustrated the capabilities of ZnO NPs as an antidiabetic agent. They confirmed in their work that the zinc possessed insulin-mimetic activity which can be useful for the treatment of diabetes mellitus type 1 and type 2 [98]. Reasons behind choosing ZnO NPs widely as antidiabetic agents are because they promote glycolysis, induce the expression of GLUT-4, and possess high permeability. They also have synergistic impact on GLUT-2, glucokinase, INS, and IRA [101].

7.4. Environmental applications

In recent studies, it has been mentioned that the algae-mediated synthesis of ZnO NPs also has the ability to eradicate dyes from polluted water and can also remove heavy metals by acting as a nano-adsorbent [102, 103]. The pollutant degradation properties of zinc oxide can be utilised in wastewater treatment as well. One of the major benefits of these NPs is that it can be reused again and again without any effect in its performance [59].

7.4.1. Pollutant removal and degradation. ZnO NPs are not only helpful in the aforementioned scenarios; they can also be used to degrade colours. Degradation of *IV2R* (azo dyes) by ZnO NPs was mentioned by Mansour *et al.* Adsorption of *IV2R* dye over NPs is thought to be facilitated by the electrostatic interactions (figure 7), H-bonds, and hydrophobic bonds between the dye's aromatic rings and the NPs' aromatic rings. According to the data, ZnO NPs exhibited a 95% absorption rate [22]. This property of removing pollutants was also described by Khalafi *et al* in 2019 where dibenzothiophene (DBT) was removed using ZnO NPs (table 3) [81]. ZnO NPs at a concentration of 0.01 g l^{-1} degraded DBT by almost 97%. In addition, using the identical ZnO NPs for 5 consecutive occasions essentially revealed no degradation in performance [81]. It has also been used to eradicate amaranth and methyl orange in addition to *IV2R* (figure 7). A maximum amount of both dyes was shown to be eliminated by 0.3 g of ZnO NPs at a pH of 6. Dye adsorption onto the NPs occurred according to pseudo-second-order kinetics and Langmuir model [25]. These NPs were also used for the degradation of methylene blue (MB) dye. ZnO NPs (<100 nm) showed better degradation than ZnO NPs (<50 nm) [104]. Apart from degradation, these NPs also exhibited adsorption properties. ZnO NPs were capable of adsorbing MB dye due to the ionic interaction between the NPs and sulfonic group present in the MB dye [105]. According to Zheng *et al*, ZnO NPs had an impact on the removal of nitrogen and phosphorus from real municipal wastewater [73, 106]. The pH of the solution performs a pivotal job in the adsorption of heavy metals on the surface of

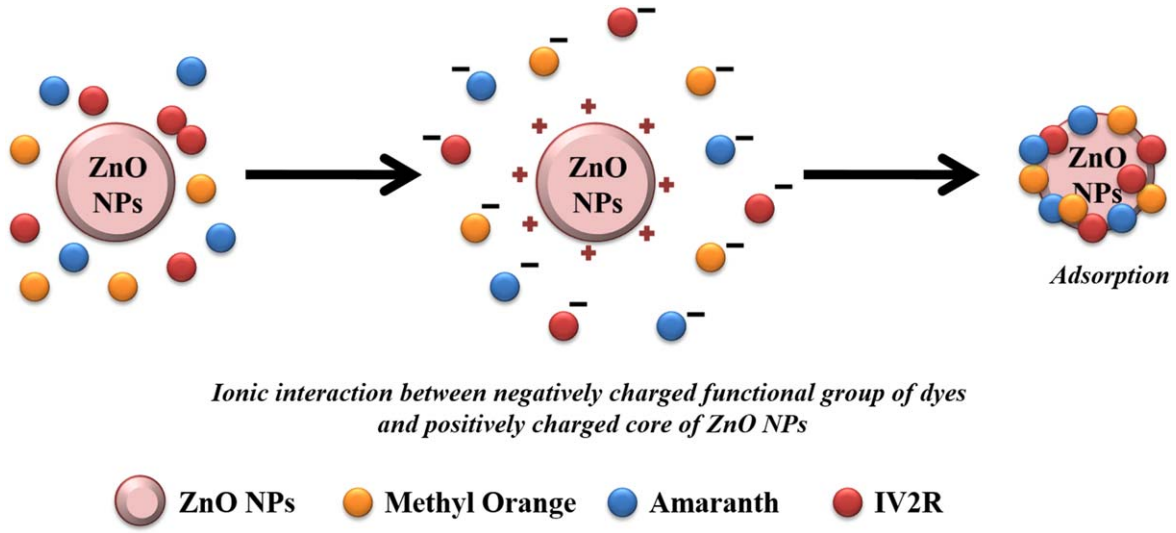


Figure 7. Possible mechanism behind the adsorption of dyes on ZnO NPs.

Table 3. The following is a listing of the algae that were utilised in the formation of ZnO NPs, along with the parameters that were optimised and their uses.

S. no.	Synthesis of ZnO NPs using algae	Optimised conditions			Implementation	References
		Temp. (℃)	Time (min)	RPM		
1.	*Sargassum muticum*	69	180–240	4000	Antiangiogenic, antiapoptotic	[2, 23]
2.	*Sargassum myriocystum*	80	5–10	—	Antimicrobial	[77]
3.	*Tetraselmis indica*	80	180	—	Antimicrobial activity, Hemolytic activity, Free radical scavenging	[70]
4.	*Cladophora glomerata*	50	15	16000	Antifungal property	[72]
5.	*Pterocladia capillacea*	65	120–180	5000	Dye degradation	[22]
6.	*Arthrospira platensis*	30 ± 2	1440	150	Antimicrobial Cytotoxic activity	[74]
7.	*Chlorella extract*	58	60	150	Dibenzothiophene decomposition (DBT)	[81]
8.	*Gracilaria gracilis*	100	10	—	Photocatalytic phenol degradation	[116]
9.	*Chlamydomonas reinhardtii*	80	60	—	Photocatalytic activity against methyl orange (MO)	[121]
10.	*Anabaena* strain L31 Blue-green algae	25	3 days	8000 g	Sunscreens	[122]
11.	*Gracilaria edulis*	25	Overnight	8000	Anti-cancer activity	[71]
12.	*Ulva fasciata* Delile	72	180–240	4100	Shrimp' biopreservation	[75]

the NPs. ZnO NPs act as a good nano-adsorbent when there is an increase in the pH [107].

7.5. Other applications

Sunscreens and other cosmetics have started to include ZnO NPs because of the UV-A and UV-B protection they offer [108]. Organosilanes are utilised to coat ZnO NPs, giving them a stable structure that enables them to be employed in sunscreens [109]. The deterioration and agglomeration of ZnO NPs are both thwarted by the covalent link established between the NPs and organosilanes [110]. Apart from shielding against UV rays, ZnO NPs also have hemolytic characteristics. They release haemoglobin when these NPs come into contact with red blood cells in the blood. The effects of ZnO NPs were studied using osmotic fragility, morphological characteristics, and hemolysis [70]. Higher concentrations (100 mg ml^{-1}) were shown to be detrimental to human health, while lower concentrations (50 mg ml^{-1}) were determined to be non-hazardous. This means that the hemolytic activity depends on the dosage (table 3) [70]. ZnO NPs were used to increase astaxanthin synthesis. The highest level of astaxanthin production was achieved when *Haematococcus pluvialis* was subjected to 100 μg ml^{-1} ZnO NPs. ZnO NPs were harmful to the microalgae at concentrations as low as 200 μg ml^{-1} [111]. For both natural rubber (NR) and nitrile rubber (NBR), ZnO NPs speed up the curing process. It was found that the NPs of ZnO varied in size from 30 nm to

69 nm [112]. There was a reduction in the swelling ratio of both the NR and NBR, illustrating enhanced mechanical and dynamic mechanical capabilities [112]. Contribution of these NPs is not only restricted to the above mentioned activities but they also contribute by stopping fouling agents from growing. This antifouling property saves the surface of the substrates from the attack of fouling agents, thus limiting/restricting unwanted accumulation of microbes [113]. ZnO NPs in the 10–15 nm size range were reported to have antibacterial activity against the *Pseudomonas aeruginosa* strain of bacterium [114]. Their activity has also been tested against pests found in tomato plants. ZnO NPs also acted as an ecofriendly nano-insecticide where they worked against *Spodoptera litura* and *Macrosiphum euphorbiae* generally known as tobacco cutworm and potato aphid respectively [115]. Recently, the photocatalytic activity of ZnO NPs has also gained importance. ZnO NPs synthesised from *Gracilaria gracilis* possessed photocatalytic activity and helped in the degradation of phenols (table 3) [116]. Similarly, ZnO nanocomposites formed by using reduced graphene oxide (ZnO-rGO) from *Sesbania bispinosa* also possessed photocatalytic properties towards methylene blue dye present in water [117]. Another approach was made recently in 2022, where ZnO NPs were doped with Cu to increase the photocatalytic efficiency of ZnO NPs. The Cu-doped-ZnO NPs degraded 90% of MB dye in 120 min [118]. In 2020, a very unique approach was made to synthesise Ni-doped-ZnO nanomaterial where the electrocatalytic activity was observed. It was observed that the Ni-doped-ZnO nanomaterial acted as an electrocatalyst for oxygen reduction reactions (ORR) in fuel cells [119]. Other recent applications of ZnO NPs involve wound healing properties, inhibitors of osteoclasts formation, bone formation, and acting as an anti-nociception agent [20]. The ZnO NPs synthesised from *Allium saralicum* reduced the wound area in rats and raised the fibrocyte, fibroblast, and hexuronic acid around the wound area [20]. Table 3 describes the synthesised algae-based zinc oxide NPs and the various fields in which they can be used.

7.5.1. Antimicrobial textiles. Utilising ZnO NPs to make adjustments to textiles was discovered to be greatly facilitated by their one-of-a-kind and advantageous qualities. In one method, the antibacterial capacity of cotton fabrics was investigated by exposing them to ZnO NPs at a concentration of 20 ppm. Cotton fabrics were found to have a low level of antibacterial activity against a variety of microorganisms, including *E. coli, S. aureus, B. subtilis,* and *P. aeruginosa*. The percentage of bacterial growth that was inhibited ranged from 82.2%, 81.9%, 74.8%, and 75.4%, respectively [78, 120]. Treating cotton increases its resistance to UV-A by around 76.3% and UV-B by about 85.4%, compared to the untreated fabrics having 79% and 67% blockage [78]. ZnO NPs act as UV absorbers and protect the human body from harmful UV radiations. It helped in increasing the mechanical strength of the fabrics (cotton, and wool), increasing their mechanical strength and reducing their elasticity [24].

8. Advantages and limitations

The biocompatible, ecofriendly, contaminant free, cost-friendly, and non-toxic nature of phycosynthesised ZnO NPs produced excellent results when used for different applications. Phycosynthesised ZnO NPs provided huge advantages and played a significant part in deteriorating microbes, pollutants, reducing growth of cancer cells, having antidiabetic activity and working as antioxidants. Apart from this, they also possessed photocatalytic activity, added in cosmetics as UV protector, osteoclast inhibitor, and many more. Moreover, among all the metal oxide NPs, ZnO NPs are considered as human-friendly; however, this area needs additional research and initiatives for the determination of the other possibilities of ZnO NPs. When it comes to the manufacturing of the NPs, there are five problems, both significant and minor: (a) The time it takes for algae to synthesise NPs is dependent on the strain of algae that is used in the synthesis process. It is possible that the algal strain chosen will not be efficient enough to produce NPs in a matter of hours, and instead the process would take several days or perhaps weeks. Therefore, the selection of the appropriate algal strain is an essential component; (b) Morphological properties of phycosynthesised ZnO NPs also vary. It is possible that the yield will be lower than expected or that the NPs produced will not be of high quality; (c) Caution is absolutely necessary when optimising, as it could decrease the likelihood of producing high-quality NPs. In order for the synthesis to be successful, the optimal pH, temperature, and contact duration must be delivered. In addition, the concentration of the precursor or the amount of precursor that is added determines not only the size but also the quantity of the NPs that are created; (d) Because the synthesis is a biological process that is dependent on the type of organism, the quantity of NPs synthesised by the organism, morphology of the NPs, and the longevity of the NPs have always been subject to variation; (e) Lastly, but certainly not least, the amount of research that has been carried out is extremely limited because most people are unaware of the potential of algae. As a result, expanding the use of this method to a larger scale requires time and additional research.

9. Conclusion and future prospective

The biogenic synthesis of ZnO NPs from algae is biocompatible and can be used for environmental remediation as well as biomedical application. ZnO NPs have shown promising results as an anticancer agent. Additionally, the size of these NPs with ability to generate ROS can easily move with the variable cell wall permeability of different microbes and reduce infections, thus showing promising ability as an antimicrobial agent. In addition, ZnO NPs were also used in sunscreens and in cosmetics as they offer protection against both ultraviolet radiations. Furthermore, functionalisation of

ZnO NPs with suitable functional group imparts additional properties that could be explored in the degradation/adsorption of toxic dyes and other contaminants. Recent studies have also revealed that these NPs also possess excellent photocatalytic activity, wound healing property, act as an inhibitor of osteoclasts formation, help in bone formation, and have anti-nociception activity as well. However, syntheses of ZnO NPs also have several challenges due to the nature of the process. For example, challenges associated with their aggregation and stability, as well as lack of sufficient support from algal strains during synthesis also lead to production of inferior quality NPs. Due to the ease of use, low cost, and diverse range of applications, this method of synthesis has the potential to play a pivotal role in the world of the near and distant future. These challenges can also be overcome by doping/functionalising ZnO NPs with metals, and other materials as it can increase the efficiency and enhance the activities of phycosynthesised ZnO NPs. Though this approach is at its infancy, its role in biomedicine and in the removal of pollutants has given a hope to treat several issues at very low cost and also at a bigger scale.

Acknowledgments

The authors would like to acknowledge the Department of Biotechnology for their constant support.

References

[1] Chouke P B *et al* 2022 *Materials Today Advances* **16** 100314
[2] Azizi S, Ahmad M B, Namvar F and Mohamad R 2014 *Mater. Lett.* **116** 275–7
[3] Majumdar M, Shivalkar S, Pal A, Verma M L, Sahoo A K and Roy D N 2020 *Biotechnological Production of Bioactive Compounds* ed M L Verma and A K Chandel (UK: Elsevier) pp 433–66
[4] Ali K, Dwivedi S, Azam A, Saquib Q, Al-Said M S, Alkhedhairy A A and Musarrat J 2016 *J. Colloid Interface Sci.* **472** 145–56
[5] Rajabairavi N, Raju C S, Karthikeyan C, Varutharaju K, Nethaji S, Hameed A S H and Shajahan A 2017 *Recent Trends in Materials Science and Applications* ed J Ebenezar (Switzerland: Springer International Publishing) pp 245–54 Springer Proceedings in Physics
[6] Jeetkar T J, Khataokar S P, Indurkar A R, Pandit A and Nimbalkar M S 2022 *Adv. Nat. Sci.: Nanosci. Nanotechnol.* **13** 033004
[7] Potbhare A K, Chouke P B, Mondal A, Thakare R U, Mondal S, Chaudhary R G and Rai A R 2020 *Mater. Today Proc.* **29** 939–45
[8] Negi S and Singh V 2018 *Journal of Applied and Natural Science* **10** 1134–40
[9] Agarwal H, Venkat Kumar S and Rajeshkumar S 2017 *Resource-Efficient Technologies* **3** 406–13
[10] Jain A 2022 *Adv. Nat. Sci.: Nanosci. Nanotechnol.* **13** 043001
[11] Zhang X, Yan S, Tyagi R D and Surampalli R Y 2011 *Chemosphere* **82** 489–94
[12] Shahzadi I *et al* 2023 *Int. J. Biol. Macromol.* **235** 123874
[13] Ikram M *et al* 2022 *RSC Adv.* **12** 32142–55
[14] Moeen S, Ikram M, Haider A, Haider J, Ul-Hamid A, Nabgan W, Shujah T, Naz M and Shahzadi I 2022 *ACS Omega* **7** 46428–39
[15] Shahzadi A, Moeen S, Khan A D, Haider A, Haider J, Ul-Hamid A, Nabgan W, Shahzadi I, Ikram M and Al-Shanini A 2023 *ACS Omega* **8** 8605–16
[16] Ikram M *et al* 2023 *Int. J. Biol. Macromol.* **230** 123190
[17] Ikram M *et al* 2023 *Surfaces and Interfaces* **37** 102710
[18] Shahzadi I *et al* 2022 *Int. J. Biol. Macromol.* **220** 1277–86
[19] Bhattacharya P, Chatterjee K, Swarnakar S and Banerjee S 2020 *Advanced Nano Research* **3** 15–27
[20] Chaudhary R, Haldar A, Dadure K and Kar Mahapatra D 2022 *Jordan Journal Of Physics* **15** 67–79
[21] Chand M, Barthwal S, Rawat A S, Khanuja M and Rawat S 2023 *Adv. Nat. Sci.: Nanosci. Nanotechnol.* **14** 015014
[22] Mansour A T, Alprol A E, Khedawy M, Abualnaja K M, Shalaby T A, Rayan G, Ramadan K M A and Ashour M 2022 *Materials* **15** 5169
[23] Sanaeimehr Z, Javadi I and Namvar F 2018 *Cancer Nanotechnol.* **9** 3
[24] Becheri A, Dürr M, Lo Nostro P and Baglioni P 2008 *J. Nanopart. Res.* **10** 679–89
[25] Zafar M N, Dar Q, Nawaz F, Zafar M N, Iqbal M and Nazar M F 2019 *Journal of Materials Research and Technology* **8** 713–25
[26] Yang J, Hou B, Wang J, Tian B, Bi J, Wang N, Li X and Huang X 2019 *Nanomaterials (Basel)* **9** 424
[27] Chaudhary R, Nawaz K, Khan A K, Hano C, Abbasi B H and Anjum S 2020 *Biomolecules* **10** 1498
[28] Siddiqi K S and Husen A 2016 *Nanoscale Res. Lett.* **11** 363
[29] Kalpana V N, Kataru B A S, Sravani N, Vigneshwari T, Panneerselvam A and Devi Rajeswari V 2018 *OpenNano* **3** 48–55
[30] Ebadi M, Zolfaghari M R, Aghaei S S, Zargar M, Shafiei M, Zahiri H S and Noghabi K A 2019 *RSC Adv.* **9** 23508–25
[31] Hassan S S M, Abdel-Shafy H I and Mansour M S M 2019 *Arabian J. Chem.* **12** 4074–83
[32] Potbhare A K, Chaudhary R G, Chouke P B, Yerpude S, Mondal A, Sonkusare V N, Rai A R and Juneja H D 2019 *Materials Science and Engineering:* C **99** 783–93
[33] Sabir S, Arshad M and Chaudhari S K 2014 *The Scientific World Journal* **2014** e925494
[34] Fatimah I, Pradita R Y and Nurfalinda A 2016 *Procedia Engineering* **148** 43–8
[35] Vijayakumar S, Vinoj G, Malaikozhundan B, Shanthi S and Vaseeharan B 2015 *Spectrochim. Acta, Part A* **137** 886–91
[36] Sharma S C 2016 *Optik* **127** 6498–512
[37] Nagajyothi P C, Cha S J, Yang I J, Sreekanth T V M, Kim K J and Shin H M 2015 *J. Photochem. Photobiol.,* B **146** 10–7
[38] Suresh D, Nethravathi P C, Udayabhanu, Rajanaika H, Nagabhushana H and Sharma S C 2015 *Mater. Sci. Semicond. Process.* **31** 446–54
[39] Azizi S, Mohamad R, Bahadoran A, Bayat S, Rahim R A, Ariff A and Saad W Z 2016 *J Photochem Photobiol* B **161** 441–9
[40] Dobrucka R and Długaszewska J 2016 *Saudi Journal of Biological Sciences* **23** 517–23
[41] Ramesh M, Anbuvannan M and Viruthagiri G 2015 *Spectrochim. Acta, Part A* **136** 864–70
[42] Anbuvannan M, Ramesh M, Viruthagiri G, Shanmugam N and Kannadasan N 2015 *Spectrochim. Acta, Part* A **143** 304–8
[43] Jafarirad S, Mehrabi M, Divband B and Kosari-Nasab M 2016 *Materials Science and Engineering:* C **59** 296–302
[44] Sri Sindhura K, Prasad T N V K V, Panner Selvam P and Hussain O M 2014 *Appl. Nanosci.* **4** 819–27
[45] Rauf M A, Owais M, Rajpoot R, Ahmad F, Khan N and Zubair S 2017 *RSC Adv.* **7** 36361–73

[46] Taran M, Rad M and Alavi M 2018 *Bioimpacts* **8** 81–9
[47] Saravanan M, Gopinath V, Chaurasia M K, Syed A, Ameen F and Purushothaman N 2018 *Microbial Pathogenesis* **115** 57–63
[48] Balraj B, Senthilkumar N, Siva C, Krithikadevi R, Julie A, Potheher I V and Arulmozhi M 2017 *Res. Chem. Intermed.* **43** 2367–76
[49] Abdelhakim H k, El-Sayed E R and Rashidi F B 2020 *J. Appl. Microbiol.* **128** 1634–46
[50] Shamsuzzaman, Mashrai A, Khanam H and Aljawfi R N 2017 *Arabian J. Chem.* **10** S1530–6
[51] Moghaddam A B, Moniri M, Azizi S, Rahim R A, Ariff A B, Saad W Z, Namvar F, Navaderi M and Mohamad R 2017 *Molecules* **22** 872
[52] Mohamed A A, Fouda A, Abdel-Rahman M A, Hassan S E-D, El-Gamal M S, Salem S S and Shaheen T I 2019 *Biocatalysis and Agricultural Biotechnology* **19** 101103
[53] Ganesan V, Hariram M, Vivekanandhan S and Muthuramkumar S 2020 *Mater. Sci. Semicond. Process.* **105** 104739
[54] Fouda A, EL-Din Hassan S, Salem S S and Shaheen T I 2018 *Microbial Pathogenesis* **125** 252–61
[55] Patel V, Berthold D, Puranik P and Gantar M 2015 *Biotechnology Reports* **5** 112–9
[56] Mohamed S, Hashim S N and Rahman H A 2012 *Trends in Food Science & Technology* **23** 83–96
[57] Chandran P R, Naseer M, Udupa N and Sandhyarani N 2011 *Nanotechnology* **23** 015602
[58] Hasan M *et al* 2021 *Materials Science and Engineering:* C **119** 111280
[59] Fawcett D, Verduin J J, Shah M, Sharma S B and Poinern G E J 2017 *J. Nanosci.* **2017** e8013850
[60] Dhillon G S, Brar S K, Kaur S and Verma M 2012 *Crit. Rev. Biotechnol.* **32** 49–73
[61] Sahoo P C, Kausar F, Lee J H and Han J I 2014 *RSC Adv.* **4** 32562–9
[62] Mondal A, Umekar M S, Bhusari G S, Chouke P B, Lambat T, Mondal S, Chaudhary R G and Mahmood S H 2021 *Curr Pharm Biotechnol* **22** 1782–93
[63] Senapati S, Syed A, Moeez S, Kumar A and Ahmad A 2012 *Mater. Lett.* **79** 116–8
[64] Khan M N A, Choi J S, Lee M C, Kim E, Nam T J, Fujii H and Hong Y K 2008 *J. Environ. Biol.* **29** 465–9
[65] Agrawal A, Sharma A, Awasthi K K and Awasthi A 2021 *Materials Today Chemistry* **21** 100532
[66] Mata Y N, Torres E, Blázquez M L, Ballester A, González F and Muñoz J A 2009 *J. Hazard. Mater.* **166** 612–8
[67] Mahdavi M, Namvar F, Ahmad M B and Mohamad R 2013 *Molecules* **18** 5954–64
[68] Uzair B, Liaqat A, Iqbal H, Menaa B, Razzaq A, Thiripuranathar G, Fatima Rana N and Menaa F 2020 *Bioengineering* **7** 129
[69] Hoang M H, Thanh H T T, Nguyen D N, Nguyen T V, Chyshankou I, Liubimau A, Dobysh U and Kulikouskaya V 2020 *Adv. Nat. Sci.: Nanosci. Nanotechnol.* **11** 025014
[70] Thirumoorthy G S, Balasubramaniam O, Kumaresan P, Muthusamy P and Subramani K 2021 *Bio. Nano Sci.* **11** 172–81
[71] Priyadharshini R I, Prasannaraj G, Geetha N and Venkatachalam P 2014 *Appl. Biochem. Biotechnol.* **174** 2777–90
[72] Abdulwahid K E, Dwaish A S and Dakhil O A 2019 *Plant Archives* **19** 3527–32
[73] Zheng X, Yang L, Shen Q and Zhou C 2019 *Ind. Eng. Chem. Res.* **58** 7929–36
[74] El-Belely E F, Farag M M S, Said H A, Amin A S, Azab E, Gobouri A A and Fouda A 2021 *Nanomaterials (Basel)* **11** 95
[75] Sportelli M C, Gaudiuso C, Volpe A, Izzi M, Picca R A, Ancona A and Cioffi N 2022 *Reactions* **3** 423–41
[76] Khanna P, Kaur A and Goyal D 2019 *J. Microbiol. Methods* **163** 105656
[77] Nagarajan S and Arumugam Kuppusamy K 2013 *Journal of Nanobiotechnology* **11** 39
[78] Fouda A, El-Din Hassan S, Salem S S and Shaheen T I 2018 *Microb. Pathog.* **125** 252–61
[79] Govindasamy G A, Mydin R B S M N, Sreekantan S and Harun N H 2021 *Adv. Nat. Sci.: Nanosci. Nanotechnol.* **12** 015013
[80] Tanna J A, Chaudhary R G, Juneja H D, Gandhare N V and Rai A R 2015 *Bio. Nano Sci.* **5** 123–34
[81] Khalafi T, Buazar F and Ghanemi K 2019 *Sci. Rep.* **9** 6866
[82] Shivalkar S, Chowdhary P, Afshan T, Chaudhary S, Roy A, Samanta S K and Sahoo A K 2023 *Colloids Surf.,* B **222** 113054
[83] Ananthi S, Raghavendran H R B, Sunil A G, Gayathri V, Ramakrishnan G and Vasanthi H R 2010 *Food Chem. Toxicol.* **48** 187–92
[84] Sivaramakrishnan T, Swain S, Saravanan K, Kiruba Sankar R, Roy S D, Biswas L and Shalini B 2017 *TrJFAS* **17** 641–51
[85] Zhou J, Hu N, Wu Y, Pan Y and Sun C 2008 *J. Zhejiang Univ. Sci.* B **9** 721–7
[86] Chouke P B, Potbhare A K, Bhusari G S, Somkuwar S, Shaik D P M D, Mishra R K and Chaudhary R G 2019 *Advanced Materials Letters* **10** 355–60
[87] Naser R, Abu-Huwaij R, Al-khateeb I, Abbas M M and Atoom A M 2021 *Appl. Nanosci.* **11** 1747–57
[88] Jiang Y, O'Neill A J and Ding Y 2015 *J. Nanopart. Res.* **17** 180
[89] Shivalkar S, Arshad F, Sahoo A K and Sk M P 2022 *ACS Omega* **7** 33358–64
[90] Zhang L, Jiang Y, Ding Y, Povey M and York D 2007 *J. Nanopart. Res.* **9** 479–89
[91] Espitia P J P *et al* 2013 *J. Nanopart. Res.* **15** 1324
[92] Gupta J, Irfan M, Ramgir N, Muthe K P, Debnath A K, Ansari S, Gandhi J, Ranjith-Kumar C T and Surjit M 2022 *Front Microbiol.* **13** 881595
[93] Arciniegas-Grijalba P A, Patiño-Portela M C, Mosquera-Sánchez L P, Guerrero-Vargas J A and Rodríguez-Páez J E 2017 *Appl. Nanosci.* **7** 225–41
[94] Patil S S, Patil R H, Kale S B, Tamboli M S, Ambekar J D, Gade W N, Kolekar S S and Kale B B 2014 *J. Nanopart. Res.* **16** 2717
[95] Brindha R, Kandeeban R, Kamal K S, Manojkumar K, Nithya V and Saminathan K 2021 *Adv. Nat. Sci.: Nanosci. Nanotechnol.* **12** 045002
[96] Chouke P B, Potbhare A K, Dadure K M, Mungole A J, Meshram N P, Chaudhary R R, Rai A R and Chaudhary R G 2020 *Mater. Today Proc.* **29** 815–21
[97] Shaheen T I and Fouda A 2018 *Int. J. Biol. Macromol.* **106** 784–92
[98] Hiromura M and Sakurai H 2008 *Pure Appl. Chem.* **80** 2727–33
[99] Bayrami A, Parvinroo S, Habibi-Yangjeh A and Rahim Pouran S 2018 *Artif Cells Nanomed Biotechnol* **46** 730–9
[100] Mohammadi Arvanag F, Bayrami A, Habibi-Yangjeh A and Rahim Pouran S 2019 *Materials Science and Engineering:* C **97** 397–405
[101] Mandal A K, Katuwal S, Tettey F, Gupta A, Bhattarai S, Jaisi S, Bhandari D P, Shah A K, Bhattarai N and Parajuli N 2022 *Nanomaterials* **12** 3066

[102] Shivalkar S, Gautam P K, Chaudhary S, Samanta S K and Sahoo A K 2021 *J. Environ. Manage.* **281** 111750
[103] Shivalkar S, Gautam P K, Verma A, Maurya K, Sk M P, Samanta S K and Sahoo A K 2021 *J. Environ. Manage.* **297** 113322
[104] Barnes R J, Molina R, Xu J, Dobson P J and Thompson I P 2013 *J. Nanopart. Res.* **15** 1432
[105] Zhang F, Lan J, Yang Y, Wei T, Tan R and Song W 2013 *J. Nanopart. Res.* **15** 2034
[106] Shivalkar S, Singh V, Sahoo A K, Samanta S K and Gautam P K 2021 *Bioremediation for Environmental Sustainability* ed V Kumar *et al* (UK: Elsevier) pp 1–21
[107] Shaba E Y, Jacob J O, Tijani J O and Suleiman M A T 2021 *Appl. Water Sci.* **11** 48
[108] Newman M D, Stotland M and Ellis J I 2009 *Journal of the American Academy of Dermatology* **61** 685–92
[109] Shivalkar S, Verma A, Singh V and Sahoo A K 2021 *Nanotechnology in Medicine* (India: John Wiley & Sons, Ltd.) pp 259–80
[110] Yung M M N, Fougères P-A, Leung Y H, Liu F, Djurišić A B, Giesy J P and Leung K M Y 2017 *Sci. Rep.* **7** 15909
[111] Nasri N, Keyhanfar M, Behbahani M and Dini G 2021 *J. Biotechnol.* **342** 72–8
[112] Sahoo S, Maiti M, Ganguly A, Jacob George J and Bhowmick A K 2007 *J. Appl. Polym. Sci.* **105** 2407–15
[113] Hölken I, Hoppe M, Mishra Y K, Gorb S N, Adelung R and Baum M J 2016 *Phys. Chem. Chem. Phys.* **18** 7114–23
[114] Dwivedi S, Wahab R, Khan F, Mishra Y K, Musarrat J and Al-Khedhairy A A 2014 *PLoS One* **9** e111289
[115] Thakur P, Thakur S, Kumari P, Shandilya M, Sharma S, Poczai P, Alarfaj A A and Sayyed R Z 2022 *Appl. Nanosci.* **12** 3835–50
[116] Francavilla M, Pineda A, Romero A A, Colmenares J C, Vargas C, Monteleone M and Luque R 2014 *Green Chem.* **16** 2876–85
[117] Chaudhary R G, Potbhare A K, Aziz S K T, Umekar M S, Bhuyar S S and Mondal A 2021 *Mater. Today Proc.* **36** 756–62
[118] Khalid A *et al* 2022 *Bioinorganic Chemistry and Applications* **2022** e9459886
[119] Mondal A, Chouke P B, Sonkusre V, Lambat T, Abdala A A, Mondal S and Chaudhary R G 2020 *Mater. Today Proc.* **29** 715–9
[120] Singh V, Dwivedi L M, Baranwal K, Asthana S and Sundaram S 2018 *Appl. Nanosci.* **8** 1149–60
[121] Rao M D and Gautam P 2016 *Environ. Prog. Sustainable Energy* **35** 1020–6
[122] Singh G, Babele P K, Kumar A, Srivastava A, Sinha R P and Tyagi M B 2014 *J. Photochem. Photobiol., B* **138** 55–62

2023 International Conference on Network, Multimedia and Information Technology (NMITCON)

RING Domain Mutation Hinders the E3 Ligase Activity of RNF8 and Affects UBE2N Binding

1st Mehar Sahu
Department of Biotechnology,
Delhi Technological University,
Delhi-110042, India
meharsahu@gmail.com

2nd Pravir Kumar
Department of Biotechnology,
Delhi Technological University,
Delhi-110042, India
pravirkumar@dtu.ac.in

2023 International Conference on Network, Multimedia and Information Technology (NMITCON) | 979-8-3503-0082-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/NMITCON58196.2023.10276155

***Abstract*—A DNA damage response (DDR), particularly on double-stranded breaks (DSBs); RNF8 and RNF168 are crucial E3 ligases that are recruited first to the site of DNA damage. RNF8 initiates the conjugation of ubiquitin with H2A and H2AX, subsequently, RNF168 is employed to amplify the K-63-linked ubiquitin chain. These two E3 ligases are necessary for the downstream signaling where 53BP1 and BRCA1 are recruited. RNF168 directs 53BP1 involvement in non-homologous end joining (NHEJ) but is unable to save BRCA1 which is involved in homologous recombination (HR); therefore, it is RNF8 that plays a huge role in the recruitment of BRCA1. As recently BRCA1 has come up as a potential E3 ligase in Alzheimer's disease (AD) and damage to its heterodimeric partner BARD1 results in mislocalization of BRCA1 to cytosol that further joins tau lesion. Therefore, in this cumbersome signaling cascade, we took RNF8 to study. RNF8 and RNF168 plus UBE2N/UBC13 carry therapeutic importance, hence, this study tries to identify potential mutants that can affect the binding affinity of RNF8 with UBE2N, thereby, disturbing its E3 ligase activity.**

Keywords—E3 ligase, RNF8, UBE2N, BRCA1, RNF168, Alzheimer's Disease, DNA damage response

I. INTRODUCTION

Eukaryotic cells have developed a very delicate mechanism to spot and repair different types of DNA damage, jointly known as DNA damage response (DDR). This in turn triggers cell cycle checkpoints to stop further development of cells and activate the DNA damage repair mechanism. Double-stranded breaks (DSBs) are one of the most lethal types of DNA damage. Where ATM initiates the response to DSBs which results in the phosphorylation of H2AX. Secondly, MDC1 directly interacts with γ-H2AX and amplifies the signal of DDR. Thereafter, RNF8 and RNF168 E3 ligases are recruited via MDC1. Chromatin bounded RNF8 attaches with E2 conjugating enzyme, UBE2N/UBC13 to ubiquitinate damaged chromatin. On the other hand, a ubiquitinated target X is identified by RNF168 that monoubiquitinates K13-15 on H2A [1]. Both the E3 ligases orchestrate to extend the ubiquitin chain on H2A and maintain genome integrity. However, the main purpose to target RNF8 is that it is the first E3 ligase that cooperates with UBE2N and is recruited to the site of DSB to start ubiquitination of H2AX, whereas, RNF168 recognizes the by-now ubiquitinated H2AX by RNF8. Later, RAP80 binds with the elongated chain created by RNF8 and RNF168 and forms a complex with BRCA1 (Fig. 1). Shreds of evidence suggest that BARD1 helps BRCA1 to attach with nucleosomes, however, impairment in BARD1 releases BRCA1 from the nucleosome core protein and binds with tau lesion in the cytosol which is a hallmark feature in Alzheimer's disease (AD). In this sophisticated DNA repair mechanism, a small segment of RNF 8 has been undertaken. Here, the study highlights the impact of the point mutation on RNF8 and its binding affinity with UBE2N. Here, protein-protein interaction, network analysis, prediction of binding sites, evolutionary conservation of predicted sites, mutagenesis, and docking were performed for E3 ligase RNF8.

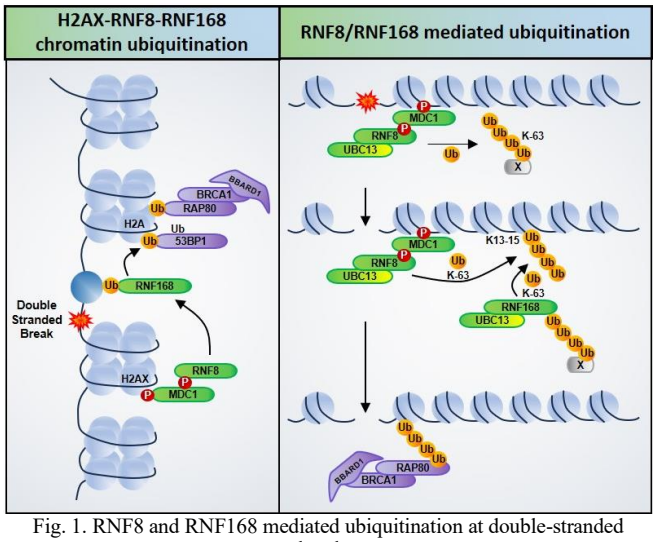


Fig. 1. RNF8 and RNF168 mediated ubiquitination at double-stranded breaks.

II. LITERATURE SURVEY

Many studies done in the past suggest that there is a strong inverse association between cancer and AD that has resulted in increased morbidity and mortality rates. Interestingly, such studies point to one common fact there is a potential link between molecular mechanisms involved in both such that in today's scenario, many anti-cancer treatments are being used against AD [2].

Breast cancer susceptibility gene 1 (BRCA1), is a tumor suppressor gene that plays a significant role in DNA damage response and many other physiological processes. However, the mutations in BRCA1 mark 80% of Breast cancer (BC) and ovarian cancer (OC) cases in females [3]. Additionally, a study also highlights increased cases of BC-AD with age in BC patients [4].

This study highlights the role of BRCA1 in regard to AD, where it has been found that BRCA1 is depleted when neuronal cultures are exposed to amyloid β (Aβ) and when human and transgenic mice are exposed to amyloid precursor protein (APP). The integral role of BRCA1 as E3 ligase involves the repair of DSBs and it is evident from past research that BRCA1 carries out ubiquitination of H2A and H2A variants at K125/K127/K129 residues. Nonetheless, the ubiquitination of H2A is governed by three major E3 ligases

979-8-3503-0082-6/23/$31.00 ©2023 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:10:12 UTC from IEEE Xplore. Restrictions apply.

namely, RNF8 and RNF168, polycomb repressive complex 1 (PRC1), and BRCA1-BARD1 heterodimer complex. These E3 ligases modify the site of H2A and enables DNA repair that further enhances chromatin compaction [5]. It has been observed that in AD, BRCA1 mislocates to cytoplastic neurofibrillary tangles (NFTs) which results in DNA fragmentation, however, BARD1 prevents BRCA1 to leave the nuclear compartment as BRCA1 contributes to DNA repair [6], [7].

The unique ubiquitination cascade that initiates on DSBs involving several E3 ligases has turned out a crucial signaling pathway that regulates DNA repair. The DDR starts with the recruitment of E3 ligase RNF8 where interaction between its FHA domain and phospho-MDC1 takes place. RNF8 forms complex with its E2 conjugating enzyme, UBE2N that ubiquitinates H2A. After this step, E3 ligase RNF168 is recruited at the DNA damage site to further amplify the effect of RNF8. Likewise, E3 ligases HERC2 and RAD18 are also called upon to intensify the ubiquitination events [8]. Finally, RAP80 is employed which keeps other DDR factors in control for instance, the involvement of BRCA1 and 53BP1 at the DNA-damaged site.

The purpose here is to study RNF8 which is one of the first E3 ligases called upon during DSBs. It has been well stated that deficiency of RNF8 is enough to cause cognitive decline, whereas, its loss potentially leads to neurodegeneration [9]. Mutation in the same might affect the forthcoming ubiquitin cascade which is to be taken forward by other E3 ligases as mentioned above. Therefore, RNF8 plays a very crucial role and could be a new therapeutic target to bring a cure to two most complicated diseases i.e., AD and BC.

III. METHODS

A. Network analysis of RNF8

The protein-protein interaction of RNF8 was performed using STRING (version 11.5) [10], where a single protein name and organism were specified as RNF8 and Homo sapiens respectively. It gave a network that summarised the association of RNF8 with other interacting partners with a confidence score. Later, gene co-expression was also studied using STRING where the intense colour of the square represents a higher association.

B. Structure retrieval and preparation of RNF8 and UBE2N

The complete structure of RNF8 is not yet available, therefore, the FASTA sequence of RNF8 was retrieved from NCBI and was submitted to ConSurf to prepare the high-resolution structure with conservation grade mapped on it [11]. RNF8 structure was validated using I-TASSER that gives top-ranked structures, estimated by C-score which is defined by equation (1) [12]. The structure of UBE2N was taken from the PDB database, 4whv.

$$\mathrm{C}-score = \ln\left(\frac{M/M_{\mathrm{tot}}}{\langle \mathrm{RMSD}\rangle} * \frac{1}{N}\sum_{i=1}^{N}\frac{z_i}{z_{\mathrm{cut},i}}\right) \quad (1)$$

C. Domain analysis

In order to understand the domain structure of RNF8 and its interacting partners, a Simple Modular Architecture Research Tool (SMART) was used that helped to understand genetically mobile domains on the basis of phyletic distribution, functional class, tertiary structure, and vital residues involved [13]. Particularly, for RNF8, its domain structure was cross-checked through UniProt (O76064). Moreover, AlphaFold which is an AI system developed by DeepMind prepared a Predicted Aligned Error (PAE) plot that further assisted in understanding useful inter-domain and intra-domain predictions [14].

D. Screening of site of interaction of UBE2N on RNF8

RNF8 is an E3 ligase and one of its potential E2 conjugating enzymes is UBE2N, therefore, it is evident that they interact with each other. Thus, the macromolecular interaction interfaces of UBE2N on RNF8 were analyzed on PDBe-KB which helped in predicting the key sites of their interaction.

E. Evolutionary conservation of interacting sites

The conservation score of each predicted site was retrieved from AlphaFold and ConSurf. The AlphaFold gave the confidence measure in the form of per-residue confidence score (pLDDT), whereas ConSurf calculated the evolutionary rates based on Multiple Sequence Alignment (MSA) which are constructed using alignment algorithms, namely, MAFFT, PRANK, T-COFFEE, MUSCLE, and CLUSTALW [15].

F. Mutational studies of RNF8

Mutagenesis of predicated sites was performed using I-Mutant 2.0 and MUpro. I-Mutant 2.0, a support vector machine-based tool assisted in predicting protein stability upon point mutation [16]. Similarly, MUpro, based on two machine learning programs namely support vector machine and neural network helped in predicting protein stability on single-site amino acid substitution [17]. Both tools gave delta delta G (DDG) values.

G. Protein-protein and protein-peptide docking

Two separate tools were used to understand RNF8 and UBE2N binding, namely, CABS-dock and HADDOCK 2.4. Protein-peptide docking was performed using CABS-dock where the targeted peptide (Wild-Type and Mutants) was uploaded separately with UBE2N that further gave cluster density, average RMSD, and Max RMSD values [18]. The CABS energy of protein-peptide contact is given by the equation (2).

$$E_{contact}(d) = \begin{cases} 0 & if\ d \le D_0 \\ s\ (d - D_0) & if\ d > D_0, \end{cases} \quad (2)$$

HADDOCK was used to dock two full-structure proteins and their binding affinity was judged based on Z-score which was calculated by equation (3) [19].

$$\mathrm{HADDOCK_{SCORE}} = 0.2 * \mathrm{E_{elec}} + 1.0 * \mathrm{E_{LJ}} + 1.0 * \mathrm{E_{AIR}} \quad (3)$$

IV. RESULTS AND DISCUSSION

A. Protein-protein interaction and gene co-expression of RNF8

Interaction between E3 ligase, RNF8, and E2 conjugating enzyme, UBE2N is indispensable for DSB and recruitment of BRCA1. However, it becomes necessary to understand other interacting partners that support this association and the STRING network analysis represented 10 networking partners of RNF8 namely UBE2N, MDC1, TP53BP1, H2AFX, RNF168, BRCA1, ATM, UIMC1, NBN, and HERC2 (Fig. 2). The p-value of this enriched network is 5.55e-16. Moreover, the number of nodes, edges, and avg.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:10:12 UTC from IEEE Xplore. Restrictions apply.

node degree of this network is 11, 55, and 10 respectively. The functional enrichment gene ontology of RNF8 suggests that it is highly involved in the histone H2A K63-linked ubiquitination biological process (GO:0070535). Molecular function suggests its direct involvement in K63-linked polyubiquitin modification-dependent protein binding (GO:0070530) and cellular component states engrossment of RNF8 in the BRCA1-A complex (GO:0070531). KEGG and Reactome pathways suggested that RNF8 is hugely involved in homologous recombination (hsa03440) and DSBs (HSA-5693548). Additionally, the subcellular localization study describes the cellular entity of RNF8 is BRCA1-BARD1 complex (GOCC:0031436). Gene co-expression analysis of RNF8 and its other interacting partners also suggested a close interaction among themselves.

B. Structural analysis of RNF8 and UBE2N

The FASTA sequence of RNF8 (485aa) was retrieved from NCBI (Accession: O76064.1 GI: 21362894), thereafter the sequence was submitted to ConSurf which gave a 3D structure of the protein. The derived 3D structure was verified through I-TASSER (Model 4, C-score -3.19). The structure of UBE2N (160aa) was retrieved from PDB (4WHV), chain B. Both the proteins are of the organism, *Homo sapiens*.

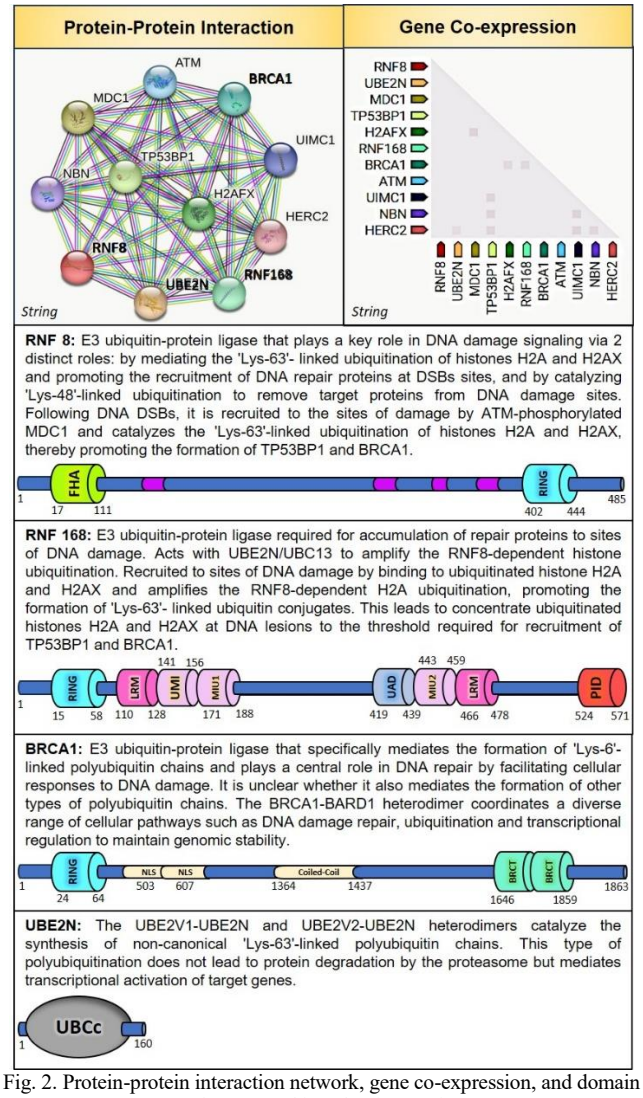


Fig. 2. Protein-protein interaction network, gene co-expression, and domain structures of RNF8 and its other interacting partners.

C. Analysis of RING domain of RNF8

The domain architecture of RNF8, RNF168, BRCA1, and UBE2N was extracted from SMART which showed the presence of the RING domain in each one of them (Fig. 2). However, keeping RNF8 as the target protein, it becomes necessary to intensely understand the domain structure of the same. Therefore, it was first checked on UniProt (O76064), where under the section of family and domain it was observed that a Zinc finger RING-type domain of RNF8 lies between 403-441aa. Later, with AlphaFold, inter and intra domains were predicted for accuracy which showed 2 distinct domains namely FHA (marked in red) (37-109aa) and RING (marked in yellow) (403-441aa) (Fig. 3). Some low-complexity regions were also observed; however, they require further validation.

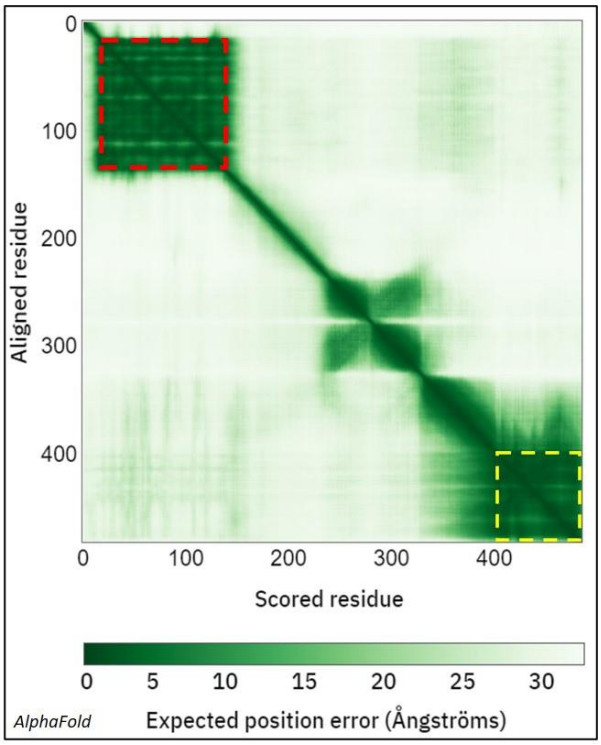


Fig. 3. Predicted Aligned Error (PAE) plot to demarcate FHA and RING domain of RNF8.

D. Interface prediction of RNF8 with UBE2N

The UniProt Id of RNF8, O76064 was searched on PDBe-KB which gave 14 interaction sites of UBE2N on RNF8 namely Q402, I404, I405, C406, S407, E408, E429, W430, R433, K434, P438, I439, R441 and R471. However, Q402 and R471 were not further analyzed as they do not lie in the RING domain. Under flexibility prediction, the results of WEBnma and DynaMine showed very less fluctuations in the residues of the RING domain as according to DynaMine, all residues have values above 0.8 suggesting a rigid conformation. Moreover, early folding residue prediction by EFoldMine depicted all key sites except R441 to have a value greater than 0.169 indicating all residues to start the early folding process. Further, through ConSurf and I-TASSER, predicted solvent accessibility of key residues exhibited whether they are structural, functional, buried, or exposed (Fig. 4). However, as for I404, it was not clear whether it is structurally or functionally involved, therefore, it was dropped for further analysis.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:10:12 UTC from IEEE Xplore. Restrictions apply.

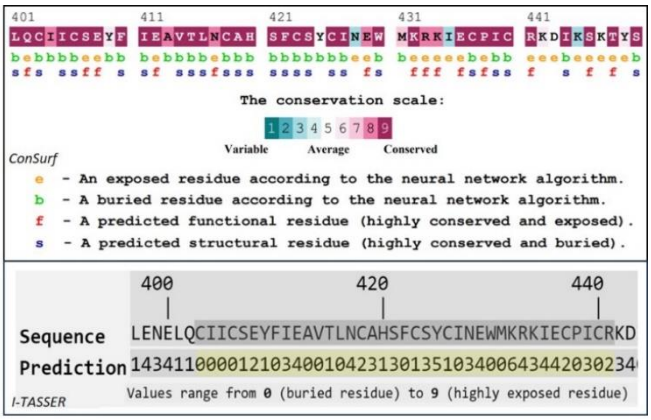


Fig. 4. Topology of predicted residues using ConSurf and I-TASSER.

E. Evolutionary conservation analysis

The evolutionary conservation score of each predicted site namely I405, C406, S407, E408, E429, W430, R433, K434, P438, I439, and R441 was first analyzed using AlphaFold which gave pLDDT score for each residue indicating that all are highly conserved. Whereas, the results of ConSurf stated that all the residues were highly conserved except for E429 and K434 which had scored a little less than the other residues (Fig. 5).

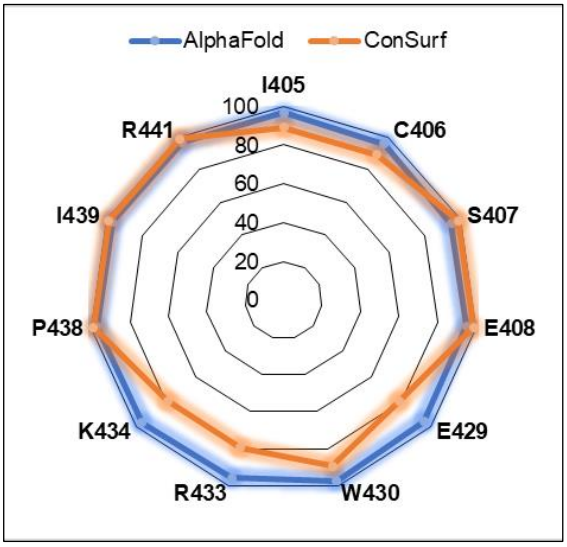


Fig. 5. Evolutionary conservation score of each precited site using AlphaFold and ConSurf.

F. Mutagenesis of qualified sites and its preparation

Despite the fluctuations in the evolutionary conservation score, all the residues were taken forth for mutational studies. The DDG values were estimated for each residue using I-Mutant 2.0 and MUpro. The average DDG value was further calculated to indicate the top mutant of each predicted site, I405 (I405G, -3.09), C406 (C406T, -1.51), S407 (S407G, -1.19), E408 (E408G, -1.34), E429 (E429G, -0.96), W430 (W430G, -2.425), R433 (R433P, -1.34), K434 (K434G, -0.95), P438 (P438A, -0.99), I439 (I439G, -1.74) and R441 (R441P, -1.61). Later, the mutant structures were prepared using PyMOL.

G. Docking analysis

1) HADDOCK2.4: Calculation of Z-score

HADDOCK2.4 server was used for protein-protein docking which gave the Z-score for each mutant of RNF8 docked with UBE2N. According to Z-score, all the mutants except P438A and R441P had Z-score less than the wild-type. P438A had similar while R441P had greater Z-score than the wild-type (Fig. 6).

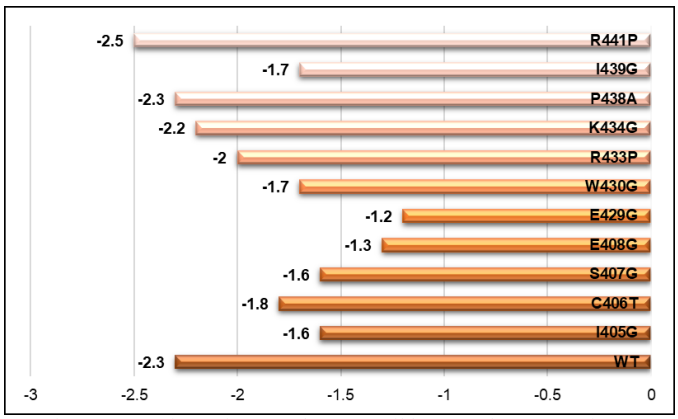


Fig. 6. Comparison of Z-score between wild-type and mutants using HADDOCK.

2) CABS-dock: Calculation of RMSD

As CABS-dock provides the platform for protein-peptide docking, therefore, the wild-type and mutant peptide sequences were prepared for each predicted residue, where the key residue was flanked by 10 residues on both sides. Contrarily, the protein structure of UBE2N was taken to be docked. The values of avg. RMSD, max RMSD, and cluster density were noted for every wild-type and mutant peptide sequence (Fig. 7). However, here the minimal avg. RMSD suggested the best binding affinity, therefore, those mutants that had avg. RMSD values less than its wild-type were dropped. Hence, suggesting that I405, S407, E408, E429, R433, P438, I439, and R441 could be potential motifs that affected the binding of UBE2N and also might hamper the E3 ligase activity of RNF8 (Table I).

TABLE I. AVERAGE RMSD OF WILD-TYPE AND MUTANTS

Sites	WT/M	Targeted Residues	Average RMSD	Max RMSD	Cluster Density
I405	Wild Type	DVLENELQCI**I**CSEYFIEAVT	2.79215	18.0271	39.038
I405G	Mutant	DVLENELQCI**G**CSEYFIEAVT	3.87994	34.6759	29.1242
C406	Wild Type	VLENELQCII**C**SEYFIEAVTL	2.16142	23.4241	47.1911
C406T	Mutant	VLENELQCII**T**SEYFIEAVTL	1.47072	7.26077	40.7964
S407	Wild Type	LENELQCIIC**S**EYFIEAVTLN	5.2558	23.2859	15.6018
S407G	Mutant	LENELQCIIC**G**EYFIEAVTLN	5.472	20.252	18.4576
E408	Wild Type	ENELQCIICS**E**YFIEAVTLNC	2.24924	23.339	47.127
E408G	Mutant	ENELQCIICS**G**YFIEAVTLNC	3.1999	29.1583	36.5636
E429	Wild Type	AHSFCSYCIN**E**WMKRKIECPI	0.305078	0.769314	118.003
E429G	Mutant	AHSFCSYCIN**G**WMKRKIECPI	5.17444	13.5865	33.6268
W430	Wild Type	HSFCSYCINE**W**MKRKIECPIC	2.48757	20.0762	43.0138
W430G	Mutant	HSFCSYCINE**G**MKRKIECPIC	2.30179	23.9917	43.879
R433	Wild Type	CSYCINEWMK**R**KIECPICRKD	0.477527	0.789012	96.3295
R433P	Mutant	CSYCINEWMK**P**KIECPICRKD	2.4398	16.632	41.3968
K434	Wild Type	SYCINEWMKR**K**IECPICRKDI	2.3946	11.9272	42.1783
K434G	Mutant	SYCINEWMKR**G**IECPICRKDI	1.72293	8.44531	38.8873
P438	Wild Type	NEWMKRKIEC**P**ICRKDIKSKT	0.835786	1.52559	92.1289
P438A	Mutant	NEWMKRKIEC**A**ICRKDIKSKT	5.33359	17.1949	38.8106
I439	Wild Type	EWMKRKIECP**I**CRKDIKSKTY	1.28767	3.23257	39.6065
I439G	Mutant	EWMKRKIECP**G**CRKDIKSKTY	1.63978	11.5038	50.0066
R441	Wild Type	MKRKIECPIC**R**KDIKSKTYSL	1.60781	3.17365	51.001
R441P	Mutant	MKRKIECPIC**P**KDIKSKTYSL	4.86235	15.3933	24.0624

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:10:12 UTC from IEEE Xplore. Restrictions apply.

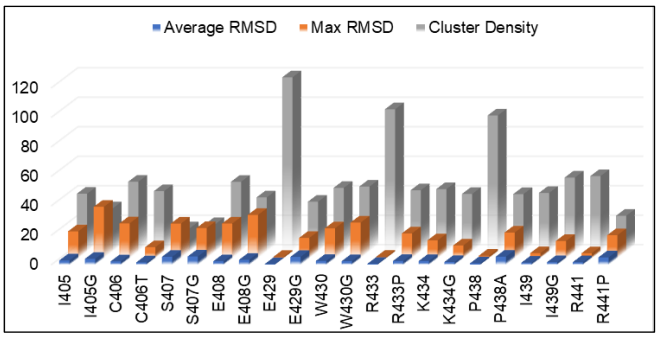


Fig. 7. Result of protein-peptide docking of wild-type and their mutants using CABS-dock.

V. Conclusion and Future Scope

Recent survey suggests BRCA1 to play a vital role in AD pathology and dysfunctioning of the same can worsen the symptoms. BRCA1 and BARD1 form a heterodimer RING that facilitates E3 ligase activity. This heterodimer RING is involved in the ubiquitination of γ-tubulin in the cytosol which is important for the centrosome complex concentration of various proteins involved in the cell cycle [20]. Dysfunction of BARD1 results in the subcellular mislocalization of BRCA1 which is highly insoluble and binds to tau lesion. BARD1 plays a vital role in binding BRCA1 with nucleosome core protein. However, before this happens, there is a huge possibility that the other interacting partners may not be functioning properly.

One of the significant E3 ligase during DSBs is RNF8 which binds with UBE2N and forms a K-63 linked ubiquitin chain that in turn acts as a framework for other proteins carrying ubiquitin-binding motifs such as RAP80 and another E3 ligase, RNF168. It is well stated in the literature that both E3 ligases, RNF8 and RNF168 are wanted at the site of DNA damage. RNF168 amplifies RNF8 ubiquitin activity and co-operates in extending the K-63-linked polyubiquitin chain that further recruits RAP80, BRCA1, and 53BP1 which altogether assists in DSB repair.

Multiple regulation of RNF8 and RNF168 represents potential therapeutic target, for instance, here it states that by hampering the E3 ligase activity of RNF8 could possibly compromise the functioning of further E3 ligases involved in the cascade. In such case, reversing the reaction by a deubiquitinating enzyme signifies a workable therapeutic solution.

The present study focused on a small segment of this huge ubiquitination cascade i.e., the impact of mutagenesis of RNF8 on its binding with E2 conjugating enzyme, UBE2N. The RING domain of RNF8 was taken and its interacting sites with UBE2N were carefully studied. After scrutinizing every site, few qualified for mutational studies. The result depicted that I405, S407, E408, E429, R433, P438, I439, and R441 could be possible sites that affect the binding of RNF8 with UBE2N and thereby hampering its E3 ligase activity. Further, validation is required to concrete the reliability of the results. Nevertheless, the limitation of this study is the impact of mutation on other diverse functions of RNF8, added, γH2AX foci do not always represent DSB formation; in such case how RNF8 would initiate ubiquitination.

Author's Contribution

P.K. – Conceptualization and supervision.
M.S. – Methodology, writing original draft, and data curation.

Acknowledgment

We thank the senior management of Delhi Technological University for their continuous support.

References

[1] C. Bartocci and E. L. Denchi, "Put a RING on it: Regulation and inhibition of RNF8 and RNF168 RING finger E3 ligases at DNA damage sites," *Front. Genet.*, vol. 4, no. JUL, p. 51975, Jul. 2013, doi: 10.3389/FGENE.2013.00128/BIBTEX.

[2] R. Mezencev and Y. O. Chernoff, "Risk of Alzheimer's Disease in Cancer Patients: Analysis of Mortality Data from the US SEER Population-Based Registries," *Cancers (Basel).*, vol. 12, no. 4, Apr. 2020, doi: 10.3390/CANCERS12040796.

[3] X. Fu, W. Tan, Q. Song, H. Pei, and J. Li, "BRCA1 and Breast Cancer: Molecular Mechanisms and Therapeutic Strategies," *Front. Cell Dev. Biol.*, vol. 10, p. 813457, Mar. 2022, doi: 10.3389/FCELL.2022.813457/BIBTEX.

[4] N. Petrović *et al.*, "MicroRNA networks linked with BRCA1/2, PTEN, and common genes for Alzheimer's disease and breast cancer share highly enriched pathways that may unravel targets for the AD/BC comorbidity treatment," *Comput. Biol. Chem.*, vol. 106, p. 107925, Oct. 2023, doi: 10.1016/J.COMPBIOLCHEM.2023.107925.

[5] Y. Guo *et al.*, "Histone H2A ubiquitination resulting from Brap loss of function connects multiple aging hallmarks and accelerates neurodegeneration," *iScience*, vol. 25, no. 7, Jul. 2022, doi: 10.1016/J.ISCI.2022.104519.

[6] M. M. Authiat *et al.*, "Preferential Involvement of BRCA1/BARD1, Not Tip60/Fe65, in DNA Double-Strand Break Repair in Presenilin-1 P117L Alzheimer Models," 2022, doi: 10.1155/2022/3172861.

[7] T. A. Evans *et al.*, "BRCA1 May Modulate Neuronal Cell Cycle Re-Entry in Alzheimer Disease," *Int. J. Med. Sci.*, vol. 4, no. 3, pp. 140–145, 2007, Accessed: Jul. 27, 2023. [Online]. Available: www.medsci.org.

[8] C. Liu, D. Wang, J. Wu, J. Keller, T. Ma, and X. Yu, "RNF168 forms a functional complex with RAD6 during the DNA damage response," *J. Cell Sci.*, vol. 126, no. 9, pp. 2042–2051, May 2013, doi: 10.1242/JCS.122945/-/DC1.

[9] S. Ouyang, Y. Song, Y. Tian, Y. Chen, X. Yu, and D. Wang, "RNF8 deficiency results in neurodegeneration in mice," *Neurobiol. Aging*, vol. 36, no. 10, p. 2850, Oct. 2015, doi: 10.1016/J.NEUROBIOLAGING.2015.07.010.

[10] D. Szklarczyk *et al.*, "STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D607–D613, Jan. 2019, doi: 10.1093/NAR/GKY1131.

[11] A. Ben Chorin *et al.*, "ConSurf-DB: An accessible repository for the evolutionary conservation patterns of the majority of PDB proteins," *Protein Sci.*, vol. 29, no. 1, pp. 258–267, Jan. 2020, doi: 10.1002/PRO.3779.

[12] W. Zheng, C. Zhang, Y. Li, R. Pearce, E. W. Bell, and Y. Zhang, "Folding non-homologous proteins by coupling deep-learning contact maps with I-TASSER assembly simulations," *Cell reports methods*, vol. 1, no. 3, Jul. 2021, doi: 10.1016/J.CRMETH.2021.100014.

[13] I. Letunic, S. Khedkar, and P. Bork, "SMART: recent updates, new developments and status in 2020," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D458–D460, Jan. 2021, doi: 10.1093/NAR/GKAA937.

[14] J. Jumper *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nat. 2021 5967873*, vol. 596, no. 7873, pp. 583–589, Jul. 2021, doi: 10.1038/s41586-021-03819-2.

[15] H. Ashkenazy, E. Erez, E. Martz, T. Pupko, and N. Ben-Tal, "ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids," *Nucleic Acids Res.*, vol. 38, no. Web Server issue, p. W529, Jul. 2010, doi: 10.1093/NAR/GKQ399.

[16] E. Capriotti, P. Fariselli, and R. Casadio, "I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure," *Nucleic Acids Res.*, vol. 33, no. Web Server issue, p. W306, Jul. 2005, doi: 10.1093/NAR/GKI375.

[17] J. Cheng, A. Randall, and P. Baldi, "Prediction of protein stability changes for single-site mutations using support vector machines," *Proteins*, vol. 62, no. 4, pp. 1125–1132, Mar. 2006, doi: 10.1002/PROT.20810.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:10:12 UTC from IEEE Xplore. Restrictions apply.

[18] M. Kurcinski, M. Jamroz, M. Blaszczyk, A. Kolinski, and S. Kmiecik, "CABS-dock web server for the flexible docking of peptides to proteins without prior knowledge of the binding site," *Nucleic Acids Res.*, vol. 43, no. W1, pp. W419–W424, Jul. 2015, doi: 10.1093/NAR/GKV456.

[19] R. V. Honorato *et al.*, "Structural Biology in the Clouds: The WeNMR-EOSC Ecosystem," *Front. Mol. Biosci.*, vol. 8, p. 729513, Jul. 2021, doi: 10.3389/FMOLB.2021.729513/BIBTEX.

[20] M. Wezyk and C. Zekanowski, "Role of BRCA1 in Neuronal Death in Alzheimer's Disease," *ACS Chem. Neurosci.*, vol. 9, no. 5, pp. 870–872, May 2018, doi: 10.1021/ACSCHEMNEURO.8B00149/ASSET/IMAGES/LARGE/CN-2018-00149N_0001.JPEG.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:10:12 UTC from IEEE Xplore. Restrictions apply.

Seismic Data Reconstruction Based on Double Sparsity Dictionary Learning With Structure Oriented Filtering

Lakshmi Kuruguntla, *Member, IEEE*, Vineela Chandra Dodda, *Member, IEEE*, Anup Kumar Mandpura, *Member, IEEE*, Sunil Chinnadurai, *Senior Member, IEEE*, and Karthikeyan Elumalai, *Member, IEEE*

***Abstract*—In seismic data processing, denoising and reconstruction are the two steps for identification of resources in the earth subsurface layers. The seismic data quality is affected by random noise and interference during acquisition. Further, the noisy data may be incomplete with missing traces. In this work, we propose a method for incomplete seismic data denoising and reconstruction based on double sparsity dictionary learning (DSDL) with structure oriented filtering (SOF). The main function of the DSDL step is denoising and SOF is used for residual noise attenuation and filling the missing data points. The proposed method is tested on 2-D synthetic and field datasets. The test results show that the DSDL-SOF method has better noise attenuation and reconstruction in terms of signal-to-noise ratio and mean squared error as compared to existing methods.**

***Index Terms*—Denoising, dictionary learning, double sparsity, seismic data reconstruction, structure oriented filtering (SOF).**

I. Introduction

IDENTIFICATION of resources present in the earth subsurface layers' is essential for the exploration of hydrocarbon and mineral deposits [1]. However, direct observation of these layers is difficult and uneconomical as the area under survey is usually vast. Seismic exploration is an efficient and economical method to determine the earth subsurface layers' information. During seismic exploration, data is collected using geophones arranged on the surface. However, the seismic data recorded by geophones is incomplete, irregular in spatial dimensions and contains anomalous traces due to imperfect instruments and the presence of noise [2], [3]. Removal of these traces from the data leads to loss of information. Furthermore, some recorded traces are lost due to obstacles in the field and ground surface limitations [4], [5], [6]. These incomplete and irregular recorded traces in seismic data severely effect the subsequent seismic data processing steps such as velocity analysis, multiples elimination, full waveform inversion, reverse time migration, amplitude variation with offset analysis, time-frequency analysis, and seismic interpretation [4], [5], [7], [8]. The missing data problems in seismic exploration are classified based on regular and irregular missing data. In regular missing data problem, the data is equidistant and missing at constant rate in a uniform grid. Whereas in irregular missing data problem, the data is randomly missing along the same uniform grid and therefore, challenging to reconstruct [2]. Therefore, reconstruction of missing data is important for accurate imaging of subsurface structures and removal of sampling artifacts.

In literature, four types of reconstruction (or) interpolation methods have been proposed for seismic data based on: Prediction Filtering (PF), Wave Equations (WE), Rank reduction (RR), and compressive sensing. The PF-based seismic data reconstruction exploits the linear predictability property of the seismic signal in different domains, namely the frequency-space (f-x), time-space (t-x), and frequency-wavenumber (f-k) domains. PF method in the f-x domain considers spatially sampled seismic data and performs the reconstruction in each frequency slice based on linear prediction (LP) theory [9]. In [10], the frequency-space domain PF method was extended to f-x-y domain in which data samples from each frequency slice are considered selectively for the designing the LP operator. Further in [11], a prediction error filter (PEF) was proposed for reconstruction of seismic data with multiple dip events having different slopes. The major limitation of PF-based methods is the need for equally spaced seismic data for reconstruction of missing traces [5].

The second category of seismic data reconstruction is based on WE techniques. The WE techniques used the Normal move out (NMO) and inverse dip move out (IDMO) jointly and performs the missing data reconstruction with the prior knowledge of the surface parameters. NMO and IDMO techniques utilized the seismic wave physical properties and depends on subsurface velocity model. The performance of seismic data reconstruction based on wave equations is dependent on the accurate velocity model which is difficult to obtain [3]. Also reconstruction

Manuscript received 31 March 2023; revised 31 May 2023, 18 July 2023, and 8 September 2023; accepted 27 September 2023. Date of publication 10 October 2023; date of current version 23 October 2023. This work was supported by the Department of Science and Technology, Science and Engineering Research Board (DST-SERB), India through Core Research under Grant CRG/2019/001234. *(Corresponding author: Karthikeyan Elumalai.)*

Lakshmi Kuruguntla, Sunil Chinnadurai, and Karthikeyan Elumalai are with the Department of Electronics and Communication Engineering, SRM University, Amaravathi 522502, India (e-mail: lakshmi_kuruguntla@srmap.edu.in; sunil.c@srmap.edu.in; imkarthi@gmail.com).

Vineela Chandra Dodda is with the Department of Electronics and Communication Engineering, Amrita School of Engineering Amaravati, Amrita Vishwa Vidyapeetham, Amaravathi 522503, India (e-mail: c_vineela@av.amrita.edu).

Anup Kumar Mandpura is with the Department of Electrical Engineering, Delhi Technological University, New Delhi 110042, India (e-mail: kanup@dtu.ac.in).

Digital Object Identifier 10.1109/JSTARS.2023.3323362

© 2023 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

technique assumes a smooth spatial spectrum of the seismic data in WE-based techniques. Further, WE-based methods are computationally expensive and not preferred for high dimensional seismic data interpolation [12].

The third category of seismic data reconstruction is based on RR method with the prior assumption of rank. In RR method, complete data is assumed to have low rank and irregular data is assumed to have a high rank [2]. The Multichannel singular spectrum analysis (MSSA) method is a rank reduction method which transforms noisy and incomplete seismic data with individual frequency into Hankel matrix [13]. Then, a rank reduction operator known as truncated singular value decomposition (TSVD) approach is applied on the Hankel matrix to decompose the noisy-incomplete data into two different subspaces namely signal subspace and noise subspace corresponding to signal and noise [14]. The noise subspace consists of both noise and incomplete seismic data [4]. However, the MSSA method is computationally expensive for the complex data with large Hankel matrix. An efficient MSSA method, where TSVD approach is replaced with randomized QR decomposition was proposed in [15]. The MSSA method can decompose seismic data into a noise subspace and a signal-plus-noise subspace. Therefore in [16], a damped multichannel singular spectrum analysis (DMSSA) method was proposed for high dimensional seismic data reconstruction with a new type of SVD known as damped SVD was applied on Hankel matrix. DMSSA method can precisely separate the seismic data into noisy and signal subspaces. The DMSSA method performs well when the data has low signal-to-noise ratio (SNR) and high dimensionality. Therefore, the DMSSA method [16] is particularly attractive for solving the 5-D offset-vector-tile (OVT) seismic data reconstruction and regularization problems in complicated area. However, in RR-based methods the selection of rank is quite difficult. A good selection of rank values provide better reconstruction results for incomplete seismic data.

The fourth category of seismic data reconstruction method is based on compressive sensing which recovers the fully sampled data by sparsity promoting methods like sparse transforms and dictionary learning (DL). Compressive sensing methods assumes that the incomplete seismic data is less sparse in comparison to complete seismic data. The incomplete data is first transformed into sparse domain using fixed basis where useful seismic signals have large amplitude coefficients and incomplete traces along with noise have low amplitude coefficients. From these coefficients, the original seismic signal is recovered using a thresholding operator. A variety of sparse transforms with predefined basis such as Fourier [17], Wavelet [18], Curvelet [19], Radon [20], Shearlet [21], Seislet [22], and Dreamlet [23] have been studied for reconstruction of seismic data. Seislet and dreamlet are two types of fixed sparse bases commonly used in seismic data processing, specifically for denoising and interpolation tasks. Compared to f-k transform-based approaches, the seislet transform-based compressed sensing method offers significantly better results in terms of data recovery, as indicated by measures such as SNR, local similarities, and visual observations [24]. To incorporate sparsity constraints in the seislet transform domain, a fast projection onto convex sets (POCS) algorithm was proposed in [24]. However, accurate estimation of the local slope is crucial for interpolation of regularly missing traces using the seislet-based POCS algorithm. The iterative slope estimation strategy employed in the algorithm is not always be reliable and can lead to a degradation of the interpolation performance. To address this issue, a more dependable technique based on the velocity-slope transformation was introduced in [22]. This alternative approach reduces the number of iterations required for interpolation and significantly improves overall performance.

The dreamlet, is a physical wavelet introduced by Wu et al. [25] and has shown remarkable capabilities in representing seismic data. The dreamlet is localized on the light cone within the 4-D Fourier space, making it an efficient tool for estimating the true signal from noisy seismic data [26]. The high sparsity observed in the dreamlet domain further enhances its effectiveness in representing seismic signals. A damped dreamlet representation was introduced in [23] to further reduce the gap between the projected signal and the true signal. By incorporating a damping factor, the damped dreamlet representation achieves more accurate estimations of the true signal. However all the aforementioned sparse transforms use orthogonal dictionaries with fixed basis which cannot match with complex seismic data structures [27]. Hence, dictionary learning (DL) methods have been proposed in seismic data reconstruction with the dictionaries learnt from training data based on data driven tight frames (DDTF), K-means singular value decomposition (K-SVD), and sequential generalized K-means (SGK) [3], [28]. The SGK algorithm is computationally efficient and used for high dimensional seismic data reconstruction.

In addition to reconstruction, noise attenuation is also a major concern for subsequent steps in seismic exploration. Henceforth, simultaneous denoising and reconstruction were performed based on rank reduction methods in [13], [29], sparse transform methods in [26] and DL-based methods in [3], [27], [28], and [30]. However, all the aforementioned methods are unsuitable for complex structured events such as curved events in seismic data. In [31], structure-oriented filtering (SOF) was used for simultaneous removal of spike-like noise and interpolation of missing points in seismic data with complex structured events. However, double sparsity dictionary learning (DSDL) methods perform better in the presence of Gaussian distributed random noise and non-Gaussian noise in seismic data [32], [33]. In [30], a joint denoising and reconstruction method for seismic data was proposed using DSDL method in the presence of spiky noise. An extended sparse K-SVD method for dictionary update was used where the atoms are updated through weighted low rank approximation. However the method in [31] was computationally complex. In recent years, deep learning methods have gained significant attention in the field of seismic data reconstruction. By training the deep learning model on large volumes of seismic data, it is possible to reconstruct missing traces, suppress noise, and improve the overall resolution of seismic images [34], [35], [36]. Deep learning models require a large amount of labeled training data to learn and generalize effectively. However, acquiring labeled seismic data for training can be challenging, time-consuming and expensive. To address

this problem, unsupervised learning or self-learning techniques are proposed in [37] and [38] for sparse time frequency analysis of seismic data.

In this work, we propose to use a computationally efficient DSDL method as a reconstruction operator in weighted iterative projection on convex sets algorithm with analytic dictionary as fast discrete curvelet transform (FDCT) and adaptive dictionary as sequential generalized k-means (SGK). Sparse representation of the data in the dictionary domain does not always guarantee a successful separation between signal and noise. Spatial coherence of the training data has a significant impact on the dictionary learning process [39]. In data sets with poor spatial coherence, a learned dictionary will not be able to represent a complex structure. In order to alleviate this shortcoming, we further applied structure oriented filtering (SOF) on the reconstructed data obtained from DSDL. The mean filter is to applied along the structural direction of seismic events in an approximately flattened gather and helps to preserve the useful signals. Existing methods such as FDCT, SGK, and SOF, individually cannot accurately recover the missing traces. In DSDL reconstruction, the sparse dictionary is based on a sparsity model of the dictionary atoms over a fixed dictionary. However, the selection of parameters in DSDL is also a challenging task. Further, the output of DSDL is filtered using SOF for further improvement in SNR. Also, in the filtering step we fine tune the number of iterations to get the accurate slope estimation for the reconstructed output of DSDL. In SOF stage, we have applied the partial denoised and reconstructed data (rather than direct noisy decimated data) which helps to reconstruct the seismic data with high level of accuracy. When compared to existing methods, the proposed DSDL-SOF method applied to noisy and incomplete seismic data has a marked improvement in terms of SNR and mean squared error (MSE).

The rest of this article is organized as follows. In Section II, we illustrate the mathematical prerequisites for the seismic signal model with noise and missing traces, analytic (FDCT) and adaptive dictionary learning (Fast dictionary learning) methods. In Section III, we describe our proposed methodology. In Section IV, the reconstruction performance of the proposed method is compared to that of existing methods. Finally, Section V concludes the article.

II. Mathematical Prerequisites

This section describes the mathematical prerequisites that will be used in the subsequent sections.

A. Signal Model

In this section, we describe the signal model for incomplete seismic data with random noise. Let $X \in \mathbb{R}^{M\times N}$ be the complete seismic data with N traces having M data points each, mathematically expressed as

$$X = [T_1, T_2, \cdots T_N] \tag{1}$$

where $T_i \in \mathbb{R}^{M\times 1}$ is the clean ith trace. In the process of acquisition, the seismic signal recorded by geophones is afflicted by noise and also some traces are lost. Let $W \in \mathbb{R}^{M\times N}$ be the additive random noise with Gaussian distribution and $S \in \mathbb{R}^{M\times M}$ be the sampling matrix with L nonzero diagonal entries and $M - L$ diagonal elements with zeros. Then, the observed noisy and incomplete seismic dataset $O \in \mathbb{R}^{M\times N}$ can be expressed as

$$O = S(X + W) \tag{2}$$

where $M - L$ traces have been lost.

B. Fast Discrete Curvelet Transform

The FDCT is one of the efficient transform used for denoising and reconstruction of seismic data. FDCT is a multiscale and multidirectional localized transform used for interpolation with POCS [40]. FDCT is also used as sparsity promoting filter for noise attenuation in seismic data [33], [41] [42]. We assume that clean seismic data X has sparse representation d in FDCT domain C is represented as

$$X = C^T d. \tag{3}$$

Then, the (2) is reformulated as

$$O = S(C^T d + W) = Ad + \tilde{W} \tag{4}$$

where $A = SC^T$, denotes the measurement matrix and $\tilde{W} = SW$. The sparse signal that can be separated from noise is then solved by minimizing the cost function as follows:

$$\hat{d} = \arg\min_d \frac{1}{2}\|O - Ad\|_2^2 + \alpha\|d\|_1 \tag{5}$$

where α is the parameter used for regularization, $\|\|_1$ and $\|\|_2$ represents L_1 and L_2 norms, respectively. The solution of above (5), provides the denoised and reconstructed seismic data.

C. Fast Dictionary Learning

In this section, we describe the fast DL method for seismic data denoising and reconstruction. In DL, sparse coding and dictionary update are the two important steps which are iteratively performed until the convergence of data model is obtained. SGK algorithm is also known as fast DL wherein the atoms of the dictionary are updated through arithmetic average of training samples with accelerated sparse coding [27], [43]. The fast DL has been used to denoise and reconstruct seismic data at the same time [27].

In fast DL method, the observed data $O(M \times N)$ is divided into patches and the ith patch is denoted as O_i. Let $D \in \mathbb{R}^{M\times P}$ represents dictionary and $G \in \mathbb{R}^{P\times N}$ is the sparse coefficient matrix. Then, the sparse form of observed seismic data based on DL method is defined as the product D and $G = [g_1, g_2, \cdots g_N]$ and mathematically expressed as

$$O = D^n G \tag{6}$$

where n represents the iteration number. The first step of fast DL method is sparse coding which can be obtained by minimizing the cost function below.

$$\forall_j g_j^n = \arg\min_{g_j} \|O - D^n G\|_F^2$$
$$\text{subject to } \forall_j \|g_j\|_1 \leqslant q \tag{7}$$

where q denotes the sparsity level and g_j^n represents the sparse coefficient corresponding to jth column of sparse coefficient matrix, G and $\forall j$ represents all columns of sparse coefficient matrix G. To solve the optimization problem in (7), an orthogonal matching pursuit (OMP) algorithm can be used until the defined sparsity constraint level is met. However, the OMP method requires q iterations and hence, more computations. To address that problem, accelerated sparse coding method was used in SGK which computes the sparse coefficients in a single iteration [27]. After sparse coding, dictionary update is the second step in SGK where the sparse coefficients G are fixed from (7) and dictionary atoms are updated by solving the cost function given below.

$$D^{n+1} = \arg\min_{D} \|O - DG^n\|_F^2 \tag{8}$$

In SGK-DL, the cost function given in (8) is solved without performing singular value decompositions (SVD) operations. Instead, an arithmetic average of training data samples is applied for updating the dictionary atoms [27]. Let d_k be the kth atom in the dictionary, D and N^k represent the number of training samples corresponding to the kth atom, then update equation of kth atom via SGK-DL method can be defined as

$$d_k = \frac{1}{N^k} \sum_{i=1}^{N^k} O_i. \tag{9}$$

The above equation is used for updating each atom in the dictionary D. After the sparse coding and dictionary update through finite number of iterations, the denoised seismic data can be reconstructed as a linear combination of the updated atoms in the dictionary and updated sparse coefficients as follows:

$$\hat{O} = \hat{D}\hat{G} \tag{10}$$

where $\hat{D}$ and $\hat{G}$ are the updated dictionary and sparse coefficient matrices, respectively.

D. Structure-Oriented Filtering

In the presence of complex events, structure oriented filtering is performed in two steps: 1) local slope estimation; and 2) filtering. Local slope is an important characteristic that can help to separate the seismic signals from noise. Based on plane wave destruction algorithm (PWD) [44], [45], the local slope is calculated. Using this slope information, the seismic data are translated to the flattened domain. Then, structural filtering is performed using a mean or median operator applied in a specific local window that is driven by the local slope at the corresponding locations [46].

The structure oriented mean filter applied to 2-D seismic data is mathematically defined [31] as

$$\hat{w}_{i,j} = \arg \min_{w_u \in W_{i,j}} \sum_{l=1}^{L+1} (w_u - w_l)^2 \tag{11}$$

where $\hat{w}_{i,j}$ is the output for observed data point at (i,j)th position in which i and j are vertical and spatial axes, respectively. $L+1$ is the filter window length, u and l are indices in filter window. The above optimization equation is equivalent to determining the mean in the filtering window. The solution of optimization equation gives the output value for the particular position (i,j) of events, where $W_{i,j}$ represents searching window with respect to slope of observed data at specified positions defined as

$$W_{i,j} = \left\{ w_{i,j-\left(\frac{L}{2}\right)\sigma_{i,j}}, \ldots, w_{i,j-\sigma_{i,j}}, w_{i,j}, w_{i,j+\sigma_{i,j}} , \ldots, w_{i,j+\left(\frac{L}{2}\right)\sigma_{i,j}} \right\} \tag{12}$$

where $\sigma_{i,j}$ represents is the local slope of the structure pattern at position i,j.

III. Methodology

In this section, we demonstrate the methodology of our work for seismic data denoising and reconstruction based on DSDL-SOF. The block diagram of the proposed method is shown in Fig. 1 which is the cascade of DSDL and SOF.

In seismic data, most of the events have complex curved structure, hence in the first step of our proposed work, we imposed the double sparsity on noisy decimated data by computational efficient fast discrete curvelet transform and fast dictionary learning algorithm. FDCT is a one type of sparse transform with high sparseness can be able to separate the signal from noise [19]. The transform-domain thresholding-based filtering works best for random noise discussed in Section II-B. However, in our work we have considered the presence of random noise in the data along with missing traces which are randomly decimated. Hence, some residual noise remaining with partial interpolation.

Therefore, we further impose the sparsity using the adaptive dictionary learning method through sequential generalized k-means algorithm given in Section II-C for the removal of the residual noise. In adaptive dictionary learning method, the atoms in the dictionary are learned from the training data which is already sparse. Therefore, the data becomes double sparse and extracts the useful seismic signal from the noise. The DSDL is a promising method for Gaussian distributed random and non-Gaussian distributed erratic noise attenuation in seismic data [32], [33]. Let $\mathcal{G}$ represents the sparse coefficient matrix, B and F are the analytic and adaptive dictionaries, respectively. The sparse representation of the observed data O based on DSDL method is mathematically expressed as

$$O = \mathcal{D}\mathcal{G} \tag{13}$$

where $\mathcal{D}$ represents the double sparsity dictionary obtained by cascading of the analytic and adaptive dictionaries, i.e.,

$$\mathcal{D} = BF. \tag{14}$$

In double sparsity dictionary, the atoms in the analytic dictionary B are fixed and predefined, and the atoms in the adaptive dictionary are updated from the sparse data provided by the fixed dictionary in adaptive manner. The sparse coefficients $\mathcal{G}$ and dictionary $\mathcal{D}$ [32] are obtained by solving the cost function below:

$$\langle \hat{\mathcal{G}}, \hat{\mathcal{D}} \rangle = \arg \min_{\mathcal{G},\mathcal{D}} \frac{1}{2} \|O - \mathcal{D}\mathcal{G}\|_2^2 + \alpha \|\mathcal{G}\|_0. \tag{15}$$

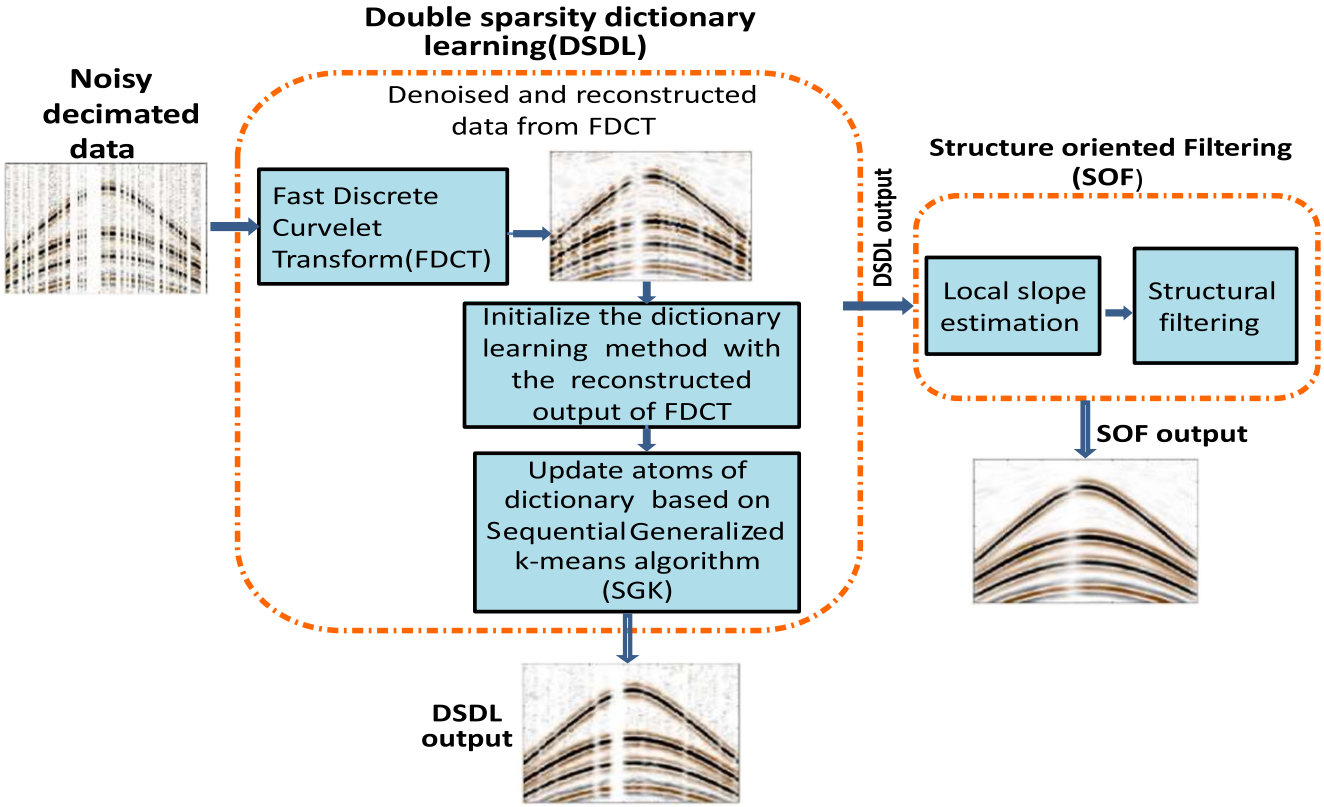


Fig. 1. Proposed DSDL with SOF method's block diagram.

The denoised seismic data R can be reconstructed using DSDL as follows:

$$R = \hat{\mathcal{D}}\hat{\mathcal{G}} \tag{16}$$

where $\hat{\mathcal{D}}$ and $\hat{\mathcal{G}}$ are updated dictionary and sparse coefficient matrices obtained using DSDL method. This DSDL method is used as a reconstruction operator in the weighted projection on convex sets algorithm for simultaneous denoising and missing traces reconstruction like in [3]. Even though, data interpolation need to be improved for further reconstruction of noiseless data with good signal preservation. Hence, finally we applied structure oriented filtering on the DSDL output with iterative slope estimation and structural filtering (Section II-D).

Let U be the SOF initialized by the DSDL output R given in (16), then the denoised and reconstructed data Y is mathematically expressed as

$$Y = U(R_{q+1}) \tag{17}$$

where q indicates the iteration number and R_{q+1} is the reconstructed data after $q+1$ iterations. The reconstructed data at $(q+1)$th iteration is obtained using the weighted iterative projection on convex sets algorithm [3] with DSDL as a reconstruction operator given as

$$R_{q+1} = \beta_q SO + (I - \beta_q S)\Gamma(O_q) \tag{18}$$

where Γ is the reconstruction operator and β_q is the relaxation parameter which decreases to 0 from 1 as the number of iterations increase. The DSDL-based reconstruction operator with sparsity constraint is represented mathematically as

$$\Gamma(O_q) = \arg\min_{\mathcal{D},\mathcal{G}} \|O - \mathcal{D}\mathcal{G}\|_2^2$$
$$\text{subject to } \forall_j g_{1j} = e_v \tag{19}$$

where e_v represents the unit vector with only vth element as 1 and remaining elements zero which is considered as sparsity constraint for DSDL method. $\mathcal{D}$ and $\mathcal{G}$ are the double sparsity dictionary and sparse coefficient matrix defined in (13), where g_1 is the coefficient vector of matrix $\mathcal{G}$.

IV. Results and Discussion

In this section, we illustrate the performance of the proposed DSDL-SOF method on incomplete synthetic and field data. We generated hyperbolic and linear crossed events similar to [47] and [48], respectively. Also, two field data examples are also used from [49] and [50]. The efficacy of our work using DSDL-SOF on synthetic data is compared based on signal to noise ratio (SNR), MSE, and local similarity map with energy [3], [33] metrics. Whereas for field data, we analyzed based on visual quality and frequency spectra. The SNR of reconstructed data can be mathematically expressed as

$$\text{SNR} = 10\log_{10}\left(\frac{\sum_k X^2(k)}{\sum_k (Y(k) - X(k))^2}\right) \tag{20}$$

where k represents the trace number, $X(k)$ is the clean seismic data, and $Y(K)$ is the reconstructed data.

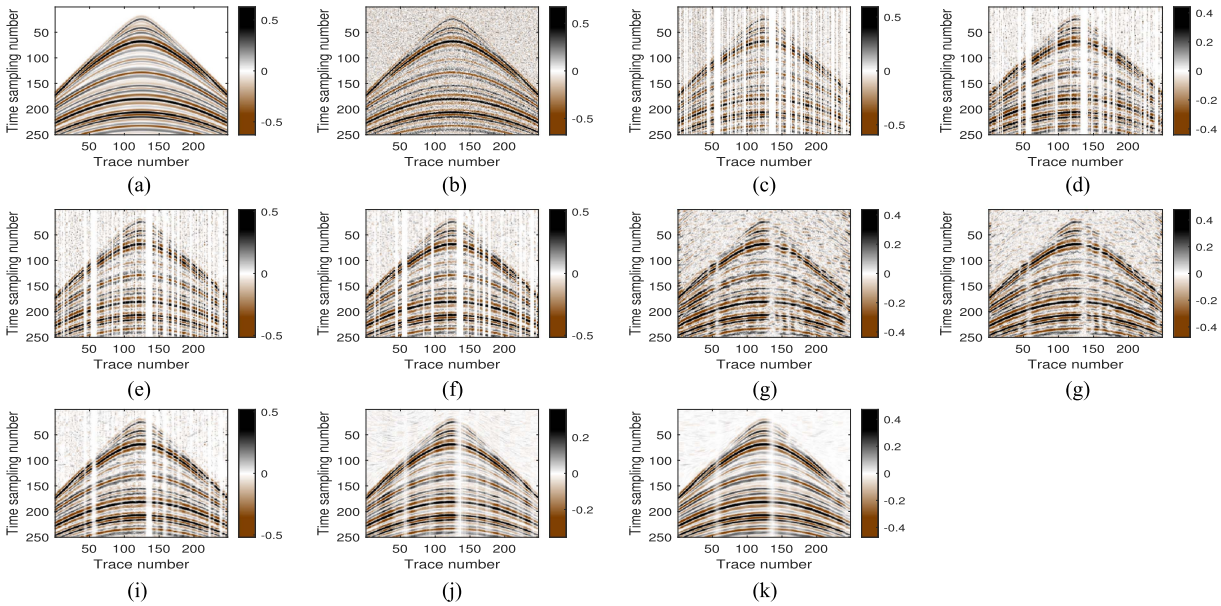


Fig. 2. Denoised and reconstruction results for existing and proposed methods on synthetic seismic data with complex hyperbolic events: (a) Clean data, (b) Noisy data, (c) noisy and incomplete seismic data with SNR 1.9863 dB, denoised and reconstructed results using:(d) DDTF, (e) K-SVD, (f) SGK, (g) MSSA, (h) DMSSA, (i) DSDL, (j) SOF, and (k) proposed DSDL-SOF.

The MSE of the reconstructed data $Y(k)$ with N traces is given by

$$\text{MSE} = \frac{1}{N}\sum_{k=1}^{N}\left(Y(k) - X(k)\right)^2. \tag{21}$$

The proposed method is studied on four different datasets as follows.

1) Synthetically generated data with complex and simple hyperbolic events;
2) synthetically generated data with linear crossed events;
3) field data-I;
4) field data-II.

A. *Study-I: Comparision of DSDL-SOF With Existing Methods on Synthetic Data With Complex Hyperbolic Events*

In this study, we analyzed the performance of proposed method with existing DL methods, structure oriented filtering method and rank-reduction methods (MSSA and DMSSA) [3] on synthetically generated seismic data with complex (more number of) hyperbolic events. The generated clean hyperbolic structured seismic events contain 250 traces with sampling duration of 4 ms as depicted in Fig. 2(a). The clean data is distorted with Gaussian distributed random noise presented in Fig. 2(b) and decimated by 50% which is shown in Fig. 2(c). Then, the noisy and incomplete synthetic data with hyperbolic events SNR is of 1.98 dB. We applied both the proposed DSDL-SOF and existing methods (DDTF, K-SVD, SGK, rank reduction methods based on MSSA and DMSSA) and individual DSDL, SOF methods on the noisy and incomplete data. Then, the obtained output SNR, MSE, energy, and computation time are tabulated in Table I.

TABLE I
COMPARISON OF PROPOSED METHOD USING DSDL-SOF ON COMPLEX HYPERBOLIC EVENTS (FIG. 2) WITH EXISTING METHODS IN TERMS OF SNR,MSE, ENERGY, AND COMPUTATION TIME

Methods	**SNR(dB)**	**MSE**	Energy	Computation time (sec)
Noisy and incomplete data	1.98	0.0241	-	-
DDTF	3.09	0.0187	0.4703	376.04
K-SVD	3.47	0.0171	0.4858	431.06
SGK	4.48	0.0136	0.5310	36.77
MSSA	4.65	0.0130	0.5618	10.32
DMSSA	5.96	0.0096	0.5922	5.76
DSDL	7.18	0.0073	0.6262	139.98
SOF	5.15	0.0116	0.6708	51.49
Proposed DSDL-SOF	8.15	0.0058	0.6764	175.17

According to Table I, DDTF, K-SVD, SGK, MSSA, DMSSA, SOF, and DSDL methods improved the SNR to 3.09 dB, 3.47 dB, 4.48 dB, 4.65 dB, 5.96 dB, 5.15 dB, and 7.18 dB, respectively. In addition, the proposed method enhances the SNR to 8.15 dB and also shows least MSE value as compared to the existing methods. The reconstructed plots of existing methods are shown in Fig. 2(d)–(h) and individual DSDL and SOF method plots are shown in Fig. 2(i)–(j) where noise can be seen in the reconstructed data and missing traces have been reconstructed partially. In Fig. 2(k), the plot of reconstructed seismic data corresponding to the proposed method based on DSDL-SOF method is presented and it is observed that it has performed better denoising and reconstruction of hyperbolic events.

It can be also observed from Fig. 2(i) that the DSDL method does not reconstruct the missing traces 55 to 59. In Fig. 2(j), SOF has a slightly better performance in reconstructing the missing

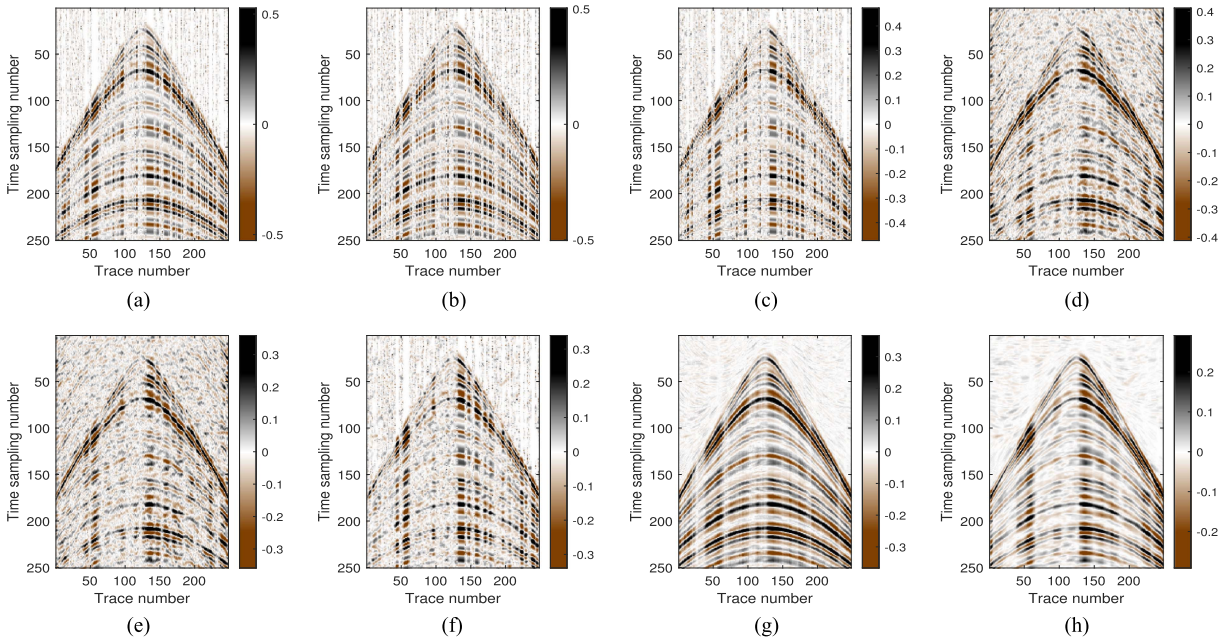


Fig. 3. Difference sections between clean and Reconstruction results for existing and proposed methods on synthetic seismic data with more hyperbolic events: Difference section of clean and reconstructed section of : (a) DDTF; (b) K-SVD; (c) SGK; (d) MSSA; (e) DMSSA (f) DSDL; (g) SOF; and (h) proposed DSDL-SOF.

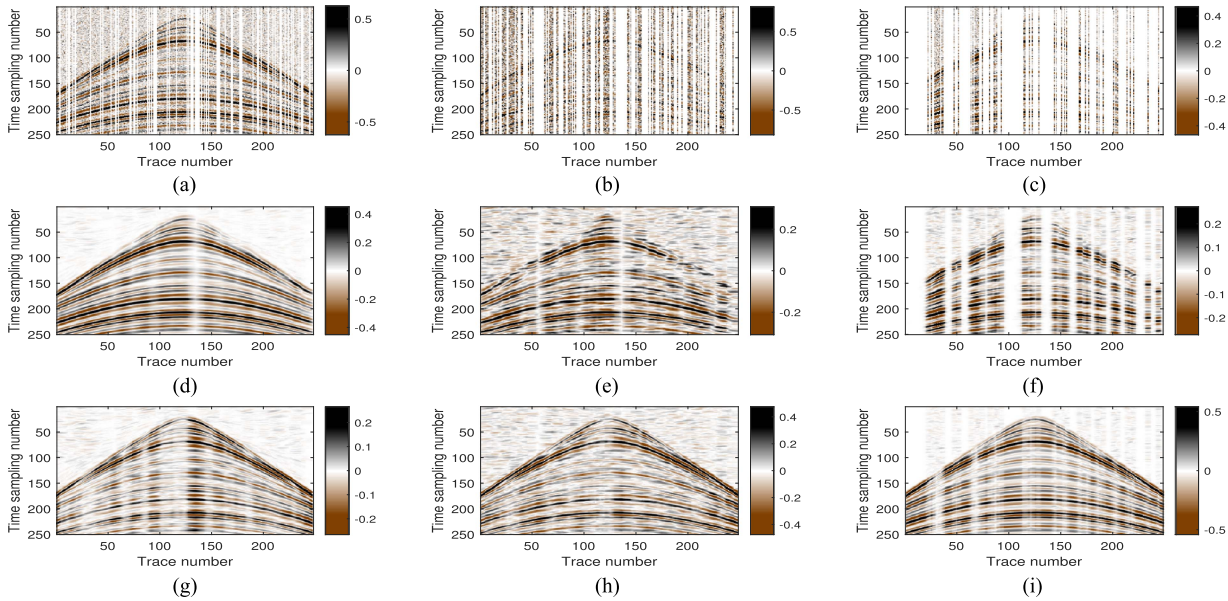


Fig. 4. Proposed method results for data with different noise levels and decimation: (a) Noisy seismic data decimated by 30% with SNR of 3.12 dB; (b) noisy seismic data decimated by 60% with SNR of –1.88 dB; (c) Noisy seismic data decimated by 80% with SNR of 0.68 dB; (d), (e), and (f) are reconstructed data of (a), (b) and (c), respectively, using the proposed method; (g), (h), and (i) are the difference sections of clean and reconstructed data of 30, 60, and 80% decimation levels.

traces 55 to 59. Similarly, the reconstruction performance for the traces 49, 50, 51, 52, 92, 93, 94, DSDL is slightly better than SOF. In Fig. 2(k), the proposed method combines the advantages of both the methods and has better reconstruction than either DSDL or SOF. The difference sections between clean and reconstructed data of all the existing and proposed methods are shown in Fig. 3(a)–(h), respectively. Further, we analyzed the performance of the proposed method on seismic data with different noise levels and sampling ratios. We generated noisy and decimated data with 30%, 60%, and 80% decimation having different noise levels (5.76 dB, –3.7 dB, and 5.82 dB). The resulting noisy and decimated data with SNR values 3.12 dB, –1.88 dB, and 0.68 dB, respectively, are shown in Fig. 4(a)–(c). The results obtained by applying the proposed method on the

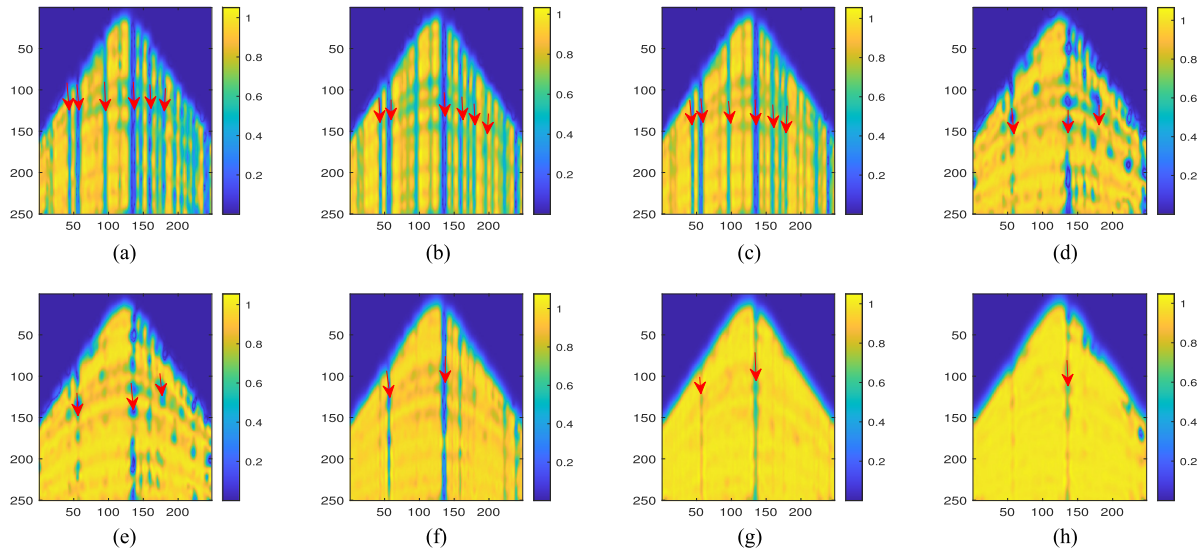


Fig. 5. Local similarity map between input and reconstructed data by (a) DDTF, (b) K-SVD, (c) SGK, (d) MSSA, (e) DMSSA, (f) DSDL, (g) SOF, and (h) proposed DSDL-SOF.

noisy decimated data are shown in Fig. 4(d)–(f), respectively. In order to illustrate the loss we consider noisy data with 50% decimation and plot the local similarity map between input and reconstructed data using the various methods which are shown in Fig. 5. The respective energy values are tabulated in Table I. From the similarity maps and Table I, we infer that the proposed method outperforms the existing methods in terms of energy. The loss in valid signal is indicated by arrows in Fig. 5 and it can be observed that signal loss in the proposed method given in Fig. 5(h) is small as compared to the existing methods. In addition, simulation results demonstrate that the DSDL method performs second best to the proposed method in terms of reconstruction and denoising. From Table I, the computation time for DSDL method is approximately 139 s and the proposed method takes approximately 175 s. While the proposed method takes slightly longer time, it achieves an improvement in SNR of nearly 1 dB as compared to DSDL method.

Furthermore, we compare the performance of our proposed method with existing methods on the synthetic data of 128 traces with simple hyperbolic events for different noise levels such as –2 dB, 0.28 dB, and 2.42 dB. The SNR and MSE of the reconstructed data with existing state-of-the-art methods and the proposed method are tabulated in Table II. From Table II, it is observed that the proposed method has an improved performance over a wide range of noise levels and decimation of seismic data.

In Fig. 6, we presented the comparative results of existing and proposed methods on synthetic seismic data with hyperbolic events of 128 traces. In Figs. 6(a)–(c) depicts noise free clean data, noisy data, noisy and incomplete data with 50% decimation having SNR of 2.42 dB, respectively. The proposed method and existing methods are applied for three different noise levels of data and the performance in terms of SNR and MSE is given in Table III. The plots of reconstructed outputs of existing and proposed methods for noisy level of 2.42 dB are shown in Fig. 6(d)–(k).

From Table III, it can be inferred that the proposed method has superior performance while compared to the existing methods

TABLE II
SNR AND MSE COMPARISON OF PROPOSED METHOD USING DSDL-SOF WITH EXISTING METHODS FOR THREE DIFFERENT NOISE LEVELS AND DECIMATION PERCENTAGES ON HYPERBOLIC EVENTS GIVEN IN FIG. 2(A)

Noisy and incomplete data with SNR(dB)	Method	Denoised and reconstructed data SNR(dB)	MSE
3.12	DDTF	4.96	0.0186
	K-SVD	5.93	0.0097
	SGK	6.54	0.0084
	MSSA	5.58	0.0105
	DMSSA	7.30	0.0071
	DSDL	7.68	0.0065
	SOF	8.61	0.0052
	Proposed DSDL-SOF	8.93	0.0050
-1.88	DDTF	-0.24	0.0403
	K-SVD	-0.28	0.0407
	SGK	1.16	0.0291
	MSSA	1.39	0.0276
	DMSSA	2.09	0.0235
	DSDL	2.10	0.0234
	SOF	2.91	0.0194
	Proposed DSDL-SOF	3.20	0.0182
0.68	DDTF	0.76	0.0320
	K-SVD	0.94	0.0306
	SGK	1.55	0.0266
	MSSA	1.25	0.0285
	DMSSA	1.34	0.0279
	DSDL	1.85	0.0024
	SOF	1.47	0.0271
	Proposed DSDL-SOF	2.11	0.0234

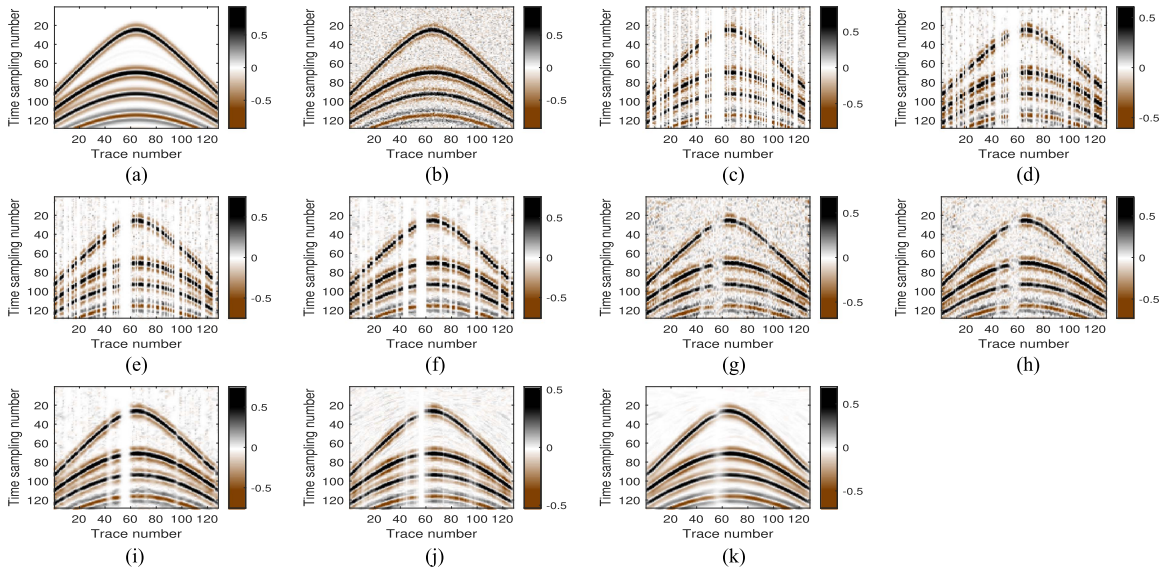


Fig. 6. Performance analysis of existing and proposed methods for data with less number of hyperbolic events: (a) Clean data with hyperbolic events, (b) noisy seismic data, (c) noisy and incomplete seismic data with SNR of 2.42 dB, denoised and reconstructed results using (d) DDTF, (e) K-SVD, (f) SGK, (g) MSSA, (h) DMSSA, (i) DSDL, (j) SOF, and (k) proposed DSDL-SOF.

TABLE III
SNR AND MSE COMPARISON OF PROPOSED METHOD USING DSDL-SOF WITH EXISTING METHODS FOR THREE DIFFERENT NOISE LEVELS ON SIMPLE HYPERBOLIC EVENTS DATA GIVEN IN FIG. 6(A)

Noisy and incomplete data SNR(dB)	Method	Denoised and reconstructed data SNR(dB)	MSE
-2	DDTF	-0.05	0.0735
	K-SVD	0.10	0.0708
	SGK	1.72	0.0488
	MSSA	1.46	0.0518
	DMSSA	1.77	0.0482
	DSDL	3.60	0.0317
	SOF	3.43	0.1173
	Proposed DSDL-SOF	5.66	0.0197
0.28	DDTF	1.23	0.0547
	K-SVD	2.24	0.0433
	SGK	3.53	0.0322
	MSSA	3.55	0.0320
	DMSSA	4.14	0.0280
	DSDL	6.03	0.0181
	SOF	4.45	0.0260
	Proposed DSDL-SOF	7.69	0.0124
2.42	DDTF	2.55	0.0403
	K-SVD	3.52	0.0323
	SGK	4.97	0.0231
	MSSA	5.39	0.0210
	DMSSA	6.65	0.0157
	DSDL	8.24	0.0109
	SOF	5.06	0.0415
	Proposed DSDL-SOF	9.05	0.0090

over wide range of noise levels in the seismic data. Further, we also compared with local similarity map between clean and reconstructed data and respective energy values are tabulated in

TABLE IV
ENERGY COMPARISON OF PROPOSED DSDL-SOF METHOD WITH EXISTING METHODS ON SIMPLE HYPERBOLIC EVENTS DATA GIVEN IN FIG. 6(A)

Methods	Energy
DDTF	0.4468
K-SVD	0.5046
SGK	0.5784
MSSA	0.5597
DMSSA	0.5912
DSDL	0.6819
SOF	0.6846
Proposed DSDL-SOF	0.7400

TABLE V
SNR AND MSE COMPARISON OF PROPOSED METHOD USING DSDL-SOF ON LINEAR CROSSED EVENTS WITH EXISTING METHODS

Methods	SNR(dB)	MSE
Noisy and incomplete data	0.44	0.0110
DDTF	1.73	0.0082
K-SVD	2.14	0.0075
SGK	3.55	0.0054
MSSA	2.47	0.0069
DMSSA	3.65	0.0053
DSDL	5.12	0.0038
SOF	5.68	0.0033
Proposed DSDL-SOF	5.85	0.0032

Table IV. From Table IV, it is clear that our proposed method has high energy value which indicates the high similarity between the clean input data and reconstructed data. The respective local similarity maps between clean and reconstructed data of existing and proposed methods are shown in Fig. 7(a)–(h).

B. Study-II: Comparision of DSDL-SOF With Existing Methods on Synthetic Data With Linear Crossed Events

In this study, we compared the proposed method results to that of existing methods on synthetic data having linear crossed

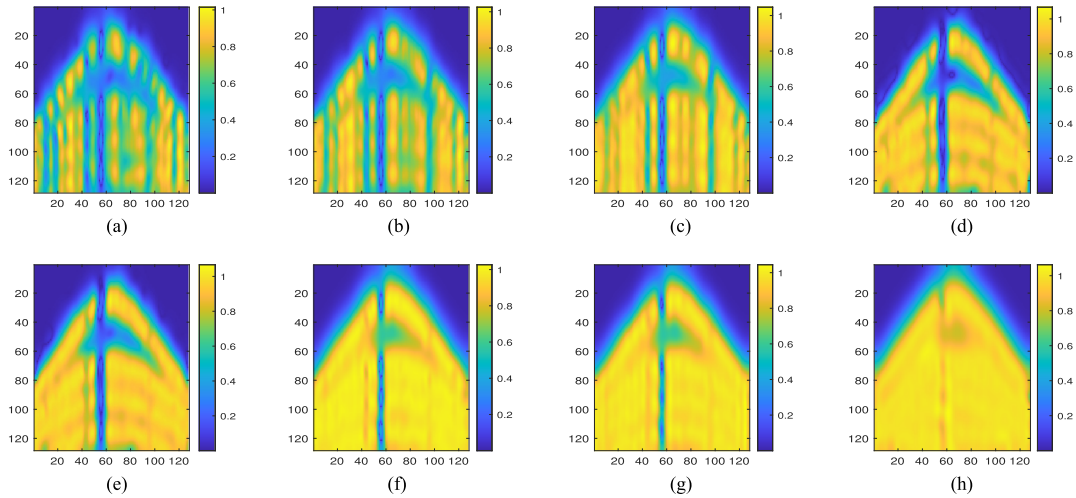


Fig. 7. Local similarity map between input and reconstructed data by (a) DDTF, (b) K-SVD, (c) SGK, (d) MSSA, (e) DMSSA, (f) DSDL, (g) SOF, and (h), proposed DSDL-SOF.

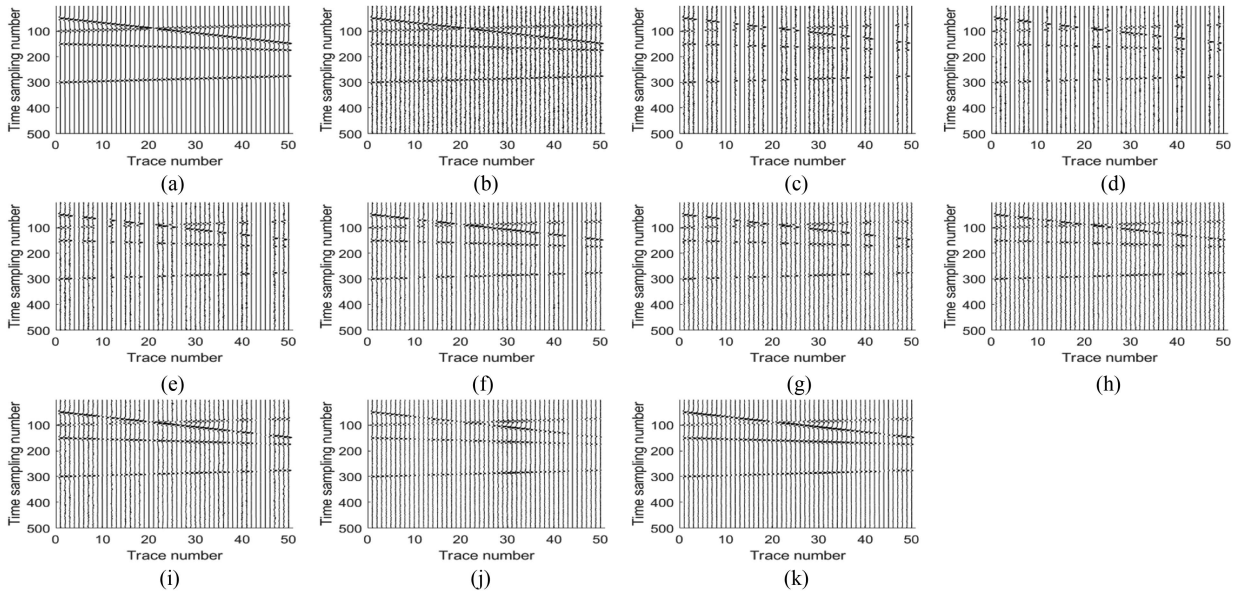


Fig. 8. Comparative results for linear-crossed events : (a) Clean data with linear-crossed events, (b) noisy seismic data with SNR 0.89 dB, (c) noisy and incomplete seismic data with SNR 0.44 dB, denoised and reconstructed data using (d) DDTF, (e) K-SVD (f) SGK, (g) MSSA, (h) DMSSA, (i) DSDL, (j) SOF, and (k) proposed DSDL-SOF.

events shown in Fig. 8. In Fig. 8(a), we presented the synthetically generated clean seismic data with linear crossed events. The clean data consists of 50 traces with sampling duration of 4 ms. The noisy data is generated by adding the Gaussian distributed Random noise to the clean data which is shown in Fig. 8(b). The noisy data is then decimated by 50% to obtain noisy incomplete data with SNR of 0.44 dB and is presented in Fig. 8(c). We applied the DDTF, K-SVD, SGK, MSSA, DMSSA, DSDL, SOF methods and the proposed method on the noisy and incomplete seismic data and the reconstructed results are shown in Figs. 8(d)–(k), respectively. The corresponding SNR and MSE values are given in Table V.

Table V clearly shows that the proposed method outperforms existing methods. The respective plot of proposed method is as shown in Fig. 8(k), where we clearly notified that the reconstructed data is more similar to the original noise free data in Fig. 8(a).

C. *Study-III: Comparision of DSDL-SOF With Existing Methods on Field Data-I*

In this study, we examine the performance of the proposed and existed methods on field data-I [49] is analyzed. The field data-I is presented in Fig. 9(a). The field data is decimated by 50% and the resultant incomplete field data is presented in Fig. 9(b). The existing and proposed methods are applied on decimated field data for noise attenuation and reconstruction. The denoised and reconstructed data of the DDTF, K-SVD, SGK, MSSA, DMSSA, DSDL, SOF and the proposed methods are presented in Fig. 9(c)–(j), respectively. From Fig. 9(j), we observed that

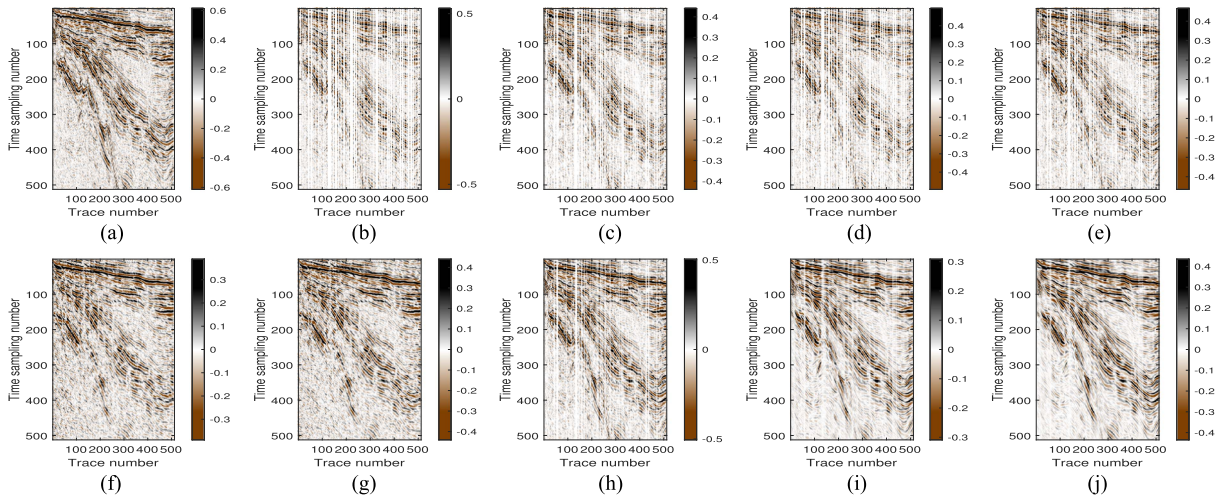


Fig. 9. Denoised and reconstructed results for field data-I (a) field data-I, (b) incomplete field data with 50% decimation, denoised and reconstructed field data using (c) DDTF, (d) K-SVD, (e) SGK, (f) MSSA, (g) DMSSA, (h) DSDL, (i) SOF, and (j) proposed DSDL-SOF.

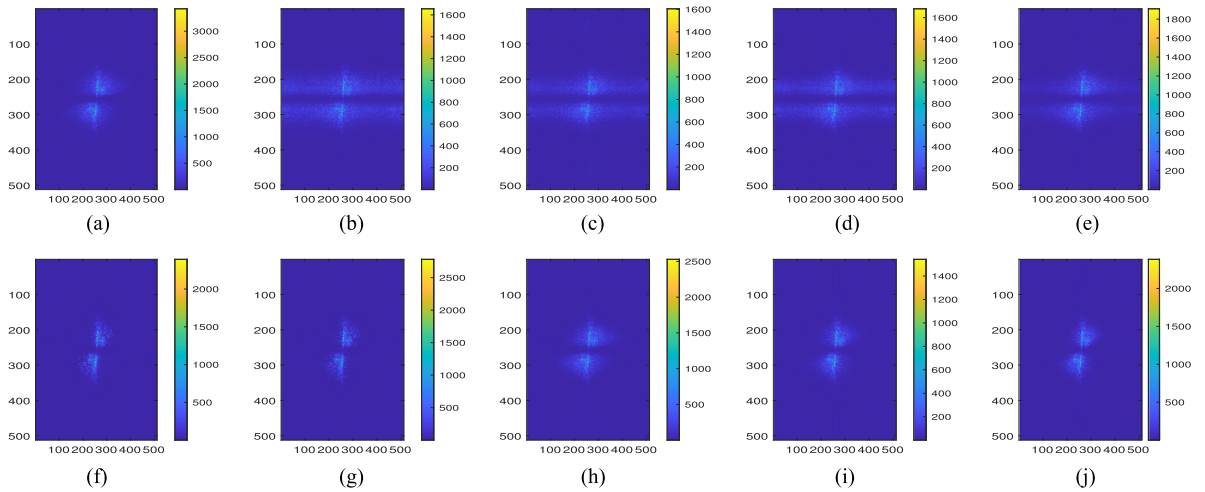


Fig. 10. Frequency spectra for input field data-I and reconstructed data through existing and proposed methods: (a) field data-I, (b) incomplete field data with 50% decimation, and reconstructed field data using (c) DDTF, (d) K-SVD, (e) SGK, (f) MSSA, (g) DMSSA, (h) DSDL, (i) SOF, and (j) proposed DSDL-SOF.

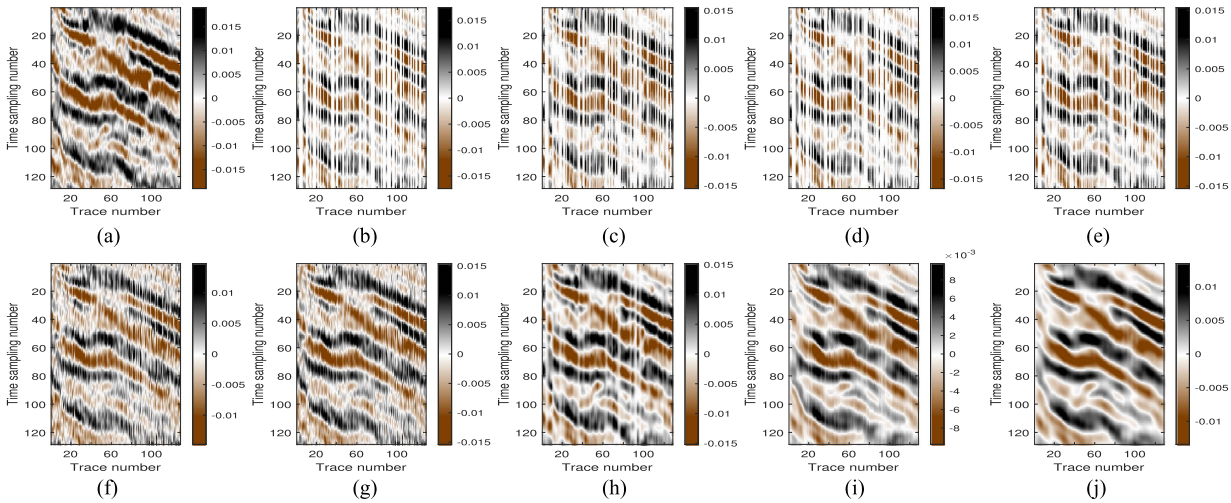


Fig. 11. Denoised and reconstructed results for (a) field data-II, (b) incomplete field data with 50% decimation, denoised and reconstructed field data using (c) DDTF, (d) K-SVD, (e) SGK, (f) MSSA, (g) DMSSA, (h) DSDL, (i) SOF, and (j) proposed DSDL-SOF.

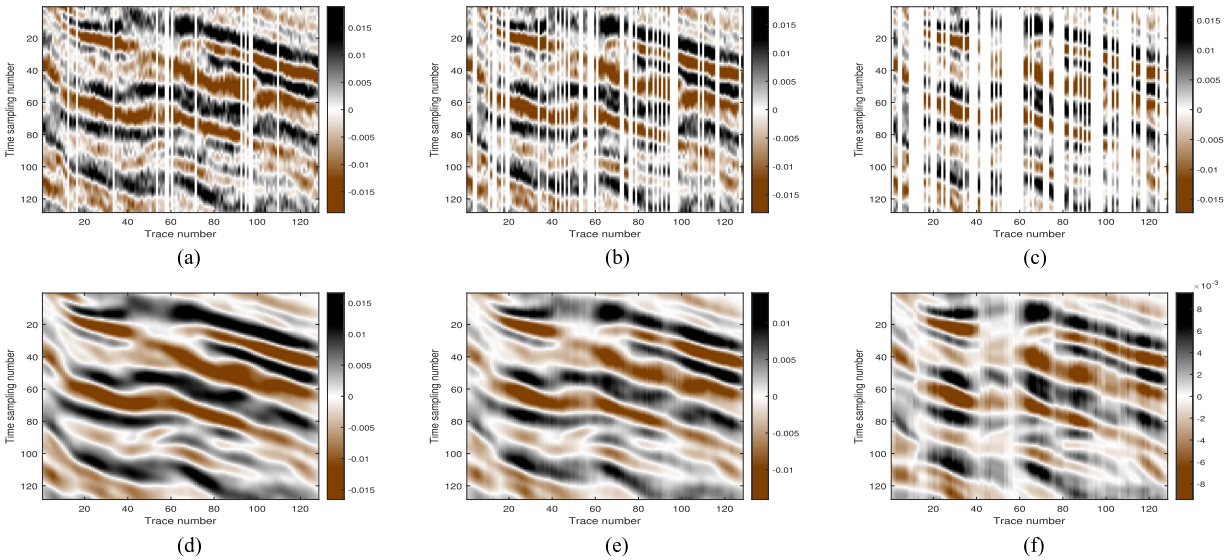


Fig. 12. Proposed method results for data with different levels of decimation: (a) Field seismic data decimated by 10%; (b) field seismic data decimated by 30%; (c) field seismic data decimated by 60%; (d), (e), and (f) are reconstructed data of (a), (b), and (c), respectively, using the proposed method.

the events in the output of proposed method are clearly visible with better noise attenuation and reconstruction. Furthermore, we also provided the frequency spectra of input field data and reconstructed data of all the existing and proposed methods in Fig. 10.

D. Study-IV: Comparision of DSDL-SOF With Existing Methods on Field Data-II

In this study, we compared the proposed method on field data-II in [50] with existed DL methods (DDTF, K-SVD, SGK, and DSDL), RR methods using MSSA and DMSSA, and filtering method based on SOF. The field data-II is depicted in Fig. 11(a). The field data decimated by 50% is presented in Fig. 11(b). The outputs of DDTF, K-SVD, SGK, MSSA, DSDL, SOF and proposed DSDL-SOF methods on incomplete field data are shown in Fig. 11(c)–(j), respectively.

The performance of the proposed method on the field data-II with three different decimation levels 10%, 30%, and 60%, respectively, shown in Fig. 12(a)–(c). The results of reconstruction from low decimation to high decimation are presented in Fig. 12(d)–(f), respectively. In Fig. 12(f), which corresponds to reconstructed data for highly decimated data, the features present in the actual are also visible.

E. Discussion

In practical scenario, the recorded seismic data is noisy and consists of missing traces. The reconstruction of irregularly decimated and noisy seismic data is a challenging problem. To address this issue, a novel method is proposed by combining the benefits of DSDL and SOF. We used the DSDL method as a reconstruction operator in weighted iterative projection on convex sets algorithm. However, the sparse representation of data in the dictionary domain does not always guarantee a successful separation between signal and noise. Spatial coherence of the training data also has a significant impact on the dictionary learning process. In datasets with poor spatial coherence, a learned dictionary will not be able to represent a complex structure. In order to alleviate this shortcoming, we further applied SOF on the reconstructed data obtained from DSDL. The mean filter is applied along the structural direction of seismic events in an approximately flattened gather and it helps to preserve the useful signals. Therefore, by combining the DSDL method and SOF method the overall performance of denoising and reconstruction has been improved. When compared to existing methods, the proposed DSDL-SOF method applied to noisy and incomplete seismic data has a marked improvement in terms of SNR and MSE. In addition, we observed aliasing in regular decimated (consecutive traces missing) seismic data with high percentage, nearly 79%. Due to aliasing, the high frequency components cannot be recovered and the edges in the reconstructed data are not sharp. Novel methods for accurate reconstruction of missing data with aliasing will be considered in the future work.

V. Conclusion

In this article, we proposed a novel method for seismic data denoising and reconstruction based on DSDL with SOF in the presence of incomplete and noisy seismic data. In our proposed method, we perform denoising and reconstruction through two cascaded steps: 1) DSDL; and 2) SOF. The first step performs denoising and reconstruction leveraging the benefits of analytic dictonary (FDCT) and adaptive dictionary (SGK), respectively. In the second step, the output of DSDL method is filtered through SOF with accurate slope estimation. Therefore, the residual noise is attenuated and simultaneous reconstruction of missing traces is performed with improved SNR. The efficacy of the proposed double sparsity DL method with SOF is studied using synthetically generated data and field data. The studies with synthetic and field data prove that the proposed method has better

denoising and reconstruction performance when compared to existing methods and has better SNR, MSE, and local similarity.

ACKNOWLEDGMENT

The authors would like to give special thanks to Prof. Yangkang Chen and his research group at the University of Texas at Austin, for their great help in providing research data, professional suggestions.

REFERENCES

[1] D. W. Steeples and R. D. Miller, "Seismic reflection methods applied to engineering, environmental, and ground-water problems," in *Proc. SEG Geotech. Environ. Geophys.*, pp. 409–461, 1988.

[2] X. Chai, H. Gu, F. Li, H. Duan, X. Hu, and K. Lin, "Deep learning for irregularly and regularly missing data reconstruction," *Sci. Rep.*, vol. 10, no. 1, pp. 1–18, 2020.

[3] H. Wang, W. Chen, Q. Zhang, X. Liu, S. Zu, and Y. Chen, "Fast dictionary learning for high-dimensional seismic reconstruction," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 7098–7108, Aug. 2021.

[4] W. Huang, D. Feng, and Y. Chen, "De-aliased and de-noise cadzow filtering for seismic data reconstruction," *Geophysical Prospecting*, vol. 68, no. 2, pp. 553–571, 2020.

[5] L. Fu, M. Zhang, Z. Liu, and H. Li, "Reconstruction of seismic data with missing traces using normalized Gaussian weighted filter," *J. Geophys. Eng.*, vol. 15, no. 5, pp. 2009–2020, 2018.

[6] S. Gan, S. Wang, Y. Chen, X. Chen, W. Huang, and H. Chen, "Compressive sensing for seismic data reconstruction via fast projection onto convex sets based on seislet transform," *J. Appl. Geophys.*, vol. 130, pp. 194–208, 2016.

[7] J. Wu, M. Bai, D. Zhang, H. Wang, G. Huang, and Y. Chen, "Fast and robust low-rank approximation for five-dimensional seismic data reconstruction," *IEEE Access*, vol. 8, pp. 175501–175512, 2020.

[8] J. Wu and M. Bai, "Adaptive rank-reduction method for seismic data reconstruction," *J. Geophys. Eng.*, vol. 15, no. 4, 2018, Art. no. 1688.

[9] S. Spitz, "Seismic trace interpolation in the FX domain," *Geophysics*, vol. 56, no. 6, pp. 785–794, 1991.

[10] Y. Wang, "Seismic trace interpolation in the FXY domain," *Geophysics*, vol. 67, no. 4, pp. 1232–1239, 2002.

[11] J. Claerbout and S. Fomel, "Geophysical image estimation by example: Stanford exploration project," Stanford Exploration Project, 2014.

[12] W. Huang and J. Liu, "Robust seismic image interpolation with mathematical morphological constraint," *IEEE Trans. Image Process.*, vol. 29, pp. 819–829, 2020.

[13] V. Oropeza and M. Sacchi, "Simultaneous seismic data denoising and reconstruction via multichannel singular spectrum analysis," *Geophysics*, vol. 76, no. 3, pp. V25–V32, 2011.

[14] M. A. N. Siahsar, S. Gholtashi, E. O. Torshizi, W. Chen, and Y. Chen, "Simultaneous denoising and interpolation of 3-D seismic data via damped data-driven optimal singular value shrinkage," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 7, pp. 1086–1090, Jul. 2017.

[15] J. Cheng and M. D. Sacchi, "Fast dual-domain reduced-rank algorithm for 3D deblending via randomized QR decomposition," *Geophysics*, vol. 81, no. 1, pp. V89–V101, 2016.

[16] W. Huang, R. Wang, Y. Chen, H. Li, and S. Gan, "Damped multichannel singular spectrum analysis for 3D random noise attenuation," *Geophysics*, vol. 81, no. 4, pp. V261–V270, 2016.

[17] H. Zhang, X.-H. Chen, and X.-M. Wu, "Seismic data reconstruction based on CS and Fourier theory," *Appl. Geophys.*, vol. 10, pp. 170–180, 2013.

[18] N. Guo, M. Chen, and Y. Cui, "Prestack seismic data reconstruction method based undecimated discrete wavelet transform," in *Proc. 77th EAGE Conf. Exhib.*, vol. 2015, no. 1. EAGE Publications BV, 2015, pp. 1–5.

[19] H. Yang, Y. Long, J. Lin, F. Zhang, and Z. Chen, "A seismic interpolation and denoising method with curvelet transform matching filter," *Acta Geophysica*, vol. 65, no. 5, pp. 1029–1042, 2017.

[20] J. Shao and Y. Wang, "Seismic data antialiasing interpolation using sparse radon transform and dynamic mask function," *Geophysics*, vol. 87, no. 5, pp. V437–V449, 2022.

[21] L. Jicheng, C. Yongxin, and Z. Jianjiang, "Interpolating seismic data via the POCS method based on shearlet transform," *J. Geophys. Eng.*, vol. 15, no. 3, pp. 852–876, 2018.

[22] W. Liu, S. Cao, S. Gan, Y. Chen, S. Zu, and Z. Jin, "One-step slope estimation for dealiased seismic data reconstruction via iterative seislet thresholding," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 10, pp. 1462–1466, Oct. 2016.

[23] W. Huang, R.-S. Wu, and R. Wang, "Damped dreamlet representation for exploration seismic data interpolation and denoising," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3159–3172, Jun. 2018.

[24] S. Gan*, S. Wang, Y. Chen, and X. Chen, "Seismic data reconstruction via fast projection onto convex sets in the seislet transform domain," in *Proc. SEG Tech. Prog. expanded Abstr. Soc. Exploration Geophysicists*, 2015, pp. 3814–3819.

[25] R.-S. Wu, Y. Geng, and B. Wu, "Physical wavelet defined on an observation plane and the dreamlet," in *Proc. SEG Int. Expo. Annu. Meeting*, 2011, pp. SEG–2011.

[26] B. Wang, R.-S. Wu, X. Chen, and J. Li, "Simultaneous seismic data interpolation and denoising with a new adaptive method based on dreamlet transform," *Geophysical J. Int.*, vol. 201, no. 2, pp. 1182–1194, 2015.

[27] J. Wu, Q. Chen, Z. Gui, and M. Bai, "Fast dictionary learning for 3D simultaneous seismic data reconstruction and denoising," *J. Appl. Geophys.*, vol. 194, 2021, Art. no. 104446.

[28] S. Yu, J. Ma, X. Zhang, and M. D. Sacchi, "Interpolation and denoising of high-dimensional seismic data by learning a tight frame," *Geophysics*, vol. 80, no. 5, pp. V119–V132, 2015.

[29] Y. Chen et al., "Simultaneous denoising and reconstruction of 5-D seismic data via damped rank-reduction method," *Geophysical J. Int.*, vol. 206, no. 3, pp. 1695–1717, 2016.

[30] L. Zhu, E. Liu, and J. H. McClellan, "Joint seismic data denoising and interpolation with double-sparsity dictionary learning," *J. Geophys. Eng.*, vol. 14, no. 4, pp. 802–810, 2017.

[31] Y. Zhou and S. Li, "Simultaneous deblending and interpolation using structure-oriented filters," *J. Appl. Geophys.*, vol. 150, pp. 230–243, 2018.

[32] Y. Chen, J. Ma, and S. Fomel, "Double-sparsity dictionary for seismic noise attenuation," *Geophysics*, vol. 81, no. 2, pp. V103–V116, 2016.

[33] L. Kuruguntla, V. C. Dodda, A. K. Mandpura, and K. Elumalai, "Erratic noise attenuation using double sparsity dictionary learning method," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5913613.

[34] N. Liu, L. Wu, J. Wang, H. Wu, J. Gao, and D. Wang, "Seismic data reconstruction via wavelet-based residual deep learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4508213.

[35] N. Liu, J. Wang, J. Gao, S. Chang, and Y. Lou, "Similarity-informed self-learning and its application on seismic image denoising," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5921113.

[36] N. Liu et al., "NS2NS: Self-learning for seismic image denoising," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5922311.

[37] N. Liu, Y. Lei, R. Liu, Y. Yang, T. Wei, and J. Gao, "Sparse time-frequency analysis of seismic data: Sparse representation to unrolled optimization," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5915010.

[38] Y. Yang, Y. Lei, N. Liu, Z. Wang, J. Gao, and J. Ding, "SparsetFNet: A physically informed autoencoder for sparse time–frequency analysis of seismic data," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4512812.

[39] H. Lv, "Sparse dictionary learning for noise attenuation in the exactly flattened dimension," *J. Seismic Exploration*, vol. 28, no. 5, pp. 449–474, 2019.

[40] H. Zhang, H. Zhang, J. Zhang, Y. Hao, and B. Wang, "An anti-aliasing pocs interpolation method for regularly undersampled seismic data using curvelet transform," *J. Appl. Geophys.*, vol. 172, 2020, Art. no. 103894.

[41] Q. Zhao, Q. Du, X. Gong, and Y. Chen, "Signal-preserving erratic noise attenuation via iterative robust sparsity-promoting filter," *IEEE Trans. Geosci. Remote Sens*, vol. 56, no. 6, pp. 3547–3560, Jun. 2018.

[42] H. Wang et al., "High-precision seismic data reconstruction with multi-domain sparsity constraints based on curvelet and high-resolution radon transforms," *J. Appl. Geophys.*, vol. 162, pp. 128–137, 2019.

[43] Y. Chen, "Fast dictionary learning for noise attenuation of multidimensional seismic data," *Geophysical J. Int.*, vol. 209, no. 1, pp. 21–31, 2017.

[44] S. Fomel, "Local seismic attributes," *Geophysics*, vol. 72, no. 3, pp. A29–A33, 2007.

[45] S. Fomel, "Applications of plane-wave destruction filters," *Geophysics*, vol. 67, no. 6, pp. 1946–1960, 2002.

[46] H. Wang et al., "A matlab code package for 2D/3D local slope estimation and structural filtering," *Geophys.*, vol. 87, no. 3, pp. 1–102, 2022.

[47] Y. Zhou, J. Yang, H. Wang, G. Huang, and Y. Chen, "Statistics-guided dictionary learning for automatic coherent noise suppression," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2020, Art. no. 5900117.

[48] G. Huang, M. Bai, Q. Zhao, W. Chen, and Y. Chen, "Erratic noise suppression using iterative structure-oriented space-varying median filtering with sparsity constraint," *Geophysical Prospecting*, vol. 69, no. 1, pp. 101–121, 2020.
[49] S. Beckouche and J. Ma, "Simultaneous dictionary learning and denoising for seismic data," *Geophysics*, vol. 79, no. 3, pp. A27–A31, 2014.
[50] O. M. Saad and Y. Chen, "Deep denoising autoencoder for seismic random noise attenuation," *Geophysics*, vol. 85, no. 4, pp. V367–V376, 2020.

Lakshmi Kuruguntla (Member, IEEE) received the B.Tech. degree in electronics and communication engineering (ECE) from Jawaharlal Nehru Technological University (JNTU), Hyderabad, India, in 2008, the M.Tech. degree in digital electronics and communication systems (DECS) from JNTU, Kakinada, India, in 2012, and the Ph.D. degree in seismic data denoising and reconstruction based on dictionary learning from SRM University, Amaravathi, India, in 2023.

She has ten years of teaching experience in reputed institutions. She is currently an Assistant Professor with Koneru Lakshmaiah Education Foundation (KLEF) deemed to be University, Guntur, India. Her research interests include seismic signal processing, compressive sensing, dictionary learning, and machine learning.

Vineela Chandra Dodda (Member, IEEE) received the B.Tech. degree in electronics and communication engineering (ECE) from Amrita Vishwa Vidyapeetham, Coimbatore, India, in 2015, the M.Tech. degree in embedded systems from KL Deemed to be University, Guntur, India, in 2019, and the Ph.D. degree in seismic data processing, inversion and interpretation from SRM University, Amaravathi, India.

From 2015 to 2017, she was an Associate Software Engineer with Robert Bosch Engineering and Business Solutions Ltd, Coimbatore, India. She is currently a Faculty Associate with the Department of ECE, Amrita Vishwa Vidyapeetham, Amaravati, India. Her research interests include neural networks, machine learning, and image processing.

Anup Kumar Mandpura (Member, IEEE) received the B.Tech degree in electronics engineering from Banaras Hindu University, Varanasi, India, in 2006, the M.Tech degree in digital signal processing from the Indian Institute of Technology, Guwahati, India, in 2009, and the Ph.D. degree in wireless communication, from the Indian Institute of Technology, Delhi, India, in 2018.

He is currently an Assistant Professor with the Department of Electrical Engineering, Delhi Technological University, Delhi, India. His research interests include performance analysis of wireless communication systems, energy harvesting communication systems, microwave/millimeter-wave systems, and signal processing.

Sunil Chinnadurai (Senior Member, IEEE) received the B.E degree in electronics and communication engineering (ECE) from Anna University, Chennai, India, in 2009, the M.S. degree in electronics and communication engineering from Mid Sweden University, Sweden, in 2012, and the Ph.D. degree in electronics and communication engineering from Jeonbuk National University, South Korea, in 2017.

For six months, he was a Postdoctoral Research Fellow with Jeonbuk National University. He was with the Signal Intelligence Research Center, Hanyang University, Seoul, South Korea, for a year, working as a Postdoctoral Research Scientist. From 2019, he is an Assistant Professor with the Department of ECE, SRM University AP, Andhra Pradesh, India. His research interests include 5G and B5G communications, intelligent reflecting surfaces, massive MIMO, NOMA, mm wave communications, machine learning, and hyperspectral image processing.

Dr. Chinnadurai was the recipient of the Best Paper Award at the 24th MSPT International Symposium in 2016. He is also a Fellow of the IETE.

Karthikeyan Elumalai (Member, IEEE) received the B.E. degreee in electronics and communication engineering and M.E. degree in communication systems from Anna University, Chennai, India, in 2007 and 2009, respectively, and the Ph.D. degree in signal processing from the Indian Institute of Technology Delhi (IIT-D), Delhi, India, in 2018.

He is currently a Assistant Professor with the Department of ECE, SRM University-AP, Amaravathi, India. Also, he received a research grant from Department of Science and Technology (DST), India. His research interests include seismic signal processing, machine learning, and seismic modeling and inversion.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/374259726

Speaker Independent Bangla Isolated Speech Recognition Using Deep Neural Network

Conference Paper · June 2023

CITATIONS
0

READS
19

3 authors, including:

Mehrab Hosain
Channel i
2 PUBLICATIONS **0** CITATIONS

SEE PROFILE

All content following this page was uploaded by Mehrab Hosain on 28 September 2023.

The user has requested enhancement of the downloaded file.

Speaker Independent Bangla Isolated Speech Recognition Using Deep Neural Network

Dipayan Bhadra [a], Mehrab Hossain [b], and Fatema Alam [c]

[a] Power Grid Company of Bangladesh Limited, email: dipu_dipayan@outlook.com
[b] Delhi Technological University, Delhi, India
[c] Graduate Student, Jahangirnagar University, Savar Dhaka, Bangladesh

Introduction: In this work, a dataset is created containing both Bangla and English words spoken by Bengali Speaker. An approach is proposed in this work to recognize isolated spoken words using the Mel Frequency Cepstral Coefficient (MFCC) and Deep Feed-Forward Fully Connected Neural Network (DFFNN) as a classifier. In this work the depth of the neural network is more than ever reported for Bangla speech recognition and the result of this work is showing better recognition accuracy compared to the most of the works done previously on Bangla.

M**aterials and methods: Dataset creation:** A dataset is created containing 36 Bangla words and 24 English words with the help of 25 different persons from different regions of Bangladesh. 30 samples per word from both male and female speakers are recorded using smartphones in room environment.

Table 1: Database size and distribution

Training Sample Size	Test Sample Size	Total Dataset Size
1200 Sample	600 Sample	1800 Sample
66.67%	33.33%	100%

Preprocessing: Down sampling, normalization, silence removal, windowing and pre-emphasis are done as preprocessing of speech signal in this work.

Down Sampling: The speech sample words are recorded with a high sampling frequency (44.10 kHz) to create this dataset. So the recorded speech signals are down-sampled at 10 kHz.

Normalization: Normalization is done by dividing the sample values of every speech signal with the maximum value present in that speech signal.Normalized signal, $\hat{x} = \frac{\text{signal, x}}{\text{maximum (x)}}$

Voice Activity Detection and Silence Removal: The voiced part of a speech signal is detected based on the energy level of that signal within a window of 108 millisecond. And then a threshold is used to determine whether it is a voiced portion or silence portion. After that, the silence portion is removed.

Windowing: In this work, Hamming window with the frame length of 300 and the 100 sample frame shift is used. 3/4 frame overlap is considered. So one frame is 30 ms and the frame shift is 10 ms.

Pre-emphasis: Pre-emphasis is done by a first-order high-pass filtering, difference equation in time domain is: Emphasis Signal (n) = Signal (n) − a ∗ Signal (n − 1); in this work '**a**' is set to 0.9375.

Feature Extraction: After pre-processing, compressed MFCCs are used as features. Here k-means clustering is used as feature compression technique.

MFCC: In this work 40 filters are used. The number of filter banks: 40, frame length: 300 and 300-100=200 is the frame overlap.The result is a N×14 matrix of MFCCs for each N frame.

Feature Compression: k-means classifier is used to cluster the vectors that contain the MFCCs to get compressed feature vectors. In this work, used 'K' value is 5 & 8 to analyze the performance. Final value of K is 8. The result is a 8×14 matrix, instead of the previous N by 14 matrix.

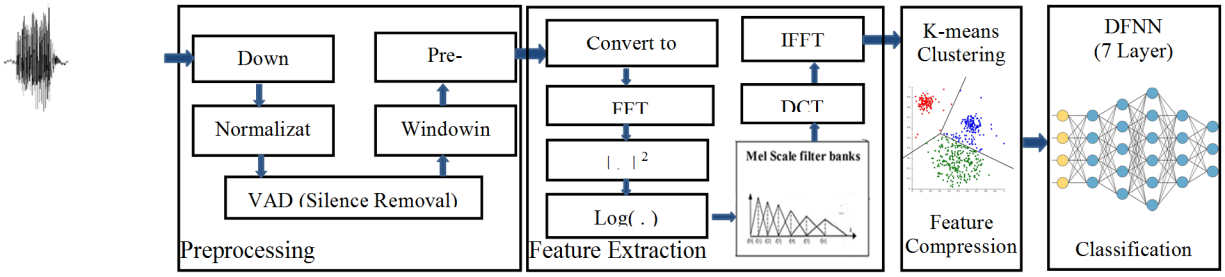


Figure 1: Block Diagram of proposed scheme

DFFNN for classification: DFFNN of 7 layers (1 input, 5 hidden and 1 output) is used as a classifier for this work. In the input layer, number of neurons is 112 (=8×14) and output layer has 60 neurons. Different number of neurons in hidden layers is investigated to get the best network architecture.

Results and Discussion: Parameters of the proposed system:

- 14 MFCCs as features (compressed into 1×112 vector with K-means classifier with K=8)
- Weight and Bias initialization method: Gaussian distribution random numbers
- Learning rate: 0.05, Decay factor: 0.95, Batch size: 1, Training algorithm: Backpropagation
- Activation function: Sigmoid, Cost function: Square Error (Euclidean distance)
- Neural Network architecture is: (112, 100, 95, 90, 95, 100, 60), total weights and biases: 53,840

Accuracy, Standard Deviation and Variance on test dataset:

Table 2: Accuracy, Standard Deviation and Variance on test dataset

Sample Information	Test set Accuracy (%)	Variance and Standard Deviation
No of Class=60 Test Sample =600	93.42%	Variance =0.02070764 Standard Deviation =0.1439

Comparative Analysis of the performance of DFFNN with different neuron number and Compressed MFCCs as input features

Table 3: Comparison with some reported work of Bangla speech recognition:

Sl. No.	Work Reference	Database information	Feature Extraction & Classifier	Reported Accuracy
1.	(Hossain, et al. 2013)	*10 classes, 10 speakers, 150 training samples, 150 test samples	*MFCC features *ANN with 1 hidden layer	**92%**
2.	(Noman & Cheng, 2022)	*7 classes (words) * 28 training samples * no other info about dataset	* absolute value of DFT *ANN with one hidden layer	**95.23%**
3.	(Sen, et al., 2022)	*100 classes, 100 samples for each class=10000 words *40,000 samples created	*MFCC spectrogram * CNN (4 Conv layer, 1 dense, 1 output layer)	**89.61%**
4.	Proposed Method	* 60 words, 30 samples/word **Total =** 1800 samples. (1200 samples for training and 600 samples for testing)	* MFCC *DFFNN (7 Layers)	**93.42%**

Table 4: Analysis of the performance of DFNN with different number of neurons

Hidden Layer and Features		No. of Neurons	Epoch	Learning Rate, ETA	Training Accuracy
5 hidden layer (MFCCs as feature)	K=5	85,90,85,90,90	50k	0.05	88.41%
	K=8	85,85,85,85,85	60k	0.06	91.75%
	K=8	100, 95, 90, 95, 100	100k	0.05	93.75%

We can see, the proposed system out-performs other prominent works on Bangla speech recognition considering database size, number of classes and recognition accuracy.

Conclusion: In this work, not only a moderate dataset has been created but also this work proposes a recognition model that gives better recognition accuracy compared to the most of the works done previously on Bangla considering the class number. 7 layer DFFNN along with MFCC as features compressed by K-means classifier with K=8 seem to provide the best output among all the trials.

Keywords: MFCC, Deep neural net, Bangla speech recognition, Speaker independent, Isolated speech

References:

Hossain, M. A., Rahman, M. M., Prodhan, U. K. & Khan, M. F. (2013). Implementation Of Back-Propagation Neural Network For Isolated Bangla Speech Recognition. *IJIST*, 3(4), 1-9.
https://arxiv.org/ftp/arxiv/papers/1308/1308.3785.pdf

Noman, A. & Cheng, X. (2022). Bengali Isolated Speech Recognition Using Artificial Neural Network. *Mechatronics and Automation Technology, 14-23*.

Sen, O., Roy, P. and Mahmud, A. (2022). A Novel Bangla Spoken Numerals Recognition System Using Convolutional Neural Network. *SSRN*.

View publication stats

Physica Scripta

INCLUSIVE PUBLISHING
TRUSTED SCIENCE

PAPER

Surface topography assessment using chemical assisted ball end magnetorheological finishing

To cite this article: Anand Sharma and Mahendra Singh Niranjan 2023 *Phys. Scr.* **98** 115961

View the article online for updates and enhancements.

You may also like

- Radio Submillimeter and -Ray Observations of the 2003 October 28 Solar Flare
G. Trottet, Säm Krucker, T. Lüthi et al.
- Computer simulation of liquid crystals
C M Care and D J Cleaver
- Investigation of the complex magnetic behavior of $Ni_{46.86}Co_{2.91}Mn_{38.17}Sn_{12.06}$ (at%) magnetic shape memory alloy at low temperatures
Ouz Yldrm, Suheyla Yuce, Nickolaus M Bruno et al.

This content was downloaded from IP address 14.139.251.98 on 03/11/2023 at 05:42

Physica Scripta

PAPER

Surface topography assessment using chemical assisted ball end magnetorheological finishing

Anand Sharma[1] **and Mahendra Singh Niranjan**[2]

[1] Research Scholar, Delhi Technological University, Delhi, India
[2] Associate Professor, Delhi Technological University, Delhi, India

E-mail: doc.asharma@yahoo.IN

RECEIVED
19 April 2023

REVISED
15 September 2023

ACCEPTED FOR PUBLICATION
10 October 2023

PUBLISHED
20 October 2023

Keywords: residual stress, chemical assisted ball end magnetorheological, aluminium, SEM, AFM, ANOVA

Abstract

The generation of residual stresses by grinding process on the surface for a soft material result in poor performance and diminished life. Ball end magnetorheological finishing (BEMRF) process is a recently developed process that is effectively used for fine figuring and polishing of a variety of magnetic and non-magnetic materials. In this paper, grinding-induced residual stresses are addressed, and an attempt is made to relieve these residual stresses while achieving high surface finish using chemical assisted ball end magnetorheological finishing (CA-BEMRF) process. The chemical used in the present work is applied first time in BEMRF process and aluminium 7075 alloy workpiece has not been finished with this novel finishing process. The impact of various CA-BEMRF variables on percentage surface roughness reduction and percentage reduction in residual stresses are discussed statistically and graphically. Residual stresses of workpiece surfaces have been measured by using portable x-ray residual stress analyzer. Scanning electron microscopy (SEM) and atomic force microscopy (AFM) of the finished workpiece before and after CA-BEMRF process is carried out to assess the effects of magnetizing current, tool rotation and working gap on surface topography of machined surface. It has been observed that reduction in residual stress is achieved along with high surface finish on aluminium workpiece surface using CA-BEMRF technique. Significant process parameters affecting the residual stresses and surface roughness during polishing of workpiece are obtained using Analysis of variance (ANOVA) and F-test. The optimization problem for maximum percentage reduction in surface roughness and maximum percentage reduction in residual stress is formulated as multi objective, multi variable, nonlinear optimization problem. Maximum percentage surface roughness reduction and minimum residual stress has been found at magnetizing current 2.3 A, rotational speed of tool 550 rpm and working gap of 0.5 mm. Confirmatory experimental tests have been conducted and results obtained were found very close to the predicted outcome.

1. Introduction

There is a significant demand for nano-meter and macro-meter items possessing excellent polished surface and strength in this age of new industrialization revolution. The surface finish has a direct impact on the optical quality, mechanical characteristics, and bio-medical functions of microparts [1, 2]. Engineering materials are utilized in a variety of applications to meet industrial objectives by leveraging qualities like conductivity, ductility, and lightweight [3]. However, some of these characteristics, such as malleability and weak scratch resistance, also provide challenges when polishing metals [4]. For the finishing of copper alloy at the nanometric level, numerous successful attempts have been made [5, 6]. Unfortunately, relatively few research has focused on using chemo-mechanical polishing (CMP) and magnetorheological finishing to achieve nanometric-level finishing on aluminium alloys (MRF). Optical mirror coatings have frequently used silver and gold. Today's high-quality mirrors, however, are coated with aluminium rather than silver to prevent tarnishing caused by air exposure, particularly sulphur, which results in the development of brown colored tarnish spots. These spots

© 2023 IOP Publishing Ltd

also result in a decrease in reflectance and a rise in emissivity [7]. Gold is a precious metal that is often used in jewelry and other decorative items. However, it is very costly as compared to other metals such as aluminium. Therefore, aluminium is often preferred over gold for various purposes. It is lighter in weight, cheaper, and more durable than gold. Additionally, it is easier to work with, making it a popular choice for many applications.

Humans' quest for solar and space exploration has led to a need for nanofinished aluminium alloys [8]. Telescope metal mirrors made of aluminium alloys are completed using the DTM process [9, 10]. These mirrors are thin, highly conductive, very resistant to corrosion, and have a longer fatigue life. Other significant uses of nanofinished aluminium surfaces include diaphragms for optical sensors [11], very corrosion-resistant marine transportation, microelectronic based devices [12], and mass level integration for conductive wires [13]. It has been discovered that finishing aluminium is a difficult process due to its malleability and flexibility. Nowadays, due to a few drawbacks including burrs, high residual stresses, sub-surface damages, etc, it is not viable to polish soft metals like aluminium up to the nanometric level using standard finishing methods like grinding, sprucing, burnishing, and lapping.

Residual stresses are stresses that remain in a material after it has undergone some process. These stresses can have harmful effects on the material, causing it to break or fail. Residual stresses can be caused by many factors, such as welding, casting, machining, or even just from the materials used in the construction of the item. The effects of residual stresses can be minimized by careful design and engineering of the item, as well as by using materials that are less likely to be affected by them.

To achieve high surface finish and minimum residual stress on aluminium alloys few finishing approaches which were able to control the finishing forces exerted on the workpiece surface are magnetic flow polishing, machining through abrasive flow, magnetorheological based abrasive flow and jet flow. In those approaches, the finishing forces are managed via way of means of controlling the magnetic flux density by means of the use of permanent magnet or an electromagnet. MRF process have been very successful in finishing of various materials such as glass, moulds made of steel, EN-31, polyamide etc but very few research has been done on aluminium alloys using this technique. However, these techniques can finish limited geometric shapes as concave, convex, planar, and symmetrical spheres.

Inspired by the favorable features of chemical mechanical polishing (CMP) process, various researchers have made an attempt to hybrid the CMP and MRF process to obtain high surface finish for different materials. Jain *et al* [14] developed chemo-mechanical MR finishing process by associating CMP and MRF process to enhance the surface finish of silicon blanks. It was concluded that operational gap, finishing period and rotational speed of magnets had significant effect and led to a surface finish of 4.8 Å.

Ranjan *et al* [5, 15] super finished a copper alloy using CMMRF technique and developed a mathematical model to study the polishing pressure and other controlling parameters. A delicate chemically passivated layer was created on copper alloy surface by chemically reacting it with ammonia, glycine and nitric acid. The results showed that optimization of pH value of the chemical is important to achieve better surface finish. Results showed that the depth and width of scratches on the surface were a few tens of nanometers size.

Ranjan *et al* [16] theoretically explored magnetism, polishing pad formation and polishing pressure using FEA-CFD simulation, the results of which were validated later experimentally on aluminium 6061-T6 alloy. The results revealed that the polishing pressure is affected by three parameters viz. Working gap, MR fluid and rotational speed.

Ghai *et al* [17] achieved a nano-level surface finish (0.597 nm) on Al 7075-T6 alloy using CMMRF technique. These results were obtained by optimizing the concentration of chemicals, abrasives, and CIP particles along with working gap. The authors used deionized water, glycine, ammonia, and citric acid to form a brittle passivated layer on workpiece surface and model the experiments with central composite design model technique. It was observed that aluminium alloy could only be finished in alkaline solution.

Kumar and Singh [18] critically assessed the benefits, futuristic opportunities, and the challenges in improving the capabilities of CMMRF process. They beautifully explained the mechanism and various process parameters affecting the capability of CMMRF process and its applications in the field of aerospace research, solar, medical implants, electronics and optical. It was also discussed that in future more focus can be given to in-depth finishing of ferrous and non-ferrous material, brittle materials, complex structures etc It was concluded that this technique can produce a surface finish on both brittle and ductile materials in the range of 2 to 6 Å.

Kumar and Singh [19] conduct experimental investigations and study the chemo-mechanical magneto-rheological finishing process of Al-6061 alloy. The authors specifically focus on the use of a composite magnetic abrasive called CIP-Al_2O_3, which is developed through the microwave-sintering route. By utilizing the composite magnetic abrasive, a damage-free Al surface with surface qualities of Ra 59 nm is obtained and the rate of material removal was steadily increased to 0.379 mg min^{-1}.

Although various researchers have carried out the finishing of materials like silicon, copper, aluminium alloy using 'CMMRF' process, but no work was found via BEMRF technique. In addition, according to the past work very limited work has been carried out using aluminium 7075 alloy. The present work is aimed to achieve a high

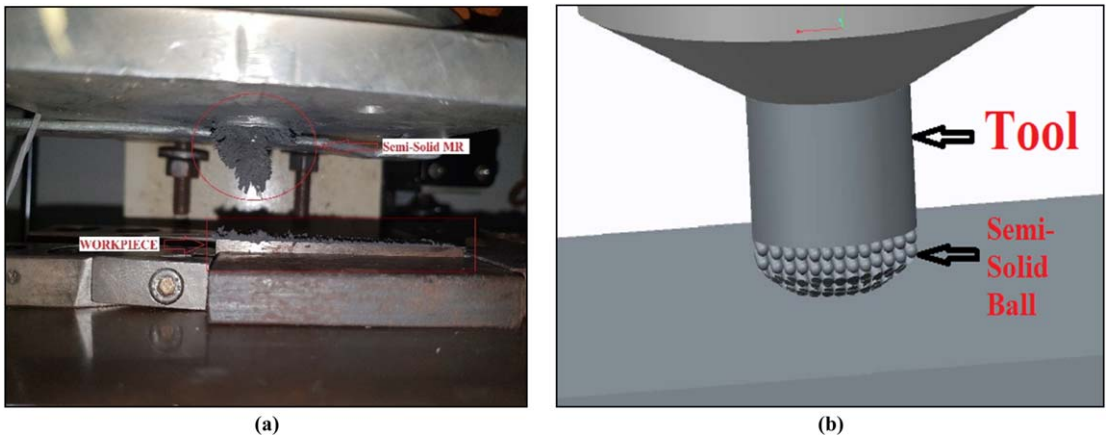


Figure 1. (a) Semi-solid MR fluid ball under influence of magnetic current, (b) schematic of semi-solid ball.

Table 1. Constituents of magnetorheological fluid.

S. no.	Constituents of MR fluid	Composition	W/W%
1	Base fluid	Paraffin oil (80 wt%) and grease (20 wt%)	55
2	Abrasive	Silicon Carbide (SiC)	25
3	Magnetic particles	Carbonyl iron particles	20

surface finish on aluminium 7075 alloy flat workpiece which can be directly used in different applications. The approach described in this study uses a newly created process called chemical assisted ball end magnetorheological finishing (CA-BEMRF), which combines the CMP and BEMRF procedures [20, 21], to finish the surface of an aluminium alloy workpiece. The chemical used in the present work is applied first time in BEMRF process and aluminium 7075 alloy workpiece has not been finished with this novel finishing process. Also, multi-optimization of residual stress and surface roughness achieved through CA-BEMRF process was not addressed earlier. To see how a chemical reaction affects the residual stresses and surface finish of an aluminium alloy workpiece, the percentage reduction in residual stress and percentage surface roughness reduction obtained before and after the CA-BEMRF process were compared.

The BEMRF process is a novel way of reducing residual stress in materials. By passing an electric current through the material, the BEMRF process can help to remove any internal stresses that may be present. This can be beneficial for a number of reasons, including improving the overall strength and durability of the material. Additionally, the BEMRF process can also help to improve the dimensional accuracy of the material, by reducing any warping or distortion that may be caused by residual stress.

2. Magnetorheological fluid

In CMP, MRF, CMMRF, and BEMRF processes, magnetorheological fluid (MR fluid) is frequently used to polish surfaces to an ultra-smooth finish. The preparation of the MR fluid is the most important factor in completing any surface in the CA-BEMRF. MR fluid enables the control of finishing forces applied on the workpiece surface while finishing through CA-BEMRF process. Table 1 describes the key constituents of MR fluid prepared for the experimentation. MR fluid consists of base fluid, abrasive particles (SiC) and carbonyl iron particles (CIPs) to form semi-solid ball under the effect of magnetic field as shown in figure 1(a). The schematic of the formation of ball shaped MR fluid is shown in figure 1(b). The MR fluid stiffens and attains the shape of a polishing ball under the action of the magnetic current before being compressed against the workpiece surface. Because the magnetic force increases as the working gap decreases, ultra-finishing can be accomplished by adjusting the distance between the tool tip and the workpiece. Controlling the tool's rotational speed, working gap, and magnetizing current the finishing forces can be controlled.

3. Mechanism of the proposed process

Finishing through CA-BEMRF process is accomplished through a mix of chemical action and mechanical erosion. Under the influence of a magnetizing current, magnetic particles create a semi-solid ball with abrasives, which causes mechanical material removal. A chemical agent then causes the production of a passivated layer, which is removed by the magnetic and abrasive particles.

3.1. Chemical reaction

Chemical reaction is essential for forming passivated layer to protect the aluminium alloy. A passive layer on the surface of aluminium alloy is formed when the workpiece is dipped in the chemical solution consisting of H_2O_2. Citric acid and phosphoric acid (H_3PO_4) were used as pH buffer solutions. In this process hydrogen peroxide (H_2O_2) was used as an oxidant for softening the aluminium alloy wokpiece surface which in turn increases the surface finish after mechanical action through BEMRF process. Kuo and Tsai (2001) [20] showed that a small concentration of H_2O_2 can substantially improve the material removal rate which in turn can quite effectively increase the surface finish of the workpiece. Equations (1)–(3) outline the mechanism's reaction pathway.

$$2AI + 3H_2O_2 \rightarrow AI_2O_3 + 3H_2O \rightarrow 2AI(OH)_3 \quad (1)$$

$$AI(OH)_3 + 3H_2O \rightarrow 2AI(OH_2)_3(OH)_3 \quad (2)$$

$$AI(OH_2)_3(OH)_3 + H_3PO_4 \rightarrow AIPO_4 + 6H_2O \quad (3)$$

This mechanism states that the surface of the Al is first oxidized by H_2O_2 to produce Al_2O_3. Al_2O_3 is next hydrated and reduced (dissolved) by H_3PO_4 to create a water-soluble phosphate salt (i.e., $AlPO_4$), which is then mechanically removed by CI and abrasive particles, concluding the removal cycle.

3.2. Mechanical action

Using an electromagnet, magnetic field is generated in CA-BEMRF process which shapes MR fluid into a semi-solid ball. Silicon carbide abrasive particles are held in chains of magnetic carbonyl iron particles when the magnetizing current is switched ON. The flexible finishing pad developed a micro-texture due to the average CIPs size of 1 to 5 μm.

Microtextures aid in retaining SiC abrasive particles and transferring finishing force to abrasive particles. By applying magnetic force to abrasive particles via CIPs, a passive layer that had formed on the alumnium workpiece surface was removed, and peak levels were reduced to a certain extent. CIPs hold the abrasive particles while also smoothing the workpiece's surface and removing any peaks. Tinier magnetic and soft abrasive particles are usually advised because hard abrasives and substantial magnetic particles will cause scratches. The final finishing force (FF) is a resultant of the normal force (N_F) and shear force (SH_F), based on the abrasive and magnetic particle microstructure. The shear force is responsible for material removal, whereas Fn is in duty of determining the depth of penetration of the workpiece surface.

4. Methodology

As a result of the restriction on the relative mobility of the workpiece and finishing tool, the limitations of finishing limited geometries were eliminated by developing ball end magnetorheological finishing (BEMRF), a variation of the MRF method. The original BEMRF configuration contained a MR finishing tool oriented vertically which is powered by a servo motor. The tool consists of a cylindrically formed central core (prepared of iron), an electromagnetic coil rapped around the tool & exterior core orientated concentrically to one another. The electromagnetic coil is intended to achieve around 0.8 Tesla flux density.

Figure 2 depicts the configuration of a 3-axis CNC BEMRF tool powered by a servo motor. The electromagnetic coil is enclosed by a tube made of copper through which cold water is pumped to keep the temperature acceptable for experiments. Once the current supply is turned on, the formation of a virtually very viscous MR fluid 'ball-shaped' at the tool tip takes place as shown in figure 1(a). The viscosity of MR fluid can be controlled or changed by real-time control of the magnetizing current, thereby controlling the intensity of the magnetic field.

4.1. Preparation of sample

Workpieces of Al 7075 alloy, each measuring 70 mm × 10 mm × 10 mm was cut and pre-finished using precision grinding technique. Table 2 shows the grinding process parameters used for pre-finishing the workpieces. The specimens were then washed with acetone and rinsed with DI water to eliminate any debris or burrs that remained after grinding. The Mitutoyo Surftest SJ-301 equipment is used to measure the first stage surface roughness acquired after grinding.

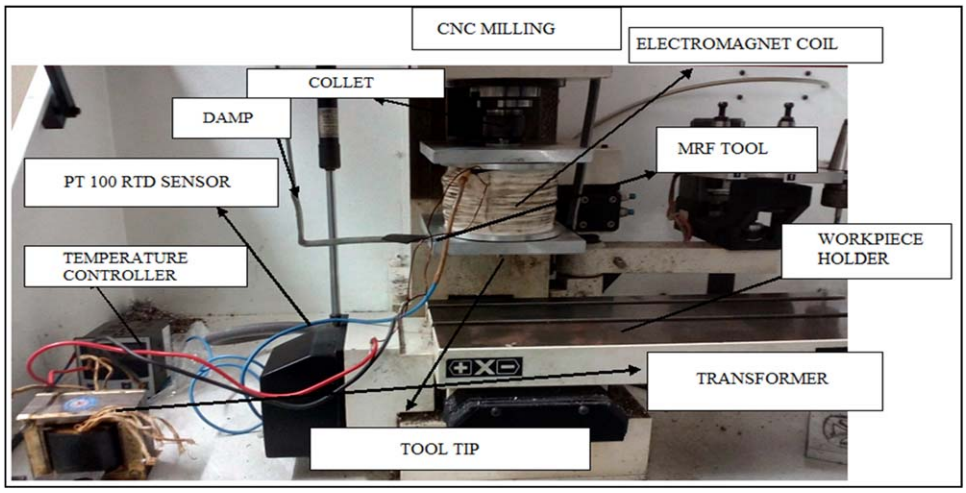


Figure 2. Setup of 3-axis BEMRF process

Table 2. Precision grinding process parameters.

Parameters	Values	Units
Wheel Velocity (m/min)	2026	m/min
Feed Rate-Work table Traverse (m/min)	0.12	m/min
Depth of cut	20	μm

Table 3. Preliminary experiments with CA-BEMRF process.

Exp no.	Chemical composition	Magnetizing current	Rotational speed of tool	Working gap	Initial *Ra* [nm]	Final *Ra* [nm]	$\%RSR_C$
1	3	2.5	500	1.5	440	278	36.82
2	4	2.5	500	1.5	442	280	36.65
3	**5**	**2.5**	**500**	**1.5**	**410**	**255**	**37.81**
4	6	2.5	500	1.5	445	280	37.10
5	7	2.5	500	1.5	435	282	35.17

4.2. Preliminary experiments

The preliminary experiments were performed on three-axis CNC BEMRF setup developed at Delhi Technological University, Delhi to obtain the ranges and significant process parameters. The effect of key parameters on percentage reduction in surface roughness and residual stress is investigated. The output response, i.e., percentage reduction in surface roughness ($\%RSR_C$) and residual stress (%RRS) after CA-BEMRF process is calculated by equation (4) and (5) respectively.

$$\%RSR_C = \frac{\text{SR after grinding process} - \text{SR after CA} - \text{BEMRF process}}{\text{SR after grinding process}} \times 100 \quad (4)$$

$$\%RRS = \frac{\text{Residual Stress after grinding process} - \text{Residual Stress after BEMRF process}}{\text{Residual Stress after grinding process}} \times 100 \quad (5)$$

Table 3 represents the preliminary experiments conducted on Al 7075 alloy workpiece surface using CA-BEMRF process. These experiments were conducted at 2.5 A magnetizing current, 500 rpm rotational speed of tool and 1.5 mm working gap with variation in chemical composition ranging from 3 to 7w/w% of MR fluid. The specimen was dipped in the chemical solution for 30 min prior to finishing through MR fluid.

It is observed from figure 3 that the highest $\%RSR_C$ is found to be 37.81% at 5% chemical composition, 2.5 A magnetizing current, 500 rpm rotational speed of tool and 1.5 mm working gap using CA-BEMRF process. The roughness profile of Al 7075 alloy workpiece surface after grinding and after CA-BEMRF process is shown in figures 4(a) and (b) respectively. The surface roughness value of 410 nm obtained after grinding the workpiece surface was reduced to 255 nm using CA-BEMRF process.

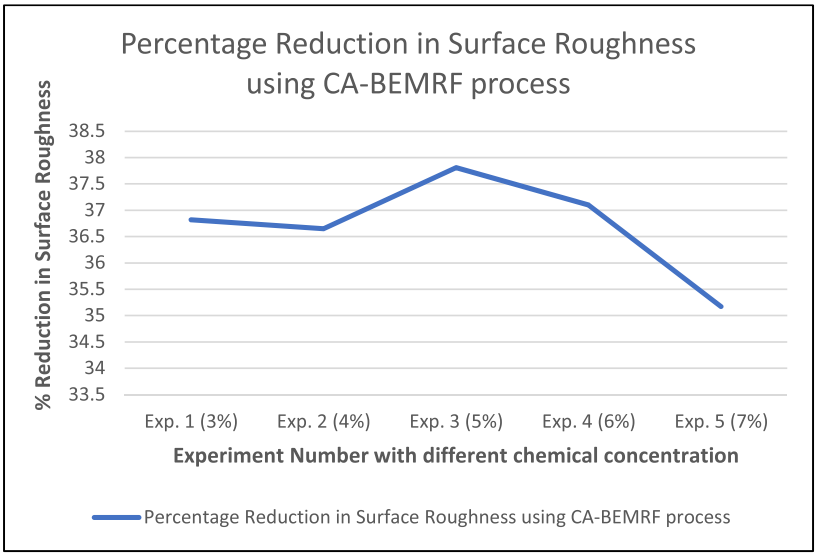


Figure 3. Variation of percentage reduction in surface roughness using CA-BEMRF process.

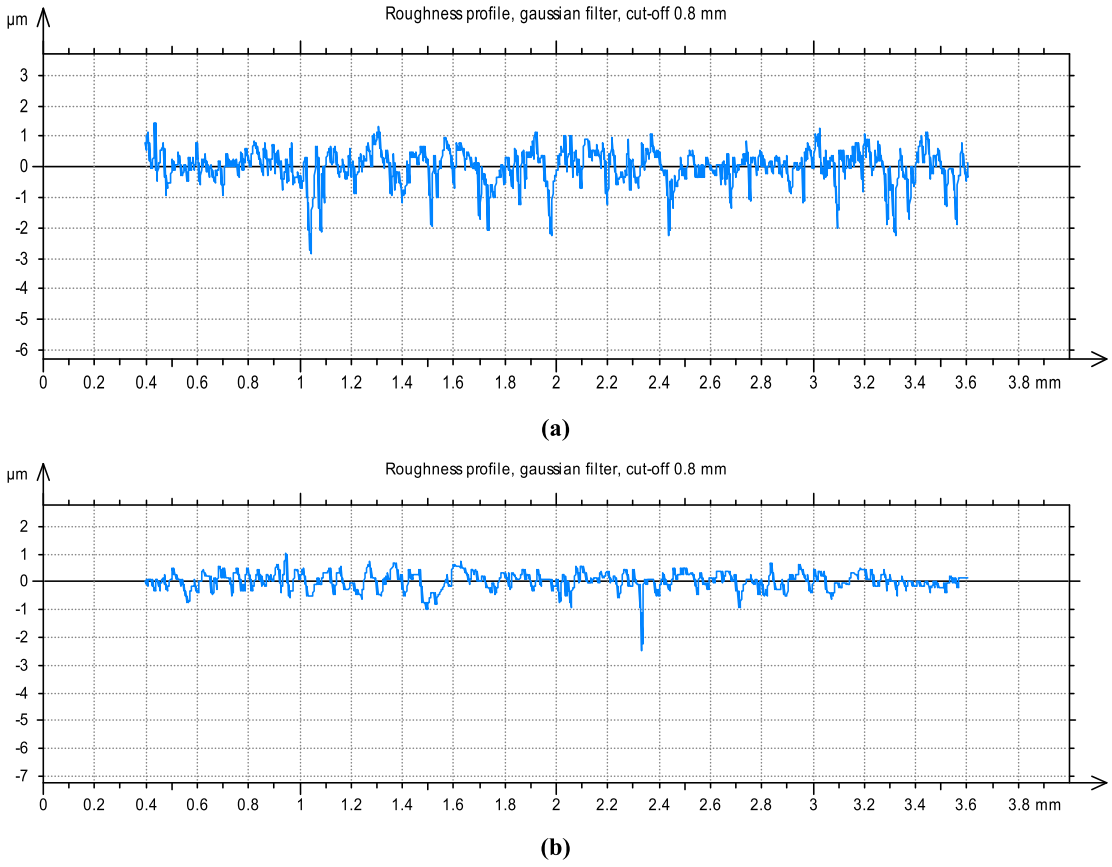


Figure 4. (a) Roughness profile of Al 7075 workpiece surface after grinding (b) Roughness profile of Al 7075 workpiece surface after CA-BEMRF process at 5% chemical composition, 2.5 A magnetizing current, 500 rpm rotational speed of tool and 1.5 mm working gap.

Comparing the percentage reduction in surface roughness obtained after grinding process and the conduction of preliminary experiments using CA-BEMRF process, it is observed that the surface finish has improved using the assistance of chemical agent. Hence, 5 w/w% chemical composition of hydrogen peroxide was fixed for the detailed experiments.

Table 4. Levels and ranges of selected process parameters for CA-BEMRF process.

Process parameter	Units	Levels I −2	II −1	III 0	IV 1	V 2	Static parameters	Description
Magnetizing Current (V_M)	A	1.5	2.0	2.5	3.0	3.5	Feed Rate	50 mm min^{-1}
Rotational Speed of Tool (T_R)	rpm	300	400	500	600	700	Finishing Time	30 min
Working gap (W_X)	mm	0.5	1.0	1.5	2.0	2.5	Chemical Composition	5 w/w%

Table 5. Percentage reduction in surface roughness and residual stress after CA-BEMRF process.

Std order	Run order	V_M	T_R	W_X	Initial Ra (nm)	Final Ra (nm)	$\%RSR_C$	Initial RS (MPa)	Final RS (MPa)	%RRS
6	1	1	−1	1	427	254	40.56	65	21	67.69
14	2	0	0	2	433	315	27.32	63	34	46.03
17	3	0	0	0	417	262	37.1	78	29	62.82
10	4	2	0	0	427	255	40.32	87	28	67.82
3	5	−1	1	−1	440	232	47.32	72	20	72.22
16	6	0	0	0	404	255	36.86	53	13	75.47
13	7	0	0	−2	442	118	57.32	46	10	78.26
8	8	1	1	1	439	311	29.12	85	55	35.29
1	9	−1	−1	−1	315	211	32.97	75	28	62.67
20	10	0	0	0	316	214	32.32	87	34	60.92
9	11	−2	0	0	433	315	27.32	52	28	46.15
7	12	−1	1	1	431	302	29.85	81	55	32.10
5	13	−1	−1	1	328	225	31.32	67	30	55.22
4	14	1	1	−1	425	238	43.97	72	21	70.83
15	15	0	0	0	429	267	37.81	80	28	65.00
18	16	0	0	0	415	260	37.32	75	28	62.67
12	17	0	2	0	435	306	29.72	71	46	35.21
11	18	0	−2	0	437	317	27.44	51	30	41.18
2	19	1	−1	−1	423	213	49.64	66	16	75.76
19	20	0	0	0	372	241	35.21	83	29	65.06

Legend: V_M = Magnetizing current, T_R = Rotational speed of tool, W_X = Working Gap, $\%RSR_C$ = Percentage reduction in surface roughness after CA-BEMRF and %RRS = Percentage reduction in residual stress after CA-BEMRF.

5. Result and discussion

In this work, the experimentation for finishing of aluminum 7075 alloy workpiece is designed according to central composite design (CCD) technique using design of expert 6.0.7 software. A total of 20 experiments designed according to central composite design (CCD) technique are conducted through CA-BEMRF process by varying the input variables viz. magnetizing current (V_M), Rotational Speed of Tool (T_R), Working gap (W_X) at five different levels as mentioned in table 4. Each specimen is submerged in chemical agent for a period of 30 min to obtain a passive layer in the surface before the mechanical abrasion through MR fluid. Table 5 represents the result of percentage reduction in surface roughness and residual stress after CA-BEMRF process on Al 7075 workpiece.

5.1. Surface roughness of Al 7075 after CA-BEMRF

A quadratic model was suggested by Design Expert 6.0.8 software for percentage reduction in surface roughness ($\%RSR_C$) after CA-BEMRF having F-value of 20.12 and the corresponding p-value which is less than 0.0001. After excluding the non-significant parameters, an empirical relationship (equation (6)) between the output response and the input process parameters after CA-BEMRF is acquired. With the prediction of the ideal values for the input process parameters, this equation may be used to achieve the maximum percentage reduction in surface roughness when CA-BEMRF of Al 7075 alloy.

$$\%RSR_C = 36.264 + 2.989*V_M + 0.021*T_R - 6.441*W_X - 1.548 *T_{R^2} + 1.887*W_{X^2} - 3.749*V_M*T_R 2.699*T_R*W_X \quad (6)$$

Figure 5(a) interaction graph between V_M and T_R on $\%RSR_C$ demonstrates unmistakably that when V_M is maintained at the greatest levels of 3.5 A, while T_R is kept at 500 rpm and W_X is maintained at 1.5 mm, $\%RSR_C$

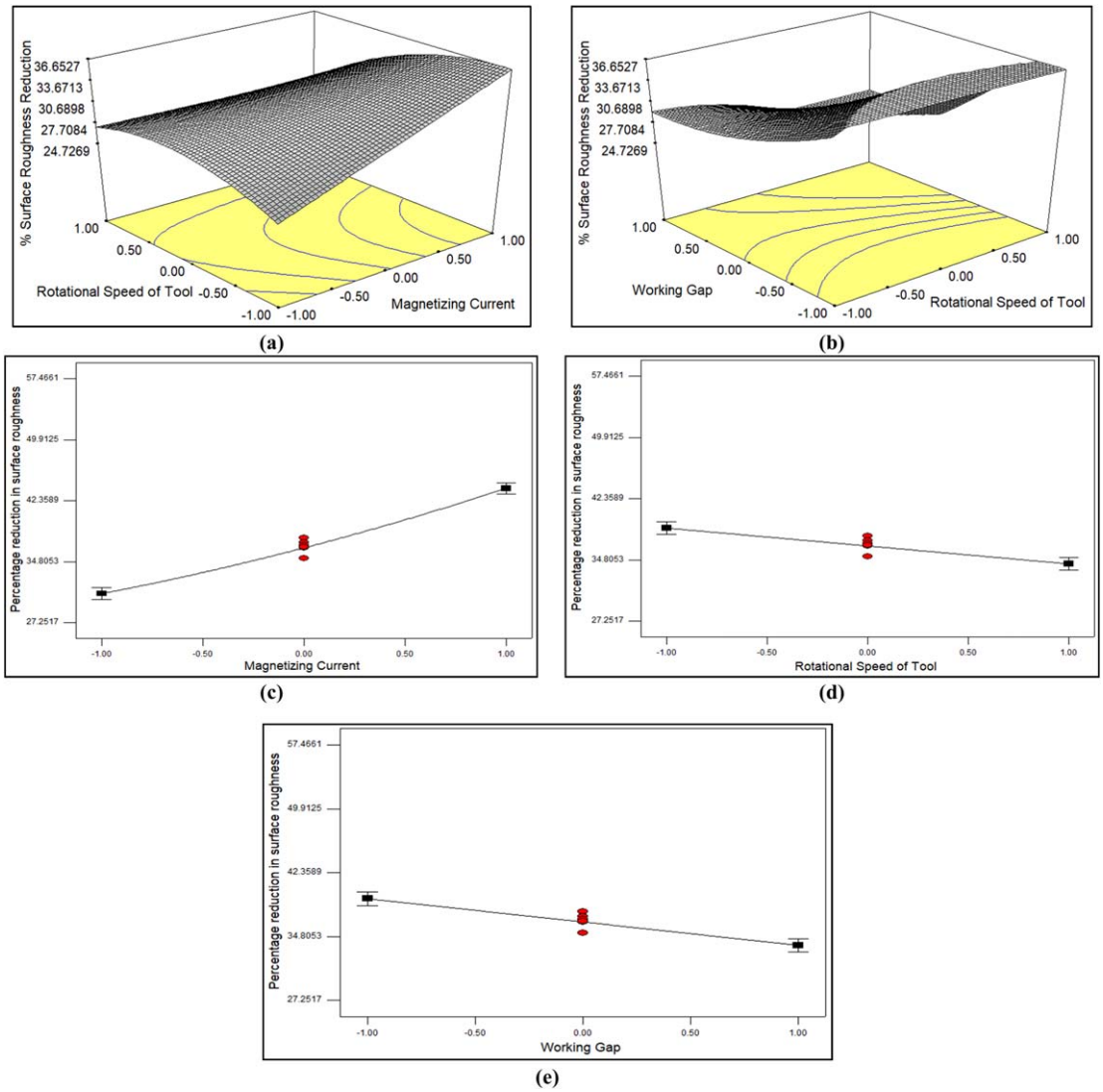


Figure 5. (a) V_M & T_R interaction graph on percent $\%RSR_C$, (b) Interaction graph of T_R & W_X on $\%RSR_C$ (c) Individual effect of V_M on $\%RSR_C$, (d) Individual effect of T_R on $\%RSR_C$ and (e) Individual effect of W_X on $\%RSR_C$.

achieves a maximum value of 45.83%. This is because hydrogen peroxide used as a chemical agent which softens the surface of the aluminium workpiece and high finishing force is exerted due to increased magnetizing current resulting in better surface finish.

The combined effect of T_R and W_X on $\%RSR_C$ as represented in figure 5(b) shows that $\%RSR_C$ value reduced from 44.15% to 34.4% on increasing T_R from 300 rpm to 700 rpm while keeping W_X at lowest level of 0.5 mm. It is also observed that $\%RSR_C$ value decreased to 33.3% from 44.15% once the working gap is increased to the highest value of 2.5 mm, keeping rotational speed of tool constant at 300 rpm.

The individual effects of chemical concentration, magnetizing current, tool rotation speed, and working gap on $\%RSR_C$ are each depicted in figures 5(c)–(e), respectively. The individual effect plot had the same slant as the interaction graph and demonstrated that an increase in V_M raised the $\%RSR_C$ while an increase in T_R and W_X decreased the $\%RSR_C$ for the previously mentioned reasons. The graphs show that the conditions of 5w/w% chemical composition, 3.5 A magnetizing current, 500 rpm rotational speed of the tool, and 1.5 mm working gap result in the highest percentage reduction in surface roughness value of 43.9%.

5.2. Residual stress of Al 7075 after CA-BEMRF

The results of the percentage reduction in residual stress of Al 7075 alloy workpiece after finishing through CA-BEMRF is shown in table 6. The quadratic model is significant, as evidenced by its F-value of 19.958 and matching p-value of less than 0.001.

After removing the non-significant process factors, an empirical relationship (equation (7)) is found between the output response and the input process parameters. By forecasting the ideal values of the input

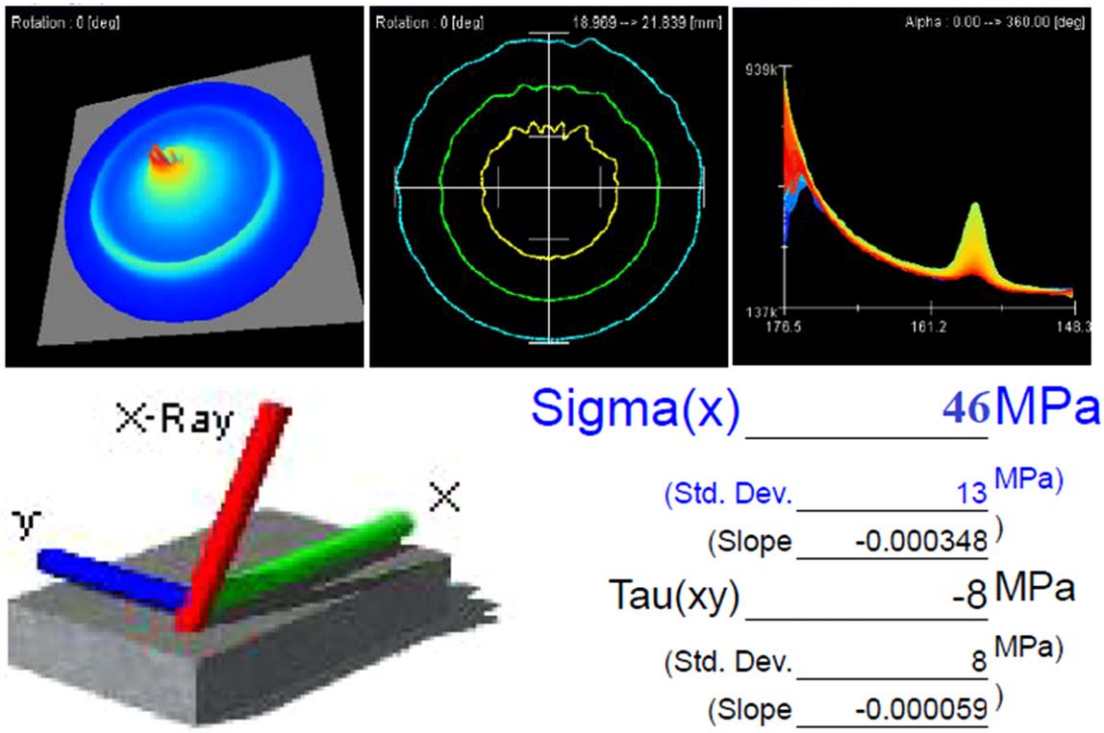


Figure 6. Residual stress on aluminium 7075 workpiece after grinding process.

Table 6. 10 best solutions for machining of Al 7075 using CA-BEMRF.

Serial no.	V_M	T_R	W_X	$\%SRR_C$	%RRS	Desirability	
1	2.3	550	0.5	58.728	80.542	1	Selected
2	3.3	510	0.6	57.732	88.457	1	Selected
3	3.4	450	0.6	59.589	83.124	1	
4	3.3	495	0.7	59.973	89.183	1	
5	3.4	470	0.8	60.717	87.568	1	
6	3.3	500	0.7	59.677	92.001	1	
7	2.6	540	0.5	59.028	87.765	1	
8	3.1	530	0.6	59.079	89.477	1	
9	3.4	400	2.5	55.138	67.531	0.844	
10	3.5	450	1.6	49.642	66.673	0.747	

process parameters, this equation can be used to achieve the maximum decrease in residual stress during BEMRF of Al7075 material.

$$\%RRS = 63.782 + 4.418*V_M - 3.926*T_R - 9.727*W_X - 6.079*C^2 - 7.519*T_R*W_X \quad (7)$$

After the initial grinding process a high level of residual stress of 46 MPa as induced in the workpiece surface as shown in figure 6. The residual stresses on the aluminium 7075 workpiece obtained after the CA-BEMRF process are shown in figure 7 while the tool's rotating speed, working gap, and magnetizing current are maintained at 2.5 mm, 500 rpm, and 0.5 mm, respectively. With the aforementioned parametric parameters, a residual stress reduction of up to 78.26% is achieved.

The combined effect of T_R and W_X on %RRS is shown in figure 8(a), and it demonstrates that when T_R is increased from 300 to 700 rpm while W_X is maintained at 2.5 mm, the value of %RRS decreases from 59.42% to 36.53%. Contrarily, when W_X was reduced from 2.5 to 0.5 mm and T_R remained constant at 700 rpm, the percentage of residual stress that was reduced dramatically increased to a high value of 71.02%. It was shown that the workpiece experiences shear and Hertzian stress as a result of vibrations with the abrasive particles when the tool's rotating speed increases while the working gap decreases.

The separate effects of V_M, T_R, and W_X on %RRS are depicted in figures 8(b)–(d). As seen in figure 9(b), an increase in magnetizing current causes a greater percentage reduction in residual stress. The results show that %

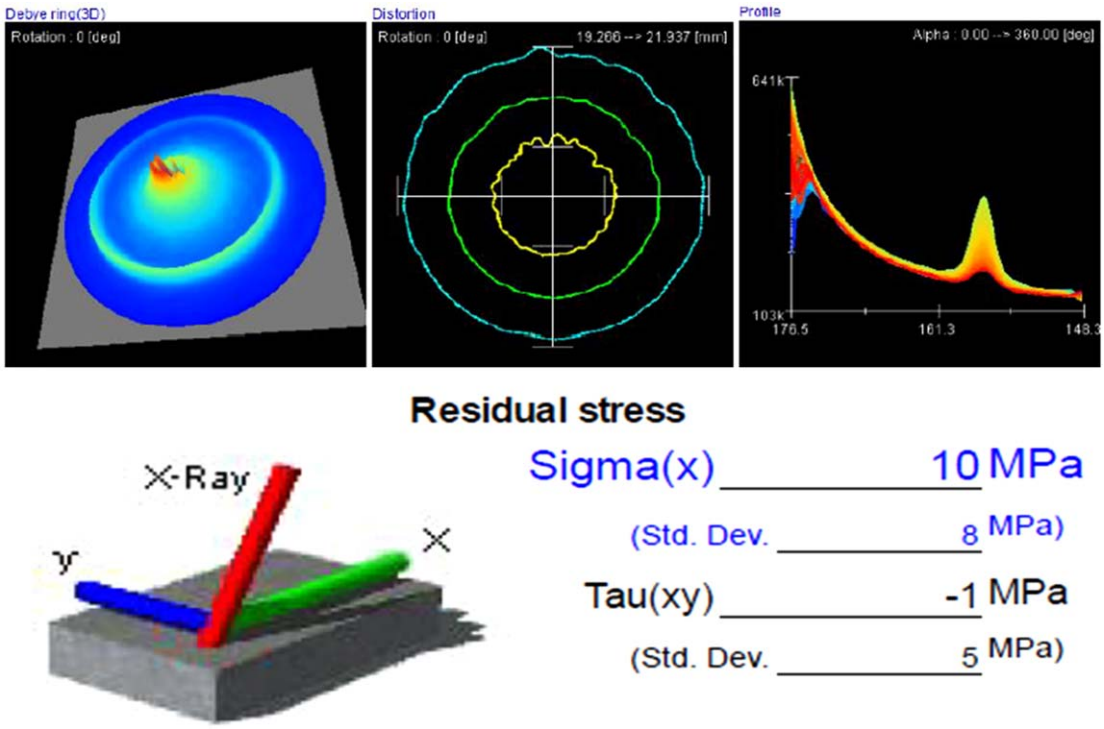


Figure 7. Residual stress on aluminium 7075 workpiece after CA-BEMRF process.

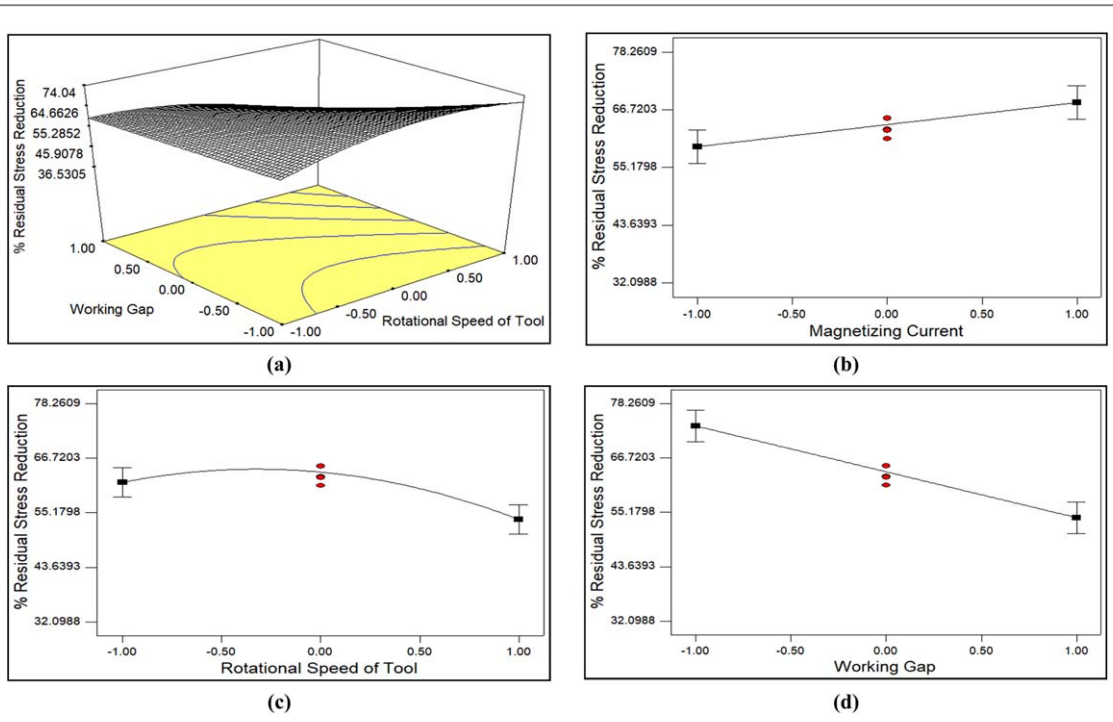


Figure 8. (a) Interaction graph of T_R and W_X on %RRS, (b) Individual effect of V_M on %RRS, (c) Individual effect of T_R on %RRS and (d) Individual effect of W_X on %RRS.

RRS increases from 59.36% to 68.2% on increasing V_M to 3.5 A from 0.5 A while keeping T_R and W_X constant at 500 rpm and 1.5 mm respectively. The individual effect plot, which had the same slant as the interaction graph of T_R and W_X, indicated that increasing T_R and W_X decreases the %RRS.

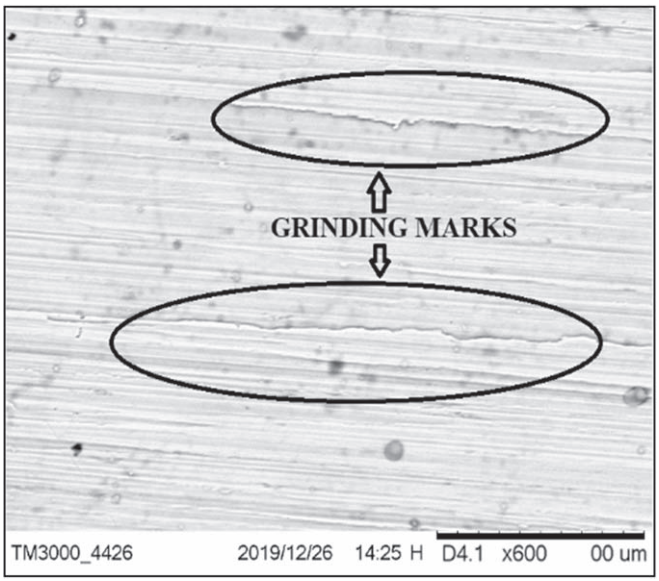


Figure 9. SEM of the workpiece after grinding process.

Table 7. Confirmatory experiments.

				Predicted		Actual		Error %	Error %
Serial no.	V_M	T_R	W_X	$\%RSR_C$	%RRS	$\%RSR_C$	%RRS	$\%RSR_C$	%RRS
				58.728	80.542	59.171	81.781	−0.75	−1.24
2	3.3	510	0.6	57.732	88.457	57.827	89.168	−0.16	−0.8

6. Output response optimization

A desirability function approach was presented by Derringer [22] and used to optimize the response to the output i.e. percentage surface roughness reduction and percentage reduction in residual stress after CA-BEMRF ($\%SRR_C$). This method assigns a dimensionless desirability value to the measured features of each expected response where d varies between 0 and 1. According to Montgomery [23], if the desirability function reaches a value of 0, it denotes an undesired response, and if it reaches a value of 1, the response is exactly what was intended. The value of d increases in direct proportion to how desirable the associated response is. The greater the scenario examined in the current work, the better the percentage reduction in surface roughness. The response is converted into a dimensionless function di using the equation (8) shown below.

$$d_i = \begin{cases} 0 Xi < Bi \\ (Xi - Bi/Hi{-}Bi)\mathrm{w} \quad Bi \leqslant Mi \leqslant Hi \\ 1 Bi > Hi \end{cases} \tag{8}$$

where Bi = Lower response limit, Mi = Upper response limit, Hi = Response goal value. B and M are selected in RSM according to the mathematical.

Determining the desirability response value and optimizing the desirability function to find the best potential response value are the first two steps in optimization.

Design Expert 6.0.8 software has been used to perform response optimization. Table 6 shows the 10 best solutions obtained using the desirability method. The highest desirability value obtained is 1, which means that when finishing Al 7075, the maximum percentage reduction in surface roughness and residual stress through CA-BEMRF process can be obtained at $V_M = 2.3$ A, $T_R = 550$ rpm and $W_X = 0.5$ mm.

A confirmation experiment was performed for solutions no. 1 and 2 to verify the above solutions. Table 7 summary of the findings from the confirmatory experiments shows that the error was less than 1.5%. These findings support the high repeatability of the beneficial optimization-derived solutions.

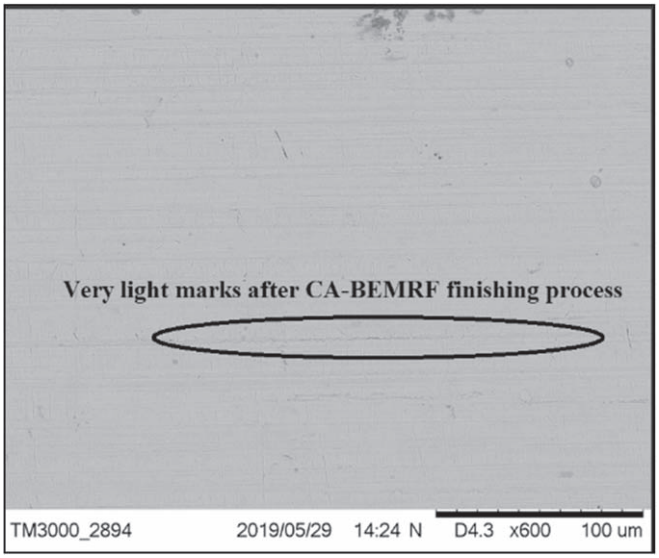


Figure 10. SEM of the workpiece after CA-BEMRF process.

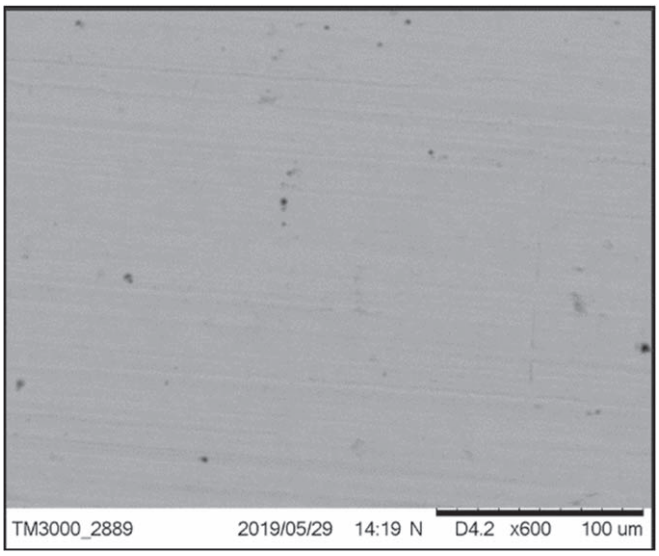


Figure 11. SEM of workpiece after confirmatory experiments.

7. Surface topography of machined surface

Scanning electron micrograph (SEM) of aluminium 7075 alloy workpiece surface after grinding process and CA-BEMRF process is shown in figures 9 and 10 respectively. The lays are visible clearly on the workpiece surface due to the pre-finishing technique i.e. grinding process as shown in figure 9. It can be observed from figure 10 that more uniform and improved surface finish with very fewer grinding marks is obtained after CA-BEMRF of the workpiece. When finished with the assistance of chemical agent hydrogen peroxide, percentage reduction in surface roughness increased to 57.32% leading to an improved surface finish as shown in figure 10. This finish was obtained at 5w/w% chemical composition, 2.5 A magnetizing current, 500 rpm rotational speed of tool and 0.5 mm working gap during CA-BEMRF process.

A 59.171% reduction in surface roughness and 81.871% reduction in residual stress was obtained after the confirmatory experiment conducted at 5 w/w% chemical composition, 2.3 A magnetizing current, 550 rpm rotational speed of tool and 0.5 mm working gap. SEM obtained for this specimen was found better as shown in figure 11 when compared with the SEM image obtained after grinding process shown in figure 9.

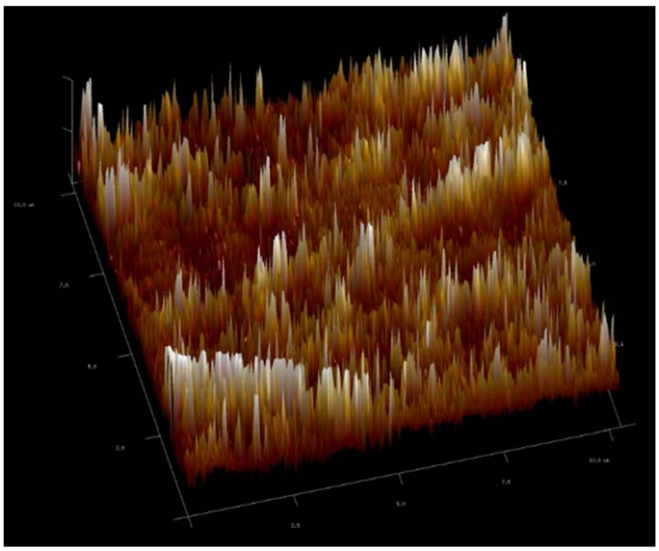


Figure12. AFM images of Al 7075 alloy workpiece surface after grinding process.

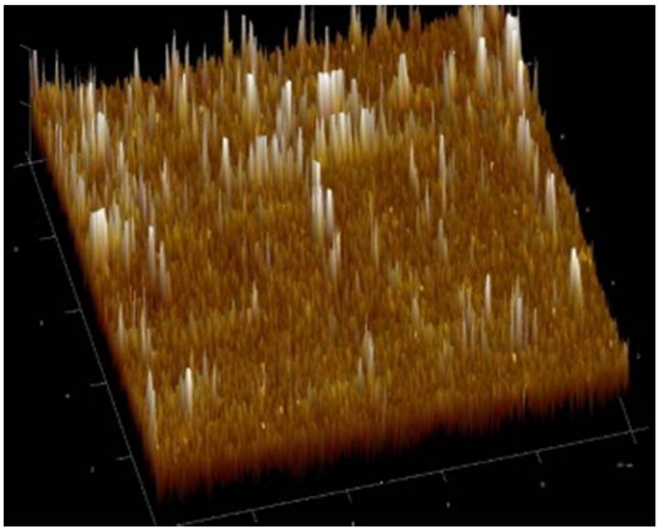


Figure13. AFM images of Al 7075 alloy workpiece surface after grinding process.

Atomic force microscopy (AFM) images of work-piece surface taken at the scale of 10 μm are shown in figures 12 and 13. From figure 12 more lays can be observed on aluminium work-piece surface after the initial grinding process. The density of lays on workpiece surface is 0.301 ($/\mu m^2$) and mean height of peak is 5.4541 (°). The surface texture produced after CA-BEMRF process at optimum value (5 w/w% chemical composition, 2.3 A magnetizing current, 550 rpm rotational speed of tool, and 0.5 mm working gap) has very fine lays as compared to the surface texture obtained after the initial grinded work-piece surface as shown in figure 13. The AFM images of work-piece surface finished through CA-BEMRF process has very fine lays with density of lays is 0.558 ($/\mu m^2$) and means height is 4.745 (°).

8. Conclusion

This study investigates how to reduce surface roughness and residual stress on the surface of an Al 7075 workpiece caused by grinding by employing the CA-BEMRF technique, which combines mechanical and chemical processes. The effects of input constraints, such as magnetizing current, tool rotation, and working gap, were studied on percentage surface roughness reduction and percentage reduction in residual stress. Additionally, the impact of chemical composition on the percentage reduction in surface roughness reduction

after the CA-BEMRF process (%SRR_C) was investigated. With design expert software, the optimization of the percentage reduction in surface roughness was carried out using the desirability technique. The following conclusions are drawn from this investigation.

- The experiments have been conducted on aluminium 7075 alloy workpiece with CA-BEMRF process.
- The CA-BEMRF process was able to provide a better surface finish and reduced residual stresses.
- The residual stress induced by the grinding process was significantly reduced by using CA-BEMRF process.
- Experimental results showed that the average percentage reduction in surface roughness of 38.19% was achieved when Al 7075 alloy workpiece was finished using CA-BEMRF process.
- The results of the confirmatory experiments conducted were found very close to the predicted outcome with a maximum of −1.24% error.
- The study also provides a set of optimized parameters for CA-BEMRF process to achieve high level surface finish and minimum residual stress for Al 7075 alloy workpiece.
- The study found that the chemical assisted BEMRF process produced a significantly smooth and more consistent surface finish on the aluminium alloy.
- SEM images of the finished workpieces showed an improved surface finish after CA-BEMRF process.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Statements and declarations

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Competing interests

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.

Availability of data and material

Not applicable.

Code availability

Not applicable.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

ORCID iDs

Anand Sharma https://orcid.org/0000-0003-4530-2531
Mahendra Singh Niranjan https://orcid.org/0000-0002-3062-8080

References

[1] Sharma A and Niranjan M S 2022 Experimental and theoretical investigation into surface roughness and residual stress using ball end magnetorheological finishing *J. Eng. Res.* (https://doi.org/10.36909/jer.14281)
[2] Jha S and Jain V K 2006 *Micromanufacturing and Nanotechnology* (Springer)
[3] Tov M S and Crabtree J A 2001 *Personal cooling apparatus and method.* US:6276155B2
[4] Chiu S Y, Wang Y L, Liu C P, Lan J K, Ay C, Feng M S and Dai B T 2003 The application of electrochemical metrologies for investigating chemical mechanical polishing of Al with a Ti barrier layer *Mater. Chem. Phys.* **82** 444–51
[5] Ranjan P, Balasubramaniam R and Suri V K 2013 Development of chemo-mechanical magnetorheological finishing process for super finishing of copper alloy *Int. J. Manuf. Technol. Manage.* **27** 130–41
[6] Babu S, Hegde S, Jha S, Patri U and Hong Y 2005 *Slurry and Method for Chemical-Mechanical Planarization of Copper.* US:20050026444 A1 Patent Application No. 10/846,718
[7] Boccas M, Vucina T, Araya C, Vera E and Ahhee C 2006 Protected-silver coatings for the 8-m Gemini telescope mirrors *Thin Solid Films* **502** 275–80
[8] Almanza R, Hernández P, Martínez I and Mazari M 2009 Development and mean life of aluminum first-surface mirrors for solar energy applications *Sol. Energy Mater. Sol. Cells* **93** 1647–51
[9] Casstevens J 2011 A 3 meter capacity diamond turning machine will be enabling technology for low cost fabrication of x-ray telescope optical components. NASA Request for Information Solicitation NNH11ZDA018L, NASA RFI
[10] Gebhardt A, Kinast J, Rohloff R R, Seifert W, Beier M, Scheiding S and Peschel T 2017 Athermal metal optics made of nickel plated AlSi40 In*International Conference on Space Optics—ICSO* **10563** 451–9
[11] Rathod A, Ghai V, Parmar P, Mishra S, Ghildiyal S, Ranjan P and Balasubramaniam R 2013 Development of intrinsic fabry–perot interferometic sensor based on a single mode—multi mode—single mode (SMS) configuration *8th International Conference on Precision, Meso. Micro and Nano Engineering, COPEN-8 (NIT, Calicut)* (https://doi.org/10.13140/RG.2.2.20602.08648)
[12] Ahn Y, Yoon J Y, Baek C W and Kim Y K 2004 Chemical mechanical polishing by colloidal silica-based slurry for micro-scratch reduction *Wear* **257** 785–9
[13] Shen M, Xiao F and Alameh K 2009 A novel reconfigurable optical interconnect architecture using an Opto-VLSI processor and a 4-f imaging system *Opt. Express* **17** 22680–8
[14] Jain V K, Ranjan P, Suri V K and Komanduri R 2010 Chemo-mechanical magneto-rheological finishing (CMMRF) of silicon for microelectronics applications *CIRP Ann.* **59** 323–8
[15] Ranjan P, Balasubramaniam R and Suri V K 2014 Modelling and simulation of chemo–mechanical magnetorheological finishing (CMMRF) process *International Journal of Precision Technology* **4** 230–46
[16] Ranjan P, Balasubramaniam R and Jain V K 2017 Analysis of magnetorheological fluid behavior in chemo-mechanical magnetorheological finishing (CMMRF) process *Precis. Eng.* **49** 122–35
[17] Ghai V, Ranjan P, Batish A and Singh H 2018 Atomic-level finishing of aluminum alloy by chemo-mechanical magneto-rheological finishing (CMMRF) for optical applications *J. Manuf. Processes* **32** 635–43
[18] Kumar Y and Singh H 2022 Chemomechanical magnetorheological finishing: Process mechanism, research trends, challenges and opportunities in surface finishing *Journal of Micromanufacturing* **5** 193–206
[19] Kumar Y and Singh H 2023 Experimental investigations on chemo-mechanical magneto-rheological finishing of Al-6061 alloy using composite magnetic abrasive (CIP-Al2O3) developed via microwave-sintering route *J. Manuf. Processes* **99** 765–80
[20] Kuo H S and Tsai W T 2001 Effects of alumina and hydrogen peroxide on the chemical-mechanical polishing of aluminum in phosphoric acid base slurry *Mater. Chem. Phys.* **69** 53–61
[21] Singh A K, Jha S and Pandey P M 2012 Magnetorheological ball end finishing process *Mater. Manuf. Processes* **27** 389–94
[22] Derringer G and Suich R 1980 Simultaneous optimization of several response variables *J. Qual. Technol.* **12** 214–9
[23] Montgomery D C 2001 *Design and Analysis of Experiments* (John Wiley & Sons Inc)

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 August 2023, accepted 25 September 2023, date of publication 4 October 2023, date of current version 13 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3321801

RESEARCH ARTICLE

SwinAnomaly: Real-Time Video Anomaly Detection Using Video Swin Transformer and SORT

ARPIT BAJGOTI[1], RISHIK GUPTA[1], PRASANALAKSHMI BALAJI[2], (Member, IEEE), RINKY DWIVEDI[1], MEENA SIWACH[3], AND DEEPAK GUPTA[4]

[1]Department of Computer Science and Engineering, Maharaja Surajmal Institute of Technology, New Delhi 110058, India
[2]Department of Computer Science, College of Computer Science, King Khalid University, Abha 61421, Saudi Arabia
[3]Department of Information Technology, Maharaja Surajmal Institute of Technology, New Delhi 110058, India
[4]Department of Computer Science and Engineering, Maharaja Agrasen Institute of Technology, Delhi 110086, India

Corresponding author: Prasanalakshmi Balaji (drsana@ieee.org)

The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through Small group Research Project under grant number RGP1/325/44.

ABSTRACT Detecting anomalous events in videos is a challenging task due to their infrequent and unpredictable nature in real-world scenarios. In this paper, we propose SwinAnomaly, a video anomaly detection approach based on a conditional GAN-based autoencoder with feature extractors based on Swin Transformers. Our approach encodes spatiotemporal features from a sequence of video frames using a 3D encoder and upsamples them to predict a future frame using a 2D decoder. We utilize patch-wise mean squared error and Simple Online and Real-time Tracking (SORT) for real-time anomaly detection and tracking. Our approach outperforms existing prediction-based video anomaly detection methods and offers flexibility in localizing anomalies through several parameters. Extensive testing shows that SwinAnomaly achieves state-of-the-art performance on public benchmarks, demonstrating the effectiveness of our approach for real-world video anomaly detection. Furthermore, our proposed approach has the potential to enhance public safety and security in various applications, including crowd surveillance, traffic monitoring, and industrial safety.

INDEX TERMS Video anomaly detection, SORT, GAN-based autoencoder, real-time tracking.

I. INTRODUCTION

Video anomaly detection is a technique used in computer vision to automatically detect unusual or anomalous events in video data. This can be useful for a wide range of applications, including surveillance, security, and traffic monitoring. However, this task has become increasingly challenging in recent times due to diversely distributed data. One method that has been employed to tackle this problem is one-class classification, where only normal videos are shown without any anomaly, and at testing the pre-trained model is required to distinguish between normal and anomalous events in complex video data.

Over the years, several methods such as reconstruction-based methods and prediction-based methods have achieved significant progress in anomaly detection. Reconstruction methods work on the principle that normal events can be reconstructed accurately whereas abnormal events will have a higher reconstruction error and hence can be detected as an anomaly. The prediction-based method on the other hand works on the principle that normal events can be accurately predicted, while abnormal events would not be correctly predicted.

Convolutional Neural Networks (CNNs) are necessary for prediction and reconstruction-based methods for their pattern recognition capabilities. However, they suffer from challenges such as the inability to find unexpected anomalies that deviate significantly from the training data since CNNs are designed to identify patterns in data. Another limitation of

The associate editor coordinating the review of this manuscript and approving it for publication was Laxmisha Rai.

© 2023 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

CNNs is the extensive computational power that they require which can be time-consuming for larger datasets such as video data.

Transformers in recent times have shown exceptional capabilities in Natural Language Processing (NLP) tasks such as language translation, text summarization, and sentiment analysis. Unlike CNNs which process input sequences sequentially, transformers process the entire sequence of input data in go. To achieve this, a mechanism called self-attention is used. Following the success of transformers in NLP, several advancements have been made in computer vision employing the use of transformers. Firstly, being introduced in Vision Transformer (ViT) [1] which takes 16×16 image patches for image classification. It was later adapted into Video Vision Transformer (ViViT) [2], which explored the application of ViT in video classification. The main idea behind using transformers in computer vision is to apply the self-attention mechanism to the spatial dimensions of the input image, which is achieved using a technique called spatial self-attention.

A. SWIN TRANSFORMER

Swin Transformer is a type of transformer architecture initially used for image classification tasks, first introduced by Microsoft Research Asia [3] in 2021. In the Swin Transformer, the input image is first divided into small patches, and then these patches are further divided into feature maps. This method allows the Swin Transformer to capture both global and local features of the image. The use of shifted windows enables the transformer to capture more fine-grained details and better handle small objects in the image. In the context of autoencoder-based methods for video anomaly detection, Swin Transformers offer several advantages such as better feature extraction as it uses self-attention to capture long-range dependencies in the input sequence, reduced computational cost, and improved reconstruction quality by reducing distortion in the reconstructed frames. The Swin Transformer supports upsampling and downsampling, unlike ViViT which makes it the optimal choice to be used as an encoder and decoder in a Generative Adversarial Network (GAN) for image generation tasks.

B. GENERATIVE ADVERSARIAL NETWORKS

Generative Adversarial Networks (GANs) have been successful in generating realistic images [4]. Recently, there has been a growing interest in combining parallel processing of transformers and realistic image generation of GANs for image segmentation tasks. In this approach, a generator network, which is a transformer-based architecture, takes an input image and produces a segmentation map. The discriminator network, which is typically a CNN-based architecture, takes both the input image and the generated image and tries to distinguish between real and fake images. Vision Transformer GAN (ViTGAN) is one such model that was proposed by

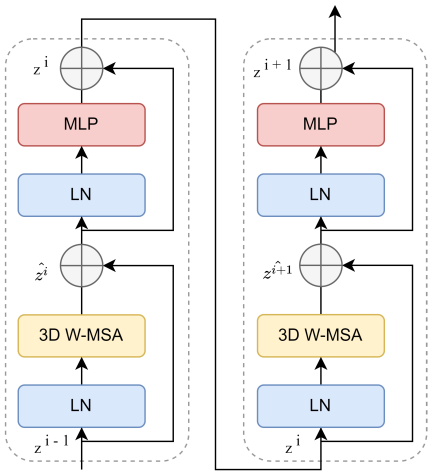


FIGURE 1. Two successive video swin transformer blocks.

Lee et al. [5] which uses a transformer-based generator and a CNN-based discriminator and has achieved state-of-the-art results. Similar to VitGAN, our model SwinAnomaly uses a swin transformer-based generator and discriminator which utilizes the benefits of using a swin transformer.

In recent years, several different architectures of GANs such as Deep Convolutional GAN (DCGANs) [6], [7] and PatchGAN [8] were introduced for image synthesis tasks. The CNN-based discriminator network type DCGAN processes the input image using many layers of convolutional and pooling procedures to produce a scalar probability score representing the realism of the image. DCGANs can have difficulty producing sharp edges and fine-grained features, but they are good at producing high-quality images with realistic textures and details. PatchGAN on the other hand concentrates on assessing the realism of local image patches. The input image is divided into several, non-overlapping patches using the PatchGAN, and each patch's level of realism is assessed independently. The discriminator can capture fine-grained information in the image and produce higher-quality photos with more realistic textures by employing this method. While PatchGANs can provide images that are clearer and more detailed than DCGANs, they may also take longer to train. The results of using both PatchGAN and DCGAN in the proposed approach have been discussed in the paper.

Although many machine learning-based anomaly detection methods have been introduced in recent years, they still suffer from some limitations. The current video anomaly detection methods struggle to generalize well to unseen or novel scenarios, performing well on the datasets they were trained on but fail to detect anomalies in new and diverse environments. Some methods, particularly those based on neural networks can be computationally intensive and require significant computational resources, making real-time deployment challenging.

In this paper, we propose SwinAnomaly, a novel architecture that aims to enhance anomaly detection in video datasets,

also allowing the model to be used in real-life scenarios. The encoder of the generator in the PatchGAN adopts the Video Swin Transformer architecture, while the decoder utilizes 2D Swin-transformer blocks to predict future frames from a sequence of video frames. Real-time anomaly detection is achieved by integrating an object detection model and the SORT algorithm, enabling the detection of anomalies using reconstruction error patches. Experimental results validate the superiority of SwinAnomaly over other prediction-based methods, highlighting its potential to enhance anomaly detection systems for more accurate and efficient real-world applications. The main contributions of the paper are as follows:

- We propose a model that combines Swin-transformers' patchwise feature extraction with the reconstruction capabilities of convolution-based autoencoders.
- SwinAnomaly operates on the current set of frames, without requiring a specific frame accumulation, allowing for the definition and continuous tracking of anomalies.
- We develop a technique that uses an object detection model such as YOLOv7 along with the SORT algorithm to identify and track anomalies in real-time.

II. RELATED WORKS

Several works have been done in the field of anomaly detection in the past decade. Generally, unsupervised video anomaly detection methods can be categorized into reconstruction-based methods and prediction-based methods.

A. RECONSTRUCTION-BASED METHODS

Reconstruction-based methods are a popular approach for video anomaly detection. The idea behind this method is to learn a model of normal behavior from a training set of videos, and then use this model to identify anomalies in new videos. It assumes that frames having anomalies will have higher reconstruction errors than frames without any anomalies. The most popular model used is the autoencoder, consisting of a convolutional encoder [11]. Other approaches by Fan et al. [12] and Li and Chang [13] used Variational Auto-Encoders (VAEs) to reconstruct the input frames, and the distribution difference of the latent representations was utilized to calculate the regularity scores. However, training VAEs was found to be difficult, particularly if the dataset was uneven or limited. Furthermore, the inability of VAEs to explicitly express temporal relationships may limit their ability to accurately capture the dynamics of a video sequence. To overcome this, Chong et al. [14] and Luo et al. [15] combined Convolutional Long Short-Term Memory (ConvLSTM) with Convolutional Auto-Encoder to extract more temporal information from input sequences. Incorporating the temporal information into the model helped improve its accuracy by allowing it to learn the dynamics of normal behavior over time. To find anomalies, certain reconstruction-based techniques also take advantage of the latent representational differences between normal and abnormal samples. However, using ConvLSTM had several challenges such as high computational cost, and overfitting, and it could not handle complex scenes with multiple objects and occlusions. In recent years, optical flow-based motion constraints have also been used in combination with Convolutional Auto-Encoder [16]. However using optical flow comes with its limitations such as high computational cost, sensitivity to noise, and lack of semantic understanding. Chang et al. [17] employed two autoencoders to individually leverage the spatial and temporal information of videos to get around the high computational cost of optical flow. While the temporal autoencoder recorded the movement data of the objects, the spatial autoencoder encoded the scenes and objects.

B. PREDICTION-BASED METHODS

Prediction-based methods for video anomaly detection involve training a model to predict future frames or events in a video sequence and detecting anomalies as deviations from this prediction. The basic assumption is that the normal event is predictable whereas the abnormal one is unpredictable [18]. These methods typically use Recurrent Neural Networks (RNNs) or Convolutional Neural Networks (CNNs) to learn the temporal dynamics of the video sequence and predict future frames or events. One common method is to predict upcoming frames in the video sequence based on previous frames using a sequence-to-sequence model, such as an encoder-decoder architecture with a Conv-LSTM [19] or Gated Recurrent Unit (GRU) [20] based RNN. The model can produce precise frame predictions under typical circumstances after being trained on a collection of typical video sequences. When there are departures from this prediction, such as abrupt changes in the projected frame or mistakes in the predicted motion, anomalies are recognized. In recent years, U-Net has been used as a generator for predicting the next frame, and a patch discriminator is used to distinguish the generated frames. Liu et al. [18] proposed a similar network architecture where U-net was used as a generator, and the model computed a regularity score with only the prediction error without reconstruction. However, using U-Net can be computationally expensive, they are not well suited for detecting anomalies that occur over longer time scales. In certain instances, reconstruction is combined with prediction models. For instance, Ye et al. [21] presented a Predictive Coding Network that, initially, predicts future frames using a ConvLSTM with predictive coding. Then, reconstruction techniques are used to reduce prediction errors. Lastly, to improve prediction performance, the predicted frames are updated with refined errors. This method still uses prediction error to get the regularity score while additionally considering reconstruction difference. In addition, Li et al. [22] replaced the original U-Net generator with a spatial-temporal U-Net that added three ConvLSTM layers between the U-Net layers to extract more temporal information. Tang et al. [23] on the other hand

followed a hybrid approach by combining both prediction and reconstruction-based methods by connecting two U-Net blocks, the first one for predicting the future frame and the second for reconstructing the future frame.

Although Convolutional-based methods work well, they require high computational power to generate reconstructed images. In our proposed approach, we use transformers to significantly reduce the processing cost of the individual frames which is achieved because of the self-attention mechanism and parallel processing capabilities of transformers.

C. OBJECT LEVEL DETECTION

The methods discussed above operate on entire video frames. This might prove to be difficult for anomaly detection as the video frames are complex, contain variations, and contain a large number of objects. More recent methods such as Object-level video anomaly detection aim to identify anomalies at the level of individual objects or regions within a video sequence [24], [25], [26]. They first extract object boundaries using object detection algorithms. Then the tracked objects are checked to be anomalous. This is an easier task since individual objects contain much less variation than the whole frame. The need for precise and effective object detection and tracking algorithms is one difficulty in object-level video anomaly detection. Occlusion, motion blur, and changes in lighting or background can all affect how accurate these algorithms are. In addition, these algorithms may be difficult to implement in real-time applications due to their computational cost.

D. TRANSFORMER BASED METHODS

Following the success of transformers in the field of NLP [27], many attempts have been made to replicate this success for computer vision tasks. Due to their capacity to represent long-term dependencies and capture intricate spatiotemporal correlations in video sequences, transformers have demonstrated promise in the field of video anomaly identification. Transformers are more effective at simulating temporal dynamics than more conventional approaches like ConvLSTM and can minimize processing time. Deshpande et al. [28] used a three-stage model consisting of a pre-trained videoswin model, an attention layer, and lastly an RFTM model for anomaly detection. TransUNet is a hybrid of U-Net with a transformer encoder, as suggested by Chen et al. [29]. TransUNet, a hybrid CNN-Transformer architecture, makes use of both the global context stored by the transformer encoder and the finely detailed, high-resolution spatial information from CNN features. Yuan et al. [30] proposed a similar architecture that combined U-Net and ViViT to capture richer temporal information and more global contexts. The ViViT was modified to make it capable of video prediction. Another strategy is to combine the usage of transformers with other machine learning methods, like object detection and tracking algorithms. This can entail projecting the future trajectory of certain objects inside a video sequence using a transformer-based model and then analyzing the behavior and motion of the items to spot anomalies.

While these methods have demonstrated impressive performance and yielded favorable outcomes across various datasets, their reliance on a substantial number of frames during inference poses certain limitations. Moreover, these approaches are primarily designed to detect local anomalies, meaning their ability to identify global anomalies is restricted. In cases where the chosen set of frames contains a considerable number of anomalies, the distinction between the minimum and maximum Peak signal-to-noise ratio (PSNR) becomes indiscernible, leading to a smooth anomaly curve. As a consequence, the effectiveness of these methods in accurately capturing and differentiating anomalies is diminished.

III. PROPOSED WORK

The overall architecture of the SwinAnomaly model is based on a Generative Adversarial Network where the generator is a Swin-UNET [31] based Autoencoder containing a 3D Encoder with Video Swin transformer layers and a decoder based on Swin-UNet decoder layers, The generator takes 3D input in the shape of (c, t, w, h) as shown in Fig. 4(a) and outputs a 2D predicted frame with dimensions (c, w, h), the generator tries to predict new frame $\hat{f}_{t+1}$ and the discriminator tries to differentiate between actual frames in the frame sequence and the predicted frame from the generator in an adversarial way, as shown in Fig. 2. The detailed overview of the proposed GAN constituents is shown in Fig. 3 and is discussed as follows:

A. FRAME ENCODER

The Encoder takes the input in the form $f_1, f_2, \ldots, f_T$ where each matrix f is a video frame of size $W \times H \times C$ and T is the number of frames taken together and concatenated to give a size of $T \times W \times H \times C$ to the input. The input is passed to a 3D patch partition layer which divides the input into $\frac{T}{t} \times \frac{W}{4} \times \frac{H}{4}$ patches where t is the number of frame partitions in the input data as shown in Fig. 4(b). The value of t is a small integer value where $kt = T$ as the frame size must be justified for input frames to be converted from a 3D input to a 2D patch where each patch contains the information content of the corresponding pixels of the frames in sequential order as shown in Fig.4(a). After the patch partition, a linear embedding layer is applied to project the patches into E-dimensional feature vectors, in the Swin-T-based architecture the embedding dimensions are taken as E = 96. The multi-head attention layer (MSA) is replaced with 3D windowed MSA and the consecutive block with cyclic shifted windowed MSA with shift space of (2, 2, 2) tokens as shown in Fig. 4(c), each encoder block's output is resized and propagated through an MLP layer to downsample the 3D input to 2D embedding which is passed to the adjacent decoder block as a skip connection to the decoder. Each encoder block is also down-sampled using a patch-merging

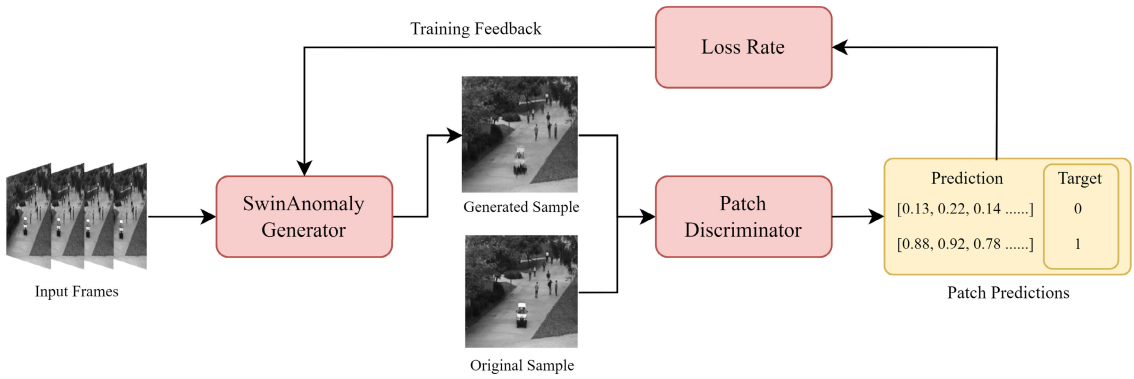


FIGURE 2. Conditional PatchGAN-based workflow of proposed approach.

layer, where the patches are reduced by two times i.e. $(T \times \frac{W}{2} \times \frac{H}{2} \times 2C)$, the consecutive blocks in each video swin transformer is shown in Fig. 1. It can be noticed that the channels and the image dimensions are altered to extract the features but the temporal frame patch size t is not reduced to maintain the integrity of our model during skip connections and upsampling [32].

1) 3D SHIFTED WINDOW BASED MSA

As the multi-head self-attention mechanism is used only within each non-overlapping 3D window, there are no connections between windows, which may reduce the architecture's capability for representation. As a result, the shifted 2D window mechanism of the Swin Transformer is converted to 3D windows to introduce cross-window connections, which is achieved by the circular shift of the patched frames as proposed by Liu et al. [32] while preserving the effective calculation of non-overlapping window-based self-attention.

$$\begin{aligned} \hat{z}^{l} &= 3DW - MSA(LN(z^{l-1})) + z^{l-1}, \\ z^{l} &= MLP(LN(\hat{z}^{l})) + \hat{z}^{l}, \\ \hat{z}^{l+1} &= 3DSW - MSA(LN(z^{l})) + z^{l}, \\ z^{l+1} &= MLP(LN(\hat{z}^{l+1})) + \hat{z}^{l+1} \end{aligned} \tag{1}$$

where $\hat{z}^{l}$ and z^{l} represent the outputs of the $(S)W - MSA$ module and the MLP module of the l_{th} block, respectively.

2) 3D RELATIVE POSITION BIAS

The addition of a relative position bias B significantly improves the performance of swin-transformers in feature extraction. For incorporating 3D patches, a new dimension P is added which represents patches in the 3rd dimension. The Attention mechanism of the attention layer in Swin block 3D is governed by the following equation:

$$Attention(Q, K, V) = SoftMax\left(\frac{QK^{T}}{\sqrt{d}} + B\right)V \tag{2}$$

where the query, key, and value matrices are $Q, K, V \in \mathbb{R}^{PM^{2} \times d}$, d is the dimension of the query and key features, and PM^{2} is the total number of tokens in a 3D window. A smaller-sized bias matrix, $\hat{B} \in \mathbb{R}^{(2P-1)\times(2M-1)\times(2M-1)}$ is parameterized, and values in B are picked from $\hat{B}$ because the relative position along each axis falls within the range of $[-P+1, P-1]$ (temporal) or $[-M+1, M-1]$ (height or width).

3) PATCH MERGING LAYER

The patch merging layer separates the input patches into 4 parts, which are then concatenated. The feature resolution will be two times down-sampled with this processing. Additionally, because the concatenate operation causes the feature dimension to increase by $4\times$, to reduce the complexity of the model, a linear layer is applied to the concatenated features to bring the feature dimension back to the original $2\times$ dimension.

B. FRAME DECODER

The symmetric decoder in Swin Transformer is constructed based on Swin Transformer blocks, similar to the encoder. However, the decoder applies attention to 2D patches instead of 3D. Unlike the patch merging layer used in the encoder, the decoder utilizes a patch-expanding layer to up-sample the extracted deep features. This layer reshapes the adjacent dimension feature maps into a higher resolution feature map with $2\times$ up-sampling while reducing the feature dimension to half of the original dimension.

1) PATCH EXPANDING LAYER

Let's take the first patch expanding layer as an example. Before up-sampling, a linear layer is applied to the input features $(\frac{W}{32} \times \frac{H}{32} \times 8C)$ to increase the feature dimension to 2x the original dimension $(\frac{W}{32} \times \frac{H}{32} \times 16C)$. Then, a rearrange operation expands the input feature resolution to 2x the original resolution and reduces the feature dimension to a quarter of the input dimension $(\frac{W}{32} \times \frac{H}{32} \times 16C \rightarrow \frac{W}{16} \times \frac{H}{16} \times 4C)$.

2) 2D SHIFTED WINDOW MSA AND POSITIONAL BIAS

The decoder uses 2D patches unlike encoders hence the equations 1 and 2 can be easily modified to sample

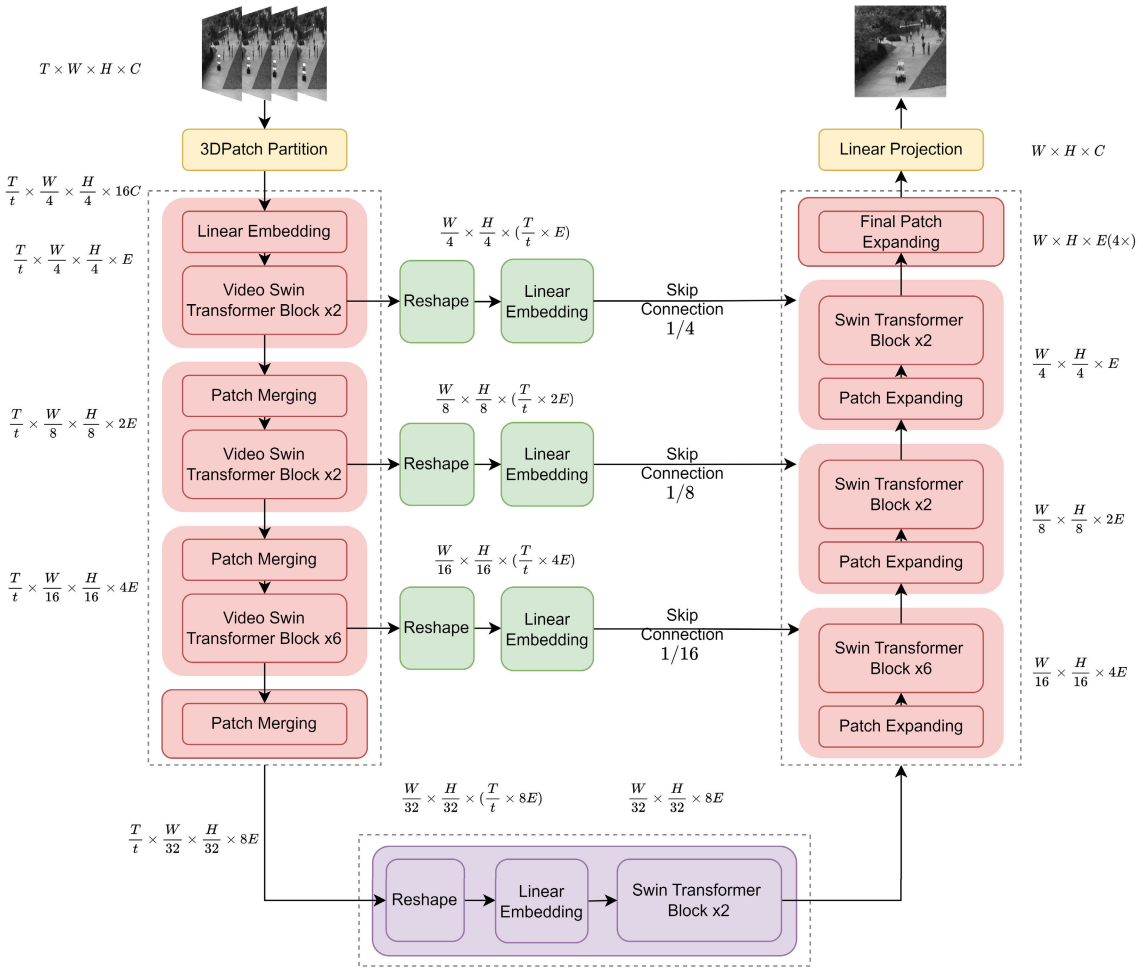


FIGURE 3. Architecture of proposed SwinAnomaly generator.

down the attention mechanism to 2D matrices, in the swin transformer(2D) block the P dimensions are discarded in eq 2, and the MSA is changed from 3D to 2D.

C. BOTTLENECK AND SKIP CONNECTIONS

The bottleneck connects the 3D encoder and 2D decoder models together, and the encoder's final layer inputs a feature space with dimensions $\frac{T}{t} \times \frac{W}{32} \times \frac{H}{32} \times 8E$ which is a $3\times$ down-sampled feature vector of the original frame sequence. To convert the features from 3D to 2D, a linear embedding on the rearranged encoder output is applied, this conscience to the invariable temporal channels which only change during the attention mechanism in the shifted-window MSA of Video Swin Transformer and doesn't affect the frame reconstruction.

The layers in the encoder part are skip-connected and concatenated with corresponding layers in the decoder part, this helps the decoder retain the information of the encoder spatial features which gets diminished during down-sampling and helps in the image or frame reconstruction. The skip connections take the encoder layer outputs before patch merging and concatenate them to their corresponding decoder swin transformer blocks by rearranging and reshaping them similar to the bottleneck block.

D. ANOMALY TRACKING USING SORT

Our proposed approach for anomaly detection and evaluation involves several steps. First, we calculate the patch-wise mean squared error on non-overlapping patches of the predicted frame and ground truth frame. This is done using eq. 3:

$$MSE_{i,j} = \frac{1}{n} \sum_{x=i}^{i+p-1} \sum_{y=j}^{j+p-1} (f_{x,y} - \hat{f}_{x,y})^2 \tag{3}$$

where $MSE_{i,j}$ represents the mean squared error for the patch at position (i, j), n is the number of pixels in the patch, p is the patch size, f is the ground truth frame, and $\hat{f}$ is the predicted frame. Next, we select the k maximum MSE patches based on the number of anomalies to capture. These patches represent the regions in the frame that have the highest anomaly scores. Then, we use the SORT algorithm to track the anomalies, as given by the equation:

$$\text{SORT}(B_i) = \hat{B}_i \tag{4}$$

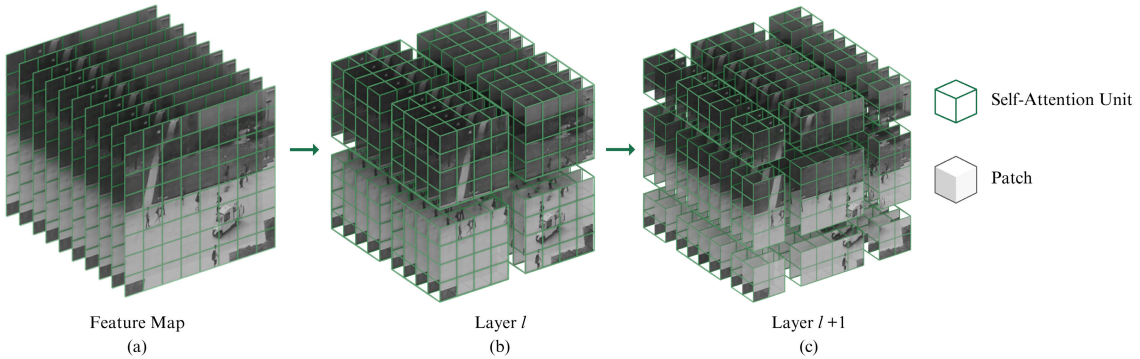


FIGURE 4. A 3D shifted window used in the generator. (a) The input of the generator with T consecutive frames (b) Patch partition of the input into tokens of shape $\frac{T}{t} \times \frac{W}{4} \times \frac{H}{4}$ (c) A cyclically shifted window by (2, 2, 2) tokens for extensive feature extraction.

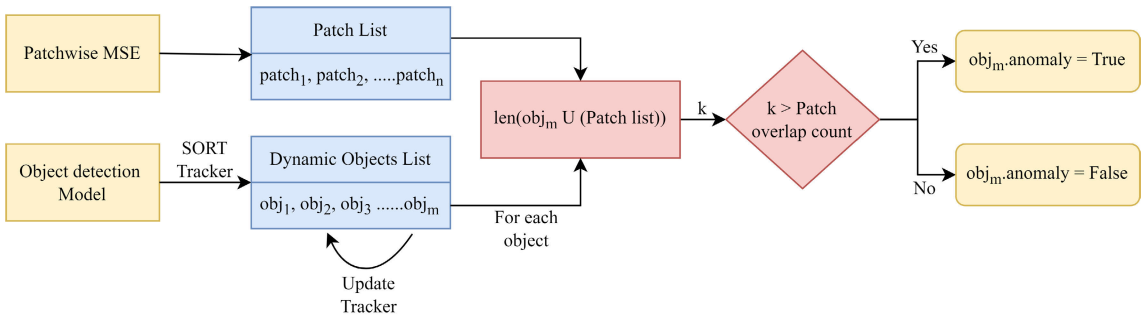


FIGURE 5. Inference workflow using patchwise MSE and SORT tracking.

where B_i represents the bounding boxes at frame f, and $\hat{B}_i$ represents the predicted bounding boxes at frame f using the SORT algorithm. The SORT algorithm helps track the movement of the anomaly regions across frames. After that, we use an object detection model such as YOLOv7 to capture the different bounding boxes in the original image, as given by the equation:

$$BB = \text{YOLOv7}(f) \tag{5}$$

where BB represents the bounding boxes detected in the original frame f. The object detection model helps detect other objects in the scene that may be relevant to the anomaly. We then find the IOU score on the rectangular bounding boxes and the max k MSE patches using the following equation:

$$\text{IOU}(BB_i, \hat{B}_j) = s_{ij} \tag{6}$$

where s_{ij} represents the IOU score between the bounding box at frame i and the predicted bounding box at frame j using the k max MSE patches. The IOU score helps us determine how well the predicted bounding boxes overlap with the actual bounding boxes. To filter out the correct anomaly paths, we take a threshold u which measures the number of consecutive times the same object was a subject of anomaly. If it exceeds the frame count, it is considered an anomaly, as given by the equation:

$$T_{i,j} = \begin{cases} 1 & \text{if } \sum_{k=i}^{i+u-1} s_{jk} > \tau \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where $T_{i,j}$ represents whether the anomaly is present in frame i and predicted in frame j, and τ represents the threshold for the sum of the IOU scores over u frames. Finally, we track the anomalies in the CCTV footage until they are out of the camera's scope using the equation:

$$\text{Track}(T_{i,j}) = A_{i,j} \tag{8}$$

where $A_{i,j}$ represents the tracked anomaly at frame i and predicted in frame j. The tracking allows us to follow the anomaly and take necessary action.

The workflow shown in Fig. 5 enables the detection and tracing of anomalies, providing flexibility regarding the type, size, and persistence of the abnormal object or activity. Unlike the sliding window PSNR approach, this method can be used in real-time analysis of surveillance cameras, and can also provide time stamps for suspicious or abnormal activities during very long video sequences in surveillance.

IV. EXPERIMENTS

In this section, the proposed generator is evaluated on the UCSD pedestrian, CUHK Avenue, and ShanghaiTech

datasets, and the effect of different discriminators and regularity scores are explored.

A. DATASETS

The CUHK Avenue dataset consists of 16 training video clips and 21 testing video clips, all captured at the CUHK campus avenue. The training videos exclusively feature regular situations, while the testing videos cover anomalous situations, including strange actions, abnormal objects, and wrong directions. The UCSD Pedestrian dataset is divided into two subsets: Ped1 and Ped2. Ped1 comprises 34 training video clips and 36 testing video clips, while Ped2 includes 16 training video clips and 12 testing video clips. The training videos in both subsets represent normal scenes, while the testing videos feature abnormal targets such as bikes, cars, and skateboards. The ShanghaiTech dataset contains a more diverse collection with 330 training videos and 107 test videos, capturing 13 different scenes with various camera angles. This dataset presents a greater challenge during inference, as it contains 130 abnormal events.

B. TRAINING

GANs refer to generative models that acquire knowledge of a mapping from a random noise vector z to output image y, which can be represented as $G : z \rightarrow y$. On the other hand, conditional GANs discover a mapping from a random noise vector z and an observed image x to produce an output image y, which can be described as $G : x, z \rightarrow y$. The generator G is designed to generate outputs that are indistinguishable from authentic images by an adversarially trained discriminator D, which is trained to identify the generator's ''fakes'' with maximum accuracy. The standard loss function for GAN is a min-max loss defined by the equation:

$$\mathcal{L}(G, D) = \mathbb{E}_{x,y}\left[log\left(D\left(x, y\right)\right)\right] + \mathbb{E}_{x,z}\left[log\left(1 - D\left(G\left(x, z\right)\right)\right)\right] \quad (9)$$

In the proposed approach, DCGAN [6] and PatchGAN [8] discriminators are used for adversarial training.

C. DCGAN LOSS AND BINARY DISCRIMINATOR

The DCGAN discriminator has a binary classifier-based architecture whose losses are defined by the equation:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[logD\left(x^{(i)}\right) + log\left(1 - D\left(G\left(z^{(i)}\right)\right)\right)\right] \quad (10)$$

The predicted image of this variation of discriminator after training for 38 epochs in the Ped1 dataset provides satisfactory results as the abnormal features or pixels from the frame are not generated, but such a predicted frame cannot be used during anomaly detection due to high noise on the static background as well as the unstructured figures of the normal scenarios, hence there is no metric that can generalize the results accounting to the high noise in the predictions, also further training the model doesn't improve the predictions but tries to overfit the data by copying the features of the corresponding frame. The result of the 38th epoch is shown in Fig 7.

D. PATCHGAN LOSS AND PATCH DISCRIMINATOR

Accounting for the limitations of the DCGAN loss, the discriminator is replaced by the PatchGAN discriminator which tries to classify each $N \times N$ in an image as real or fake. For example, if the input frame shape is $224 \times 224 \times 3$ and the linear embedding dimension $k = 4$ then the output shape of the discriminator is of shape 28×28 as shown in Fig. 6. In addition to the conditional GAN loss, the L1 loss is also used to retain the sharpness of the image as shown:

$$\mathcal{L}_{L1} = \mathbb{E}_{x,y,z}\left[\|y - G(x, z)\|_1\right] \quad (11)$$

The overall loss is the combined effect of both the losses from equations 9 and 11 as:

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L1}(G) \quad (12)$$

Fig 6 shows that the full convolution discriminator in [30] is replaced by the 2D swin transformer blocks and patch merging layer for feature extraction and down-sampling which takes $224 \times 224 \times 3$ frame as input and outputs a 28×28 patch as a patch for further prediction. The result of the 45th epoch is shown in Figure 8, which shows that the predicted frame has better sharpness and reconstruction of the static pixels and a blurry prediction for the anomalous pixels.

E. INFERENCE

The output of the trained SwinAnomaly model is further processed for anomaly detection, where the output generated from the Generator of the GAN is considered for a temporal sequence of 4 frames per input, and the blurred pixels from the predicted frame are taken as input which is passed through the inference pipeline as shown in Fig. 10. For inference, different anomaly detection strategies are used, but the common image difference metric used is the PSNR which is defined as follows:

$$PSNR(f, \hat{f}) = 10 \log_{10} \frac{[\max_{\hat{f}}]^2}{\frac{1}{N}\left\|f - \hat{f}\right\|_2^2} \quad (13)$$

1) SLIDING WINDOW BASED PSNR

The initial metric is a regularity score that relies on the sliding window-based PSNR. This score calculates the mean squared error between corresponding pixels of a predicted frame $\hat{f}$ and its ground truth f, where the sliding window determines the patches. Specifically, the p patches with the highest mean squared error are identified as $MSE_{p_1}, MSE_{p_2}, \ldots MSE_{p_p}$, and the PSNR of $\hat{f}$ and f is then determined using the following formula:

$$PSNR_{SW}(f, \hat{f}) = 10 \log_{10} \frac{[\max_{\hat{f}}]^2}{\frac{1}{p}\sum_{q=1}^{p} MSE_{p_q}} \quad (14)$$

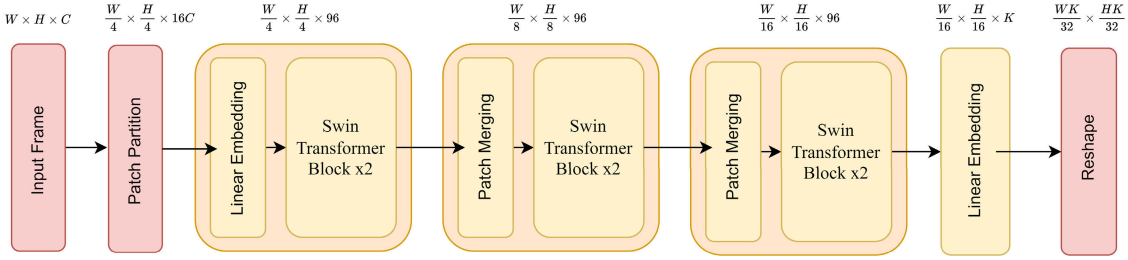


FIGURE 6. PatchGAN discriminator architecture.

FIGURE 7. Real and generated images using Binary discriminator.

FIGURE 8. Real and generated images using Patch discriminator.

after the sliding window PSNR, the regularity score of the f_{th} frame in a video is calculated as follows:

$$R(i) = \frac{PSNR_{SW}(f_i, \hat{f}_i) - min(PSNR_{SW})}{\max(PSNR_{SW}) - min(PSNR_{SW})} \tag{15}$$

The results computed on the metric mentioned in eq 15 indicate the presence of an anomaly in a long sequence of video frames and do not require a threshold to be set beforehand as an upper limit for the PSNR difference score for checking the presence of an anomaly. One significant drawback of this method is its reliance on a lengthy sequence of frames, and its inability to consider the anomaly region or type within the frame sequence. Furthermore, this approach cannot be used for real-time threat detection since the regularity score is calculated based on an accumulated sliding PSNR metric, which is determined by the minimum and maximum values of the stored data. Therefore, the method is not suitable for applications that require real-time anomaly detection, such as security systems or surveillance. However, this approach may be useful for offline analysis of video footage, where the sequence of frames can be analyzed in batches. Additionally, it is worth noting that the accuracy of the regularity score heavily depends on the quality of the ground truth frames used in the calculation. Thus, in cases where the ground truth frames are noisy or inaccurate, the regularity score may not be reliable.

TABLE 1. Comparing sliding PSNR regularity score (AUC) for different patch sizes.

Patch Size	Average Sliding PSNR AUC
7 x 7	**83.37**
14 x 14	80.75
28 x 28	73.02
56 x 56	62.84

2) ANOMALY TRACKING USING PATCH-WISE MSE AND SORT

The SORT-based anomaly detector proposed in section III-D is used during inference, and the patch-wise Mean Squared Error (MSE) is calculated for all the independent patches of size 7×7 between the original feature frame, and its corresponding patch in the predicted frame. The threshold values are taken as follows:

- Max Patches = 12
- Patch Overlap Count = 6
- Intersection Threshold between bounding box and patch = 0.7

The inference is done with a frame rate of 10fps for each test folder in all four datasets. The regularity score is a boolean value that indicates the presence of an anomaly in a frame of the sequence and the AUC score is calculated by passing the true labels and predicted anomaly labels in the AUC function to get results, the visualization is also saved as a video, that detects and tracks the anomaly bounding box.

V. RESULTS

This section presents the results of the proposed approach on the Ped1, Ped2, Avenue, and ShanghaiTech datasets. The proposed approach is evaluated using different threshold values and compared with other video anomaly detection methods.

The Area Under Curve (AUC) score was used for evaluating the model. AUC measures the overall performance of the model by calculating the area under the Receiver Operating Characteristic (ROC) curve. Generally, a higher AUC score indicates better performance of a model.

Table 1 shows the corresponding value of the average sliding PSNR AUC score of the four different datasets

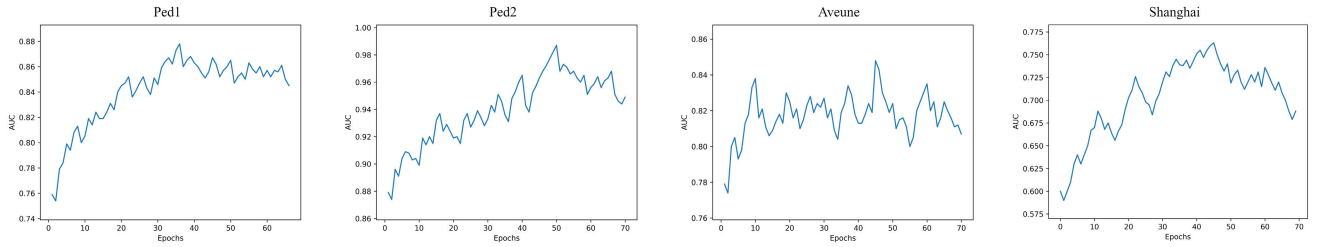


FIGURE 9. AUC curves after 70 epochs.

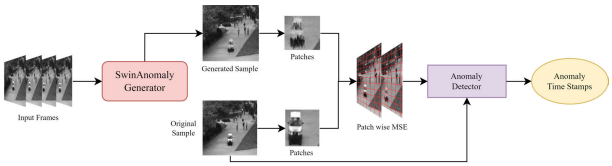


FIGURE 10. Anomaly detection pipeline.

TABLE 2. Comparing the value of tunable hyperparameter for a patch size of 7 × 7, Patch Overlap Count.

Max Patches	Patch OC	Ped1	Ped2	Avenue	Shanghai
12	6	**87.3**	**98.2**	83.02	**76.3**
	4	85.9	97.73	**84.8**	75.81
	2	79.1	92.2	80.09	71.85
8	6	84.26	93.34	78.89	69.07
	4	87.23	94.88	78.02	69.38
	2	86.11	89.86	75.98	65.6
4	6	-	-	-	-
	4	72.68	71.97	65.24	56.42
	2	71.03	73	62.66	57.9

calculated using eq. 15 for different window patch sizes. Through the table, we can observe that the patch size of 7 × 7 gives the best PSNR score for the frame size of 224 × 224 and is hence selected for further evaluation.

After establishing a consistent patch size of 7 × 7 across all evaluations (as illustrated in Table 1), we conducted additional tests to determine the optimal values of custom parameters for the total number of high MSE patches to consider in the reconstructed frame(Max Patches) and the number of such patches that overlap with bounding boxes of the objects detected in the frame by the object detection model(Patch Overlap Count thresholds). The former represents the total number of individual patches within a given frame that is considered anomalous, while the latter specifies the minimum number of these patches that must appear within the bounding box of an object for it to be classified as anomalous.

Through an analysis of Table 2, we observed that the maximum AUC score was obtained by selecting 12 patches and a minimum of 6 patch overlaps for each bounding box for the Ped1, Ped2, and Shanghai datasets. However, for the Avenue dataset, the ideal AUC score was achieved with

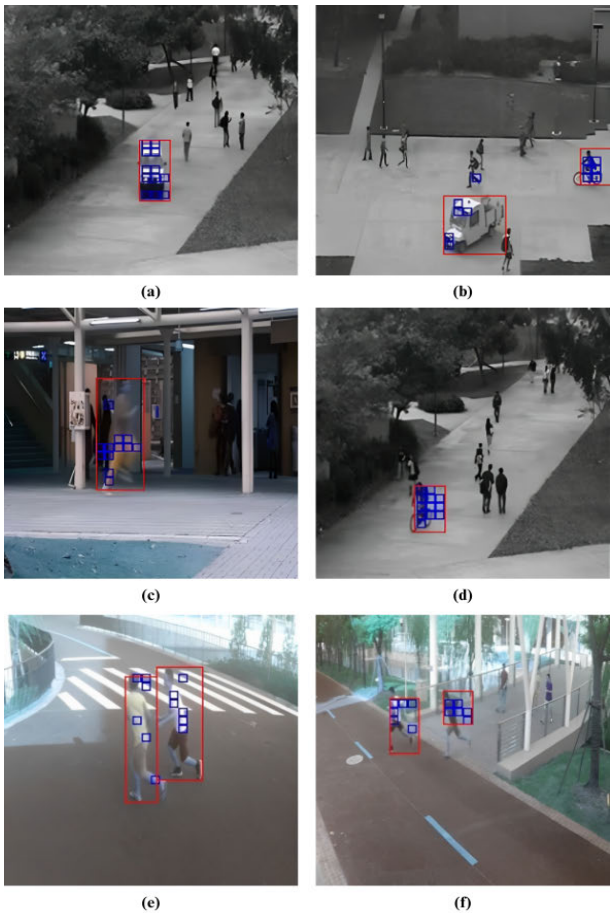


FIGURE 11. Visualization of video output showcasing anomaly detection in real-time.

12 patches and a minimum of 4 patch overlaps. By averaging the threshold values of (12, 6) and (12, 4), we determined that (12, 6) ultimately produced a superior score and, therefore, was chosen as the optimal threshold for both Max Patches and Patch Overlap Count. After getting the value of all thresholds, the model training is resumed with the validation set being evaluated on the chosen hyperparameters from Table 1 and 2 up to 70 epochs on all datasets. Figure 9 shows the AUC trend across all datasets. From the graphs, it can be seen that there was an initial spike in the AUC score for the first few epochs, then the score became constant followed by a downward trend with some variation in graph pattern in all the datasets. The

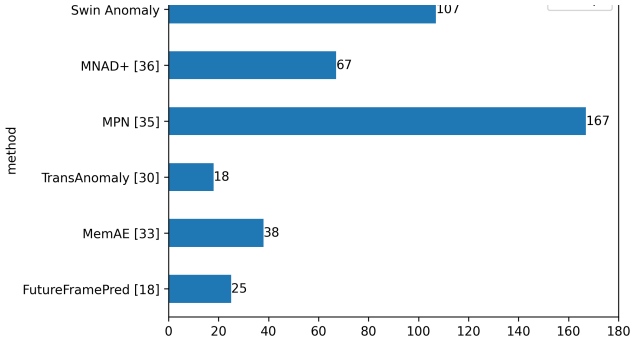


FIGURE 12. computation time comparison of different methods.

TABLE 3. AUC scores compared with different related works.

Method	Ped1	Ped2	Avenue	Shanghai
Conv2D [11]	81	90	70.2	-
ConvLSTM [15]	77.5	88.1	77	-
FutureFramePred [18]	83.1	95.4	85.1	72.8
MemAE [33]	-	94.1	83.3	71.2
TransAnomaly [30]	86.7	96.4	87	-
Dual Discriminator [34]	-	95.6	84.9	73.7
MPN [35]	85.1	96.9	89.5	73.8
MNAD+ [36]	-	97.8	88.5	70.5
RAD [40]	-	97.4	86.7	73.6
Proposed Approach	**87.3**	**98.2**	**84.8**	**76.3**

highest AUC score was achieved in the 36th, 50th, 45th, and 47th epoch on Ped1, Ped2, Avenue, and ShanghaiTech datasets respectively. A visualization of the bounding box and patch overlap using SORT tracking is shown in Fig 11.

A. COMPUTATION TIME

The inference results for the selected threshold configurations indicate that the proposed pipeline, which involves applying the SwinAnomaly generator model for future frame prediction followed by the SORT tracker, exhibits varying processing speeds based on the configuration. Our model is trained on an RTX 2060 GPU and inferred on an i7 CPU. In the CPU configuration, the pipeline achieves a rate of 1.96 fps, while in the GPU configuration, the speed increases to 4.32 fps. These results are calculated with no latency in input frame data and without utilizing the parallelization capabilities of a transformer architecture model. However, when we constrain the result to a latency of two seconds allowing frames to accumulated and then sending them together in batches, we observe an incredible boost in performance, achieving 41.66 fps on an RTX 2060 GPU and 107 fps on an RTX 3080 GPU. The comparison of the computation cost of various models is discussed in Fig 12.

Table 3 compares the proposed approach with various other approaches using the same four datasets. Latest models such as Transanomaly [30], and Meta Prototype Network [35] were also compared. As evident from the table, our method gives better AUCs for Ped1, Ped2, and ShanghaiTech however performs poorly on the Avenue dataset as compared to other state-of-the-art models.

VI. DISCUSSION

A. WHY SWIN-TRANSFORMER WAS SELECTED AS THE ENCODER

With the introduction of a hierarchical framework, swin Transformers separate the input image into more manageable, non-overlapping patches. To encode images in GANs, this hierarchical structure effectively captures local and global information. This can be extended further into PatchGAN where dividing the image segment into patches helps in capturing more context. The use of shifted windows in Swin Transformer significantly reduces the computational complexity compared to the standard self-attention used in traditional transformers. Swin Transformers' patch-based approach makes it more scalable, as it can encode the images in a parallel and memory-efficient manner. These properties of the Swin Transformer make it the best choice for the task at hand.

B. REAL-TIME ANOMALY DETECTION

The methods discussed in the related works II-D proposed an inference technique based on the ratio of the maximum and minimum Peak Signal-to-Noise Ratio (PSNR). This approach enables the processing of long sequences of frames in batches, offering a tradeoff between latency and achieving better frame rates. In contrast, our method employs a real-time processing approach, handling one input frame sequence at a time. However, we have optimized the inference pipeline to ensure efficient and streamlined processing, achieving a frame rate of 4.32 fps. Moreover, by introducing a small latency of 2 seconds, we can leverage the batch inference of transformers and obtain frame rates up to 107 fps, as discussed in Section V. Overall, while the related works focus on batch processing with a latency-fps tradeoff, our method prioritizes real-time processing.

C. LOW AUC SCORE ON AVENUE AND SHANGHAI DATASETS

In our inference, we applied a resizing operation to both the Avenue and Shanghai datasets, setting the frame size to (224×224) pixels. This step was necessary as it aligns with the image input requirements of our model. However, we observed that this resizing process led to a loss of fine-grained details in the frames due to dilation. Given that the images were in color and captured at a high frame rate, a strong correlation between consecutive frames emerged. Consequently, during training, the model tended to prioritize copying frame features rather than focusing on the accurate reconstruction of future frames. Moreover, we encountered limitations with the object detection model when restricting our results to an average selection of patch overlap count. This restriction hindered the detection of shadowed individuals in the background and small segment anomalies, leading to occasional failure in anomaly detection. To enhance our anomaly detection system, we can address the limitations by exploring alternative object detection models

and leveraging a combination of diverse architectures for improved accuracy. To reduce frame correlation, we will carefully select non-similar frames during training, promoting feature learning over copying. Additionally, we plan to upscale image size from swin-tiny to swin-large to capture finer details and achieve better detection results.

VII. CONCLUSION

In this paper, we have introduced SwinAnomaly, an anomaly detection model based on Swin Transformers. SwinAnomaly utilizes the parallelization feature of transformers and tracking capabilities of the SORT algorithm for real-time anomaly detection and tracking. To improve frame reconstruction, we modified the skip connections and bottleneck layer to extend the window attention of 3D and connect it to 2D Swin blocks in the autoencoder. Real-time anomaly analysis and recognition are achieved by non-overlapping patch intersection with the bounding boxes of the tracking boundaries of Kalman filters using a fast object detection model. Our experiments on four baseline datasets demonstrate that our model outperforms state-of-the-art prediction and reconstruction-based anomaly detection methods and provides flexibility to filter the size and persistence of anomalies using tunable parameters. However, a major drawback of our approach is the static settings of parameters at inference time, which are unresponsive to outlier anomalies. Another limitation of the proposed approach is the high sensitivity of Swin Transformers towards different lighting conditions, which can affect the performance of the model if it is not trained on different lighting conditions. Future work includes extending this approach to address its limitations by exploring other object detection models or tracking techniques and developing a framework to track anomalies in multiple CCTV cameras.

ACKNOWLEDGMENT

The authors would like to thank King Khalid University, Saudi Arabia, for providing administrative and technical support and also would like to thank the editors and the reviewers for their helpful comments and suggestions on the manuscript.

REFERENCES

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[2] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid, "ViViT: A video vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6816–6826.

[3] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[4] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, Jan. 2018.

[5] K. Lee, H. Chang, L. Jiang, H. Zhang, Z. Tu, and C. Liu, "ViTGAN: Training GANs with vision transformers," 2021, *arXiv:2107.04589*.

[6] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*.

[7] C. Li, Y. Jiang, and M. Cheslyar, "Embedding image through generated intermediate medium using deep convolutional generative adversarial network," *Comput., Mater. Continua*, vol. 56, no. 2, pp. 313–324, 2018.

[8] U. Demir and G. Unal, "Patch-based image inpainting with generative adversarial networks," 2018, *arXiv:1803.07422*.

[9] G. Welch and G. Bishop, "An introduction to the Kalman filter," Tech. Rep., 1995.

[10] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 3464–3468.

[11] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 733–742.

[12] Y. Fan, G. Wen, D. Li, S. Qiu, M. D. Levine, and F. Xiao, "Video anomaly detection and localization via Gaussian mixture fully convolutional variational autoencoder," *Comput. Vis. Image Understand.*, vol. 195, Jun. 2020, Art. no. 102920.

[13] N. Li and F. Chang, "Video anomaly detection and localization via multivariate Gaussian fully convolution adversarial autoencoder," *Neurocomputing*, vol. 369, pp. 92–105, Dec. 2019.

[14] Y. S. Chong and Y. H. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," in *Proc. Int. Symp. Neural Netw.* Cham, Switzerland: Springer, 2017, pp. 189–196.

[15] W. Luo, W. Liu, and S. Gao, "Remembering history with convolutional LSTM for anomaly detection," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 439–444.

[16] T. N. Nguyen and J. Meunier, "Anomaly detection in video sequence with appearance-motion correspondence," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1273–1283.

[17] Y. Chang, Z. Tu, W. Xie, and J. Yuan, "Clustering driven deep autoencoder for video anomaly detection," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 329–345.

[18] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection—A new baseline," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6536–6545.

[19] J. R. Medel and A. Savakis, "Anomaly detection in video using predictive convolutional long short-term memory networks," 2016, *arXiv:1612.00390*.

[20] Y. Guo, W. Liao, Q. Wang, L. Yu, T. Ji, and P. Li, "Multidimensional time series anomaly detection: A GRU-based Gaussian mixture variational autoencoder approach," in *Proc. Asian Conf. Mach. Learn.*, 2018, pp. 97–112.

[21] M. Ye, X. Peng, W. Gan, W. Wu, and Y. Qiao, "AnoPCN: Video anomaly detection via deep predictive coding network," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1805–1813.

[22] Y. Li, Y. Cai, J. Liu, S. Lang, and X. Zhang, "Spatio-temporal unity networking for video anomaly detection," *IEEE Access*, vol. 7, pp. 172425–172432, 2019.

[23] Y. Tang, L. Zhao, S. Zhang, C. Gong, G. Li, and J. Yang, "Integrating prediction and reconstruction for anomaly detection," *Pattern Recognit. Lett.*, vol. 129, pp. 123–130, Jan. 2020.

[24] M.-I. Georgescu, A. Barbalau, R. T. Ionescu, F. S. Khan, M. Popescu, and M. Shah, "Anomaly detection in video via self-supervised and multi-task learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12737–12747.

[25] Z. Liu, Y. Nie, C. Long, Q. Zhang, and G. Li, "A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13568–13577.

[26] G. Wang, Y. Wang, J. Qin, D. Zhang, X. Bao, and Di Huang, "Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2022, pp. 494–511.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[28] K. Deshpande, N. S. Punn, S. K. Sonbhadra, and S. Agarwal, "Anomaly detection in surveillance videos using transformer based attention model," 2022, *arXiv:2206.01524*.

[29] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.
[30] H. Yuan, Z. Cai, H. Zhou, Y. Wang, and X. Chen, "TransAnomaly: Video anomaly detection using video vision transformer," *IEEE Access*, vol. 9, pp. 123977–123986, 2021, doi: 10.1109/ACCESS.2021.3109102.
[31] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-Unet: Unet-like pure transformer for medical image segmentation," in *Computer Vision—ECCV 2022*. Cham, Switzerland: Springer, 2023, pp. 205–218.
[32] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video Swin transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3192–3201.
[33] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. Van Den Hengel, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1705–1714.
[34] F. Dong, Y. Zhang, and X. Nie, "Dual discriminator generative adversarial network for video anomaly detection," *IEEE Access*, vol. 8, pp. 88170–88176, 2020.
[35] H. Lv, C. Chen, Z. Cui, C. Xu, Y. Li, and J. Yang, "Learning normal dynamics in videos with meta prototype network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15420–15429.
[36] H. Park, J. Noh, and B. Ham, "Learning memory-guided normality for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14360–14369.
[37] J. Yao and S. Jin, "Multi-category segmentation of Sentinel-2 images based on the Swin UNet method," *Remote Sens.*, vol. 14, no. 14, p. 3382, Jul. 2022.
[38] W. Liu, H. Chang, B. Ma, S. Shan, and X. Chen, "Diversity-measurable anomaly detection," 2023, *arXiv:2303.05047*.
[39] T. Reiss and Y. Hoshen, "Attribute-based representations for accurate and interpretable video anomaly detection," 2022, *arXiv:2212.00789*.
[40] V.-T. Le and Y.-G. Kim, "Attention-based residual autoencoder for video anomaly detection," *IntAppl Intell.*, vol. 53, no. 3, pp. 3240–3254, Feb. 2023, doi: 10.1007/S10489-022-03613-1.

ARPIT BAJGOTI was born in Delhi, India, in 2002. He is currently pursuing the B.Tech. degree in computer science and engineering with the Maharaja Surajmal Institute of Technology. He has published a research paper on acoustic modeling and ASR systems and has professional experience in audio processing and computer vision, specifically in large language models.

RISHIK GUPTA was born in Delhi, India, in 2002. He received the Diploma degree in data science from the Indian Institute of Technology Madras. He is currently pursuing the B.Tech. degree in computer science and engineering with the Maharaja Surajmal Institute of Technology. He has published papers in audio signal processing and has a keen academic interest. His research interests include speech detection, computer vision, and natural language processing.

PRASANALAKSHMI BALAJI (Member, IEEE) received the master's degree in computer science from Bharathidasan University, the master's degree in computer engineering from Anna University, India, in 2008, and the Ph.D. degree in computer science from the Research & Development Centre, Bharathiar University, India, in 2014. She is currently an academician and a research fellow with over 16 years experience. From 2019 to 2021, she was a Research Fellow with the Centre for Artificial Intelligence, Saudi Arabia. Since 2017, she has been with the Computer Science Department, King Khalid University, Abha, Saudi Arabia. She holds three executive education courses on artificial intelligence from IIIT Hyderabad, Artificial Intelligence Engineer Master's Program from IBM, and data analytics with Python from IBM Watson. She has filed seven patents with Indian and international scope. She has presented/published 40 research papers at both national and international level. She has various appraisals and achievements to her credit that includes the Young Scientist Award, Best Researcher Award, and won the first prize for Research and Innovation Award from King Khalid University, in 2022. Her research interests include cryptography, machine learning, deep learning, medical imaging, and neural networks.

RINKY DWIVEDI received the B.Tech. degree in computer science and engineering from Guru Gobind Singh Indraprastha University, Delhi, in 2004, the M.E. degree in computer technology and application from the Delhi College of Engineering, Delhi, in 2008, and the Doctorate degree from Delhi Technological University, New Delhi, in 2016. She has over 19 years of experience in academics, currently an Associate Professor and the Head of the Department of Computer Science Engineering, Maharaja Surajmal Institute of Technology, New Delhi. She has published more than 20 research papers in reputed journals and conference proceedings and has also authored books.

MEENA SIWACH is currently pursuing the Ph.D. degree from Guru Gobind Singh Indraprastha University, Delhi. She has over 15 years of experience in academics and also an Assistant Professor with the Maharaja Surajmal Institute of Technology, Delhi. She is highly interested in participating in the development of new interdisciplinary programs of study and highly enthusiastic about research and innovative ideas. She has published over 15 research papers in national/international journals and conferences. Her research interests include data mining, machine learning, and deep learning.

DEEPAK GUPTA received the B.Tech. degree from Guru Gobind Singh Indraprastha University, the M.E. degree from Delhi Technological University, and the Ph.D. degree from Dr. A. P. J. Abdul Kalam Technical University. He is currently an accomplished academic with a diverse educational background. He has completed a postdoctoral research with the National Institute of Telecommunications, Brazil, currently working as an Assistant Professor in the Computer Science and Engineering Department of Maharaja Agrasen Institute of Technology, New Delhi, India. He has authored numerous journal articles and books, secured patents, and received the 2021 IEEE System Council Best Paper Award. He is involved in organizing conferences, serves as an associate editor for respected journals and has been recognized as a top researcher in healthcare applications. Additionally, he promotes startups and holds a series editor roles for prominent publishers. He has secured research grants from international funding agencies and is a co-PI in an Indo–Russian joint project.

Journal of Fluorescence
https://doi.org/10.1007/s10895-023-03449-z

RESEARCH

Synthesis, Structural and Fluorescence Investigations of Novel $Li_2Ba_5W_3O_{15}$:Sm^{3+} Phosphors for Photonic Device Applications

Anu[1] · A. S. Rao[1]

Received: 24 August 2023 / Accepted: 21 September 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

In the current study, Sm^{3+} ions doped Lithium Barium Tungstate ($Li_2Ba_5W_3O_{15}$) (LBW) phosphors with the ability to emit orange-red light were made using the traditional high-temperature solid-state reaction technique. The structure and phase of the as-synthesized phosphor samples were examined via X-ray diffraction (XRD) patterns. The diffraction peaks of the undoped LBW and Sm^{3+} ions doped LBW phosphors closely resemble those of the Joint Committee on Powder Diffraction Standards (JCPDS) pattern with card number 01–072-1717. Scanning electron microscopy (SEM) was employed for the analysis of the morphological characteristics of the synthesized phosphor material. Fourier Transform Infrared (FT-IR) spectroscopy was used to study several vibrational and molecular bands present in the host matrix. Using diffuse reflectance spectra (DRS), the optical band gap values (E_g) were evaluated by applying Tauc's method. The photoluminescence (PL) spectra characteristics at $\lambda_{ex} = 336$ nm indicate the emission of dopant ions (Sm^{3+}) in the deep orange-red region corresponding to $^4G_{5/2} \rightarrow {}^6H_{5/2}$ transition (at 581 nm) with concentration quenching after 2 mol % of Sm^{3+} ions. Using the PL spectra, the CIE chromaticity coordinates of LBWS2.0 phosphor were estimated and found in the deep visible orange-red area, indicating the potential use of the prepared phosphor material for phosphor-converted white light emitting diodes (w-LEDs) applications. Double exponential behaviour can be seen in the PL decay spectral profiles obtained under $\lambda_{em} = 581$ nm and $\lambda_{ex} = 336$ nm. The experimental lifetimes (τ_{exp}) decrease as the concentration of Sm^{3+} ions rise. The temperature-dependent PL (TDPL) and activation energy results show that the as-synthesized phosphor has considerably superior thermal stability. The results of the current research contemplate us the applicability of Sm^{3+} ions doped LBW phosphor for photonic devices such as w-LEDs.

Keywords Sm^{3+} · Rietveld, scanning electron microscopy (SEM) · Bi-exponential · TDPL

Introduction

White light-emitting diodes (w- LEDs) have taken the position of traditional lighting devices in the SSL industry due to their superior qualities, especially extended durability, great brightness, sustainability, and low power consumption [1, 2]. Commercial w-LEDs built on a lnGaN chip with a $Y_3Al_5O_{12}$:Ce^{3+}($YAG:Ce^{3+}$) phosphor have high correlated colour temperatures (CCT) and poor colour rendering indices (R_a) due to the lack of red element in the visible light region, which restricts their further growth [3, 4]. Another alternate technique for creating w-LEDs involves coating an InGaN chip with trichromatic phosphors (phosphors that emit blue, green, and red light). Excellent red phosphors like nitride are produced under extreme conditions (high pressures and high temperatures). To address the demand for high-performance w-LEDs, it is crucial to create red-emitting phosphors that can be efficiently activated by ultraviolet (UV) or near ultraviolet (n-UV) light. Rare-earth (RE) ions are renowned for their exceptional efficacy as dopant ions within host lattices. The role of RE ions in visible luminescence is crucial because of f-f transitions. These f-f transitions lead to luminous sharp emission when RE ions are introduced as a dopant to an appropriate crystal lattice [5–8]. Furthermore, these RE ions are easily energized by UV/n-UV radiation absorption, which results in the down-conversion process releasing light in the visible area. Among all the RE activator ions, Sm^{3+} ions stand out as highly effective and radiation-stable candidates, proficiently emitting red or

✉ A. S. Rao
drsrallam@gmail.com

[1] Department of Applied Physics, Delhi Technological University, Bawana Road, New Delhi 110042, India

Published online: 07 October 2023

orange-red colour and having remarkable colour purity (CP) [9, 10]. Additionally, it is crucial to choose an activator for phosphor materials that can easily ignite Sm^{3+} doped materials when using commercially available chips. Moreover, substances containing Sm^{3+} ions have been the subject of much research owing to their $^4G_{5/2}$ to $^4H_{J/2}$ (J = 5, 7, 9, and 11) emissions in the visible red region [11]. Several Sm^{3+} ions activated host matrices have already been reported, such as $Ca_2Gd_8Si_6O_{26}$ [12], $Ba_3La(PO_4)_3$ [13], and $Ba_5Zn_4Y_8O_{21}$:Sm^{3+} [14], which results in absorption of n-UV light and emission in the intense orange-red region. Ekmekçi and his colleagues prepared a single phase BTO:Sm^{3+} ceramic phosphor and studies their structural and luminescent properties [15]. Poonam and her team synthesized BBiVO:Sm^{3+} nanomaterial; an orange light emitters, for various lighting applications [16]. Zhou et al. successfully prepared a series of Sm^{3+} ions doped $NaGdMgTeO_6$ phosphor via traditional solid-state reaction method and found them suitable for horticultural and w-LED applications [17]. Using a hydrothermal process, Yang and colleagues successfully synthesized orange-red phosphors of $Sr_3Al_2O_6$ doped with Sm^{3+} ions and their investigations revealed that the Sm^{3+} ions readily occupy the symmetric center site [18]. Li and co-researchers developed a Sm^{3+}-doped $Ba_3LaNb_3O_{12}$ phosphor that exhibits red–orange emission unaffected by thermal quenching and resistant to moisture [19]. Govind B. Nair and his co-author prepared $CaZr_4(PO_4)_6$:Sm^{3+} phosphor which can develop a w-LED of high CRI with the combination of green phosphor [20].

Currently, a wide variety of host materials, such as phosphates, silicates, fluorides, ceramics, and tungstates, can be doped with RE ions. Tungstate phosphor is an example of a typical self-activated luminous material with superior optical characteristics, low phono energy, great mechanical strength, high thermal and chemical stability, and eco-friendly nature [21–24]. Despite possessing good luminescence properties, physical stability, and chemical stability, the use of poly tungstate matrix is rather unusual [24–28] and single and double tungstate with Scheelite structures as a host are frequently used for SSL devices. To our current understanding, the utilization of Sm^{3+} ion-doped Lithium Barium Tungstate (LBW) phosphor was initially introduced to explore its potential uses in white light-emitting diodes (w-LEDs).

Synthesis and Instrumentations

The LBW phosphors doped with samerium ions were synthesized using a typical high-temperature reaction approach. The starting ingredients included Li_2CO_3 (99.99%), $BaCO_3$ (99.99%), Sm_2O_3 (99.99%), and WO_3 (analytical grade) acquired from commercial sources. An electronic balance was used to weigh the stoichiometric reactants, which were then completely combined and pulverized in an agate mortar for 45 min. The homogenous mixture was then put into the alumina crucible and heated in a programmable electric muffle furnace for 360 min at 1473 K. The final products were re-grounded once again when they reached room temperature for further analysis. The flowchart of the steps taken throughout the synthesis process is shown in Fig. 1.

The XRD patterns of the synthesized samples were recorded using the Bruker Model D8 Advance instrument (with an error of 0.3 degree deviation & employing Cu-K radiation (λ = 1.54 Å) filtered through Nickel). To investigate the structural parameters thoroughly, the Rietveld refinement using the FullProf Suite program was used. Scanning electron microscopy (SEM) recordings were performed to study the surface morphology by using JEOL JSM-6490 operated at 10 kV accelerating voltage. Using a Perkin-Elmer Frontier spectrometer, the FT-IR spectra for the as-prepared phosphor have been obtained. The DRS was recorded using a JASCO model V-923 UV–Vis-NIR spectrophotometer (with 2.0 nm resolution). PL excitation and emission patterns were studied using a JASCO FP-8300 spectrofluorometer (1.0 nm resolution) and a Xenon excitation source. To study the PL decay profiles, a Fluorescence Spectrometer (Edinburgh FL 920 with an error of 0.004 ms) was used. All measurements were made in a temperature-controlled environment. TDPL investigations were carried out using an Ocean Optics Flame-S-XR1-ES spectrometer.

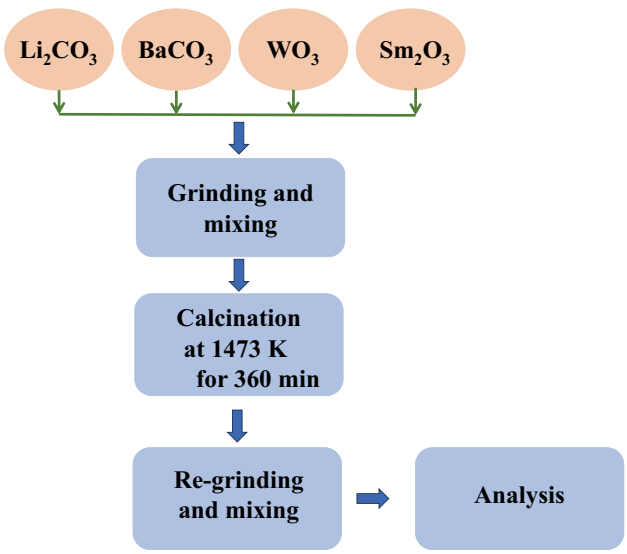


Fig. 1 The flowchart of the steps involved in the synthesis process

Results and Their Discussion

X-ray Diffraction Analysis

The un-doped LBW and Sm^{3+} ions doped LBW phosphors' crystalline structure and phase purity were examined using XRD patterns. The XRD patterns for un-doped LBW and LBWS1.0 phosphors are shown in Fig. 2(a), together with the appropriate standard JCPDS data (01–072-1717), which contains the lattice parameters a = 5.760 Å = b, c = 23.742 Å, α = β = 90° and γ = 120° and V = 682.17 $Å^3$, as well as the space group $P6_3mc$ and space group number is 186, was found to closely match all the peaks. The well-matched XRD patterns of the phosphor samples with the JCPDS database showed their high purity and therefore, enabled us to speculate the absence of any secondary or impurity phases in the prepared phosphors even with the addition of Sm^{3+} ions. The synthesized phosphors exhibit strong and precise diffraction peaks, demonstrating their high crystallinity, which is necessary for significantly better illumination. It is also required to determine the radius percentage difference (D_r) between substituted ions and activator ions to better understand the dopant Sm^{3+} ions occupying Ba^{2+} sites in the LBW host lattice. The value of D_r is defined by the following expression [29]:

$$D_r = \frac{R_1(CN) - R_2(CN)}{R_1(CN)} \times 100\% \quad (1)$$

here, $R_1(CN)$ and $R_2(CN)$ are the radii of substituted (Ba^{2+}) ions and activated Sm^{3+} ions, respectively. The value of D_r was determined using Eq. (1) to be 26.30%, indicating that Sm^{3+} could effectively replace Ba^{2+} ion positions in LBW host lattices [29].

The well-known Debye–Scherrer formula described below was used to determine the crystallite size (D) while taking into consideration the primary diffraction peaks of the recorded XRD data [30]:

$$D = \frac{k\lambda}{\beta\cos\theta} \quad (2)$$

where k is the form factor (= 0.94), D is the crystallite size, λ signifies the wavelength of the X-rays (1.54 Å), and θ is the angle of diffraction, β means full width at half maximum (FWHM). The necessary values of β and θ were extracted from the XRD patterns to determine crystallite size. Using the Debye-Scherer relation, the evaluated value of the average crystallite size for the titled glass was found to be equal to 82.87 nm.

The Rietveld refinement was conducted on the as-prepared LBW sample using software from the Fullprof package to investigate the crystal structure and phase. The profile shape refinement has been defined using the pseudo-Voigt function. Rietveld refinement pattern of an un-doped LBW sample is shown in Fig. 2(b). The red solid line with a circle depicts the experimentally determined data, and the black line reflects the observed data pattern. The green vertical lines show the LBW phosphor Bragg angles (2θ). The solid blue line illustrates the distinction between calculated and experimental patterns.

Table 1 displays the improved parameters for the crystal structure. The calculated goodness of fit (χ^2) was found to be 1.64. This outcome demonstrates that pure phases devoid of secondary phases can be produced by LBW phosphor samples with octahedral structures.

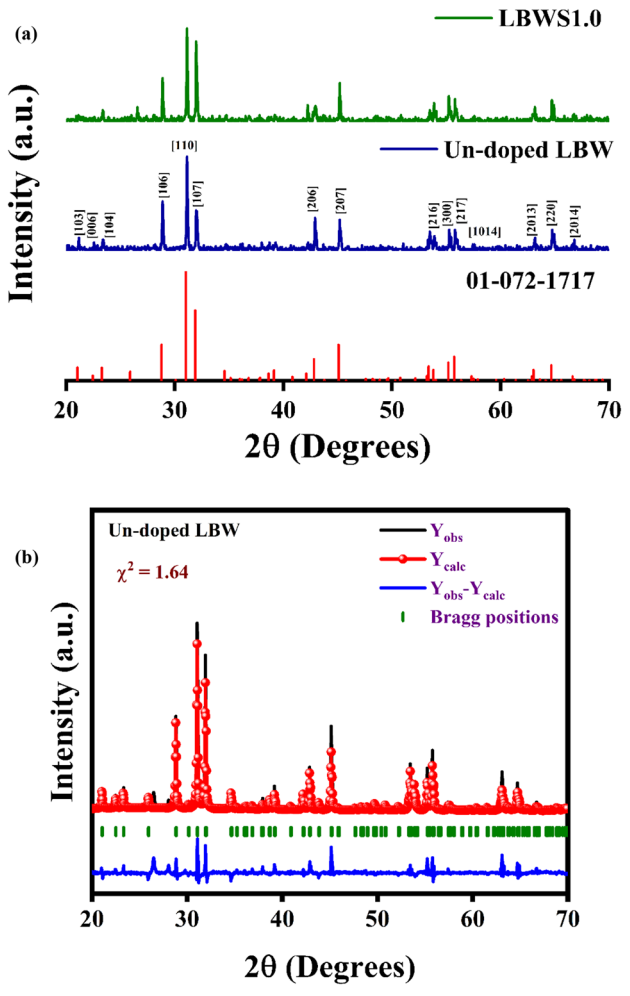


Fig. 2 **a** The XRD patterns for an un-doped LBW and LBWS1.0 phosphor. **b** Rietveld refinement of an un-doped LBW sample

SEM: Morphological Studies

The SEM metaphors of an un-doped LBW and LBWS2.0 phosphor were obtained to better comprehend the data pertaining to the surface morphology and the particle size of the

Table 1 Refinement parameters for un-doped LBW material

Sample	JCPDS	Un-doped LBW
a = b (Å)	5.76	5.757
c (Å)	23.742	23.714
V (Å^3)	682.17	680.863
R_B	-	12.4
R_p (%)	-	30.3
R_{wp} (%)	-	33.3
R_{exp} (%)	-	11.3
χ2	-	1.64
Density (g/cm^3)	7.264	7.387

titled phosphor samples. According to Fig. 3(a) and (b), the LBW phosphor materials have unequal polyhedron shapes and particle sizes. Additionally, owing to the hand grinding and a very high-temperature reaction process, some phosphor particles also show low agglomeration. The bulk of commercial phosphor particles available today has a size between 2 and 10 µm (µm). As a result, the samples' particle sizes are in the micrometer (µm) range, which makes them perfect for coating in displays and white LEDs [24].

FT-IR Spectral Analysis

To further understand the type of bonding in an un-doped LBW sample, infrared spectroscopy (IR) investigations were conducted. The various vibrational bands are represented by the FT-IR spectra of an un-doped LBW phosphor in the range 400–4000 cm^{-1} in Fig. 4. As seen in Fig. 4, the FT-IR spectra have bands close to 576, 700, 836, 1064, 1413, 2886 and 2984 cm^{-1}. The bands at 500–836 cm^{-1} are connected to the W–O symmetric and antisymmetric stretching and O-W–O symmetric and antisymmetric stretching deformation vibrations of the tungstate group [31]. The peak between 1064 cm^{-1} seems to be caused by the W–O bond's asymmetric stretching vibration [32]. Surface CO_3^{2-} species can be attributed to the band at 1413 cm^{-1} [31]. The band at 2886 cm^{-1} may represent the presence of hydroxyl radicals, and the band at 2984 cm^{-1} is believed to depict the antisymmetric stretching of adsorbed water molecules. As a result, the FT-IR spectrum verified the distinctive bonding found in the synthesized LBW samples.

Diffuse Reflectance Spectral Analysis (DRS)

One critical factor for phosphors used in w-LEDs is the optical band gap (E_g). The diffuse reflectance spectra (DRS) for LBW: x Sm^{3+} phosphors (x = 0, 0.5, 1.0, 2.0, 3.0 and 5.0 mol%) in the 200–1500 nm range are shown in Fig. 5. The spectra consists of a broad absorption band with a maximum wavelength (λ_{max}) of 314 nm, lying within 200 to 400 nm. This should be attributed to the charge transfer between the 2p orbital of oxygen to 5d orbital of tungsten, causing the absorption zone in the range of 280–330 nm. At wavelengths 404, 474, 944, 1072, 1236, 1362, and 1474 nm, respectively, the spectra show a transition from the ground level $^6H_{5/2}$ to various excited levels $^4F_{7/2}$, $^4I_{11/2}$, $^6F_{11/2}$, $^6F_{9/2}$, $^6F_{7/2}$ and $^6H_{13/2}$ [33].

The reflectance spectrum is transformed to absorbance data using the Kubelka–Munk function given as [34]:

$$F(R) = \frac{(1-R^2)}{2R} = \frac{K}{S} \tag{3}$$

which relates the ratio of the absorption coefficient (K) and scattering coefficient (S) of the sample with diffuse reflectance (R).

In a parabolic band structure, the optical band gap (E_g) and the semiconductor oxides absorption factor (B) can be interrelated as follows [35, 36]:

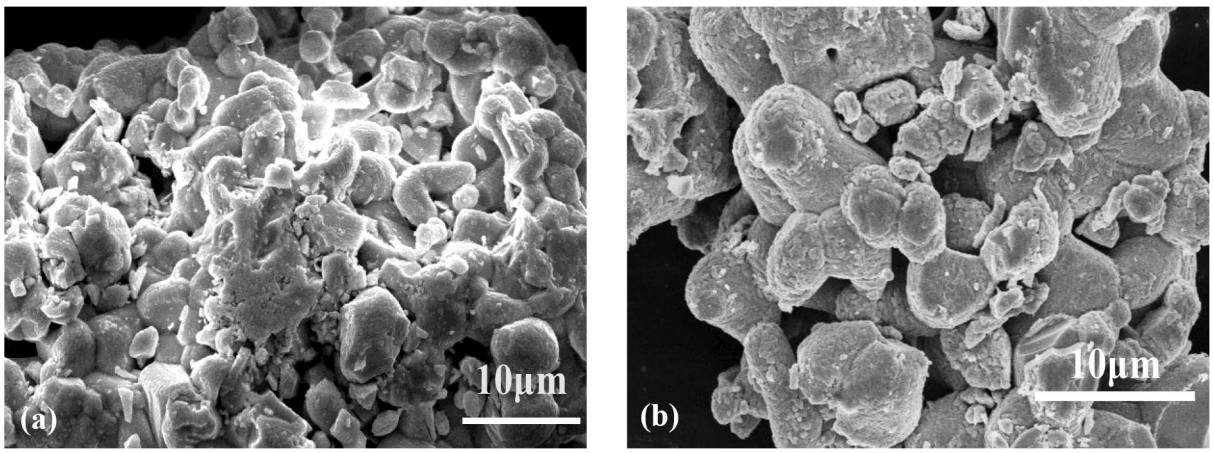


Fig. 3 **a**, **b** The SEM metaphors of an un-doped LBW and LBWS2.0 phosphor

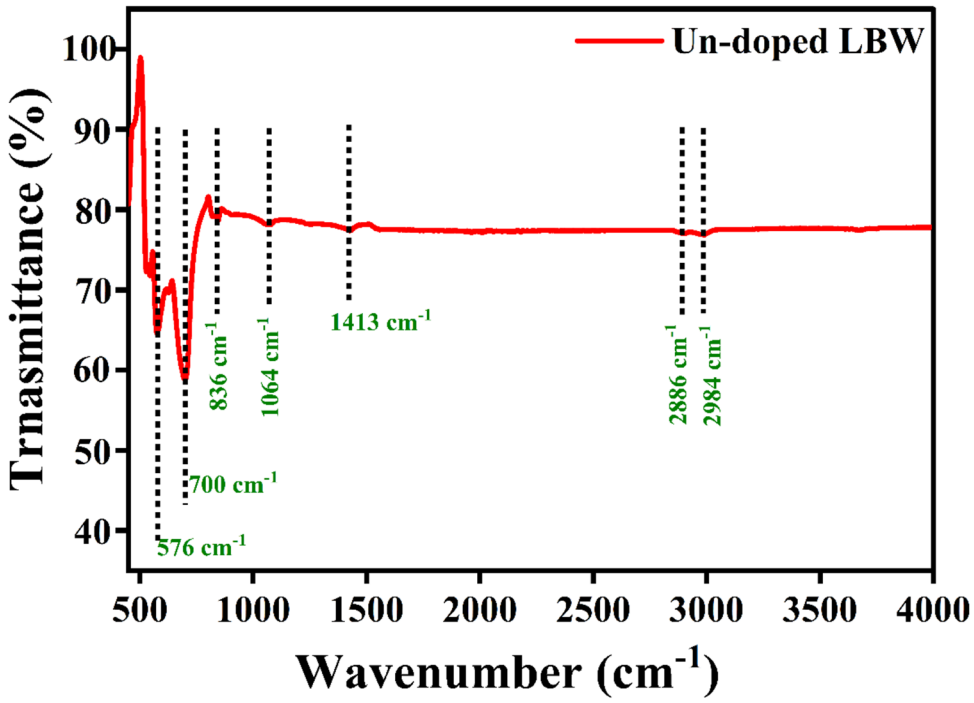


Fig. 4 FT-IR spectrum of an un-doped LBW sample

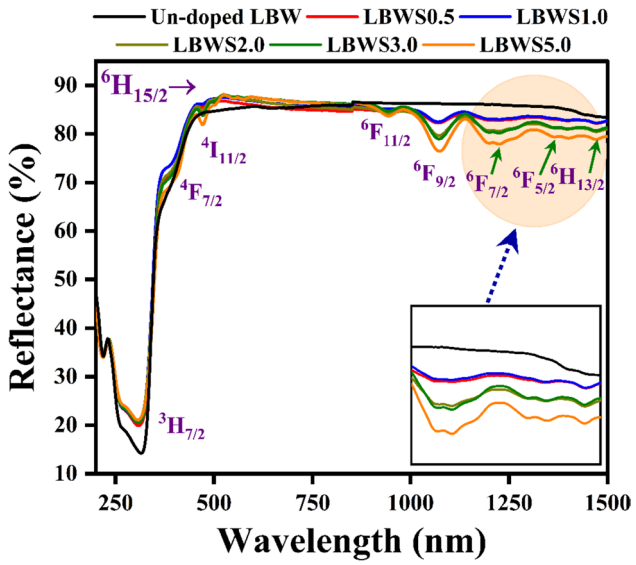


Fig. 5 UV–vis DRS of an un-doped LBW and Sm^{3+} ions doped LBW phosphors

$$\alpha = \frac{B}{h\nu}(h\nu - E_g)^n \tag{4}$$

where n is a constant related to various types of electronic transitions, $h\nu$ is the photon energy, B is a proportionality constant, E_g is the optical band gap, and α is the linear absorption coefficient of the material; $n = 1/2$, 2 for a direct and indirect allowed transition and $n = 1/3$ and 3 for direct and indirect forbidden transition, respectively.

The absorbance coefficient is found to be proportional to absorbance, which is connected to reflectance, $F(R)$ [37]. Therefore, in the calculations, α is changed to $F(R)$, and the optical bandgap (E_g) of undoped LBW and Sm^{3+} ions doped LBW phosphors were then determined using the following expression [34, 38, 39]:

$$F(R)h\nu = C(h\nu - E_g)^n \tag{5}$$

where C is a constant and $h\nu$ denotes the incident photon's energy. The values of n characterized the nature of the transition.

Figure 6 represents a graph between $h\nu$ (incident energy) and $[F(R)h\nu]^2$, called Tauc's Plot for an un-doped LBW and Sm^{3+} ions doped LBW phosphor samples. By extrapolation of the linear portion of the graph to the x-axis, the bandgap values of phosphor samples were evaluated. The results show that the E_g decreases with increasing Sm^{3+} ions concentration due to the formation of the intermediate energy levels of the Sm^{3+} ions in the LBW host matrix. The variation of bandgap with Sm^{3+} ions concentration is presented in Table 2.

Photoluminescence Spectral Analysis

The study PL spectrum features at room temperature allowed us to investigate the PL characteristics of Sm^{3+} ions doped LBW phosphors. Figure 7 displays the PL excitation spectrum of LBWS2.0 phosphor in the spectral range of 250–500 nm at $\lambda_{em} = 581$ nm (i.e., Em = 581 nm). Three excitation bands corresponding to transitions from $^6H_{5/2}$ to

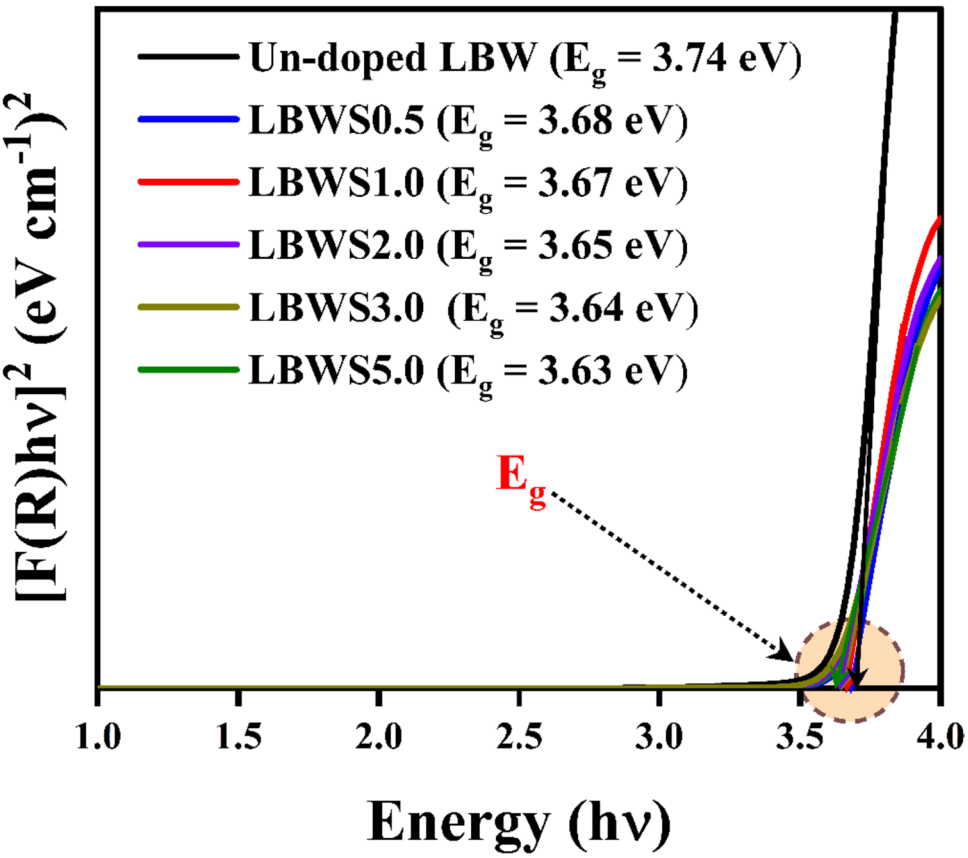


Fig. 6 Direct band gap for of the un-doped LBW and Sm^{3+} ions doped LBW phosphors

Table 2 Band gap (E_g, eV) of an un-doped LBW and Sm^{3+} ions doped LBW phosphors

Sm^{3+} (mol %)	System Name	Band Gap (eV)
0	**Un-doped LBW**	3.74
1	**LBWS0.5**	3.68
2	**LBWS1.0**	3.67
3	**LBWS2.0**	3.65
4	**LBWS3.0**	3.64
5	**LBWD5.0**	3.63

$^6P_{7/2}$ (378 nm), $^4F_{7/2}$ (407 nm), and $^4I_{11/2}$ (467 nm), respectively, along with a charge transfer band (CTB, 336 nm) attributed to the O^{2-}-W^{6+} bond comprise the PL spectra [40, 41]. The CTB is the strongest peak among all the excitation peaks therefore considered as the best excitation wavelength for capturing PL emission spectra. The CTB, can be seen in the excitation spectra of Sm^{3+}-activated phosphors, serves as evidence that energy is transferred from the host to the activator Sm^{3+}. To address the needs of UV LEDs, the near-UV excitation range between 345 and 420 nm is quite suitable [42].

The PL emission spectra for the as-prepared phosphor LBW:xSm^{3+} (x = 0.5, 1.0, 2.0, 3.0, and 5.0 mol%) at

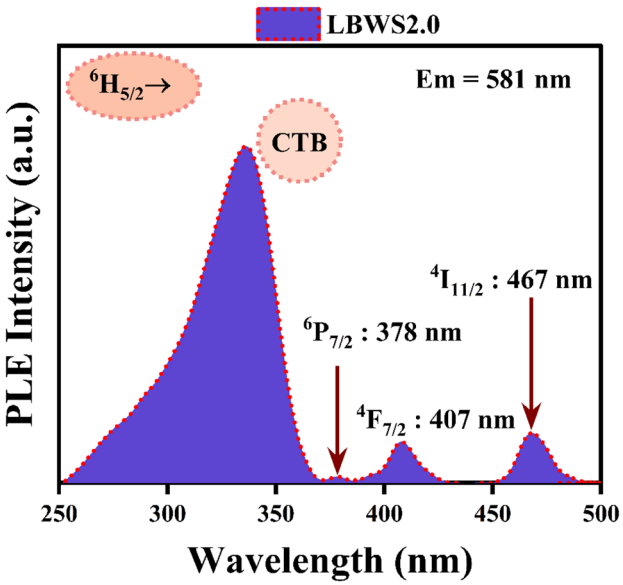


Fig. 7 PLE spectra for LBWS2.0 phosphor

336 nm excitation wavelength (i.e., Ex = 336 nm) are shown in Fig. 8. Three unique sharp peaks of Sm^{3+} ion emission in the PL spectra are attributed to the $^4G_{5/2}$ to $^6H_{5/2}$ (581 nm), $^6H_{7/2}$ (613 nm), and $^6H_{9/2}$ (668 nm) transitions, respectively

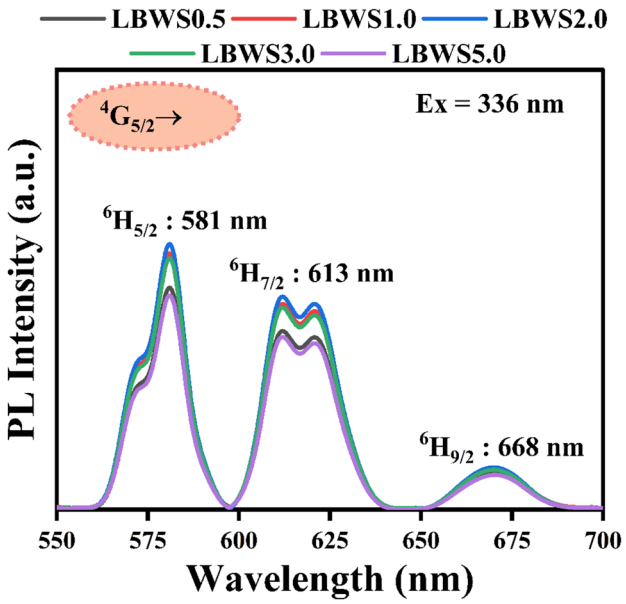


Fig. 8 PL emission spectra of the Sm^{3+} ions doped LBW phosphors at Ex = 336 nm

in 550–700 nm wavelength range [42]. Amid all, the peak corresponding to $^4G_{5/2} \rightarrow {}^6H_{5/2}$ (581 nm) transition is considerably more efficient. The shifting in the PL peaks in the PL spectra can be attributed to the stark splitting [43]**.** The PL emission peak at 668 nm, attributed to the $^4G_{5/2} \rightarrow {}^6H_{9/2}$ transition, is widely recognized as being solely an electric dipole (ED) transition ($\Delta J \leq 6$). Conversely, the emission peak equivalent to $^4G_{5/2} \rightarrow {}^6H_{5/2}$ transition (at 581 nm), is exclusively a magnetic dipole (MD) transition ($\Delta J = 0, \pm 1$). However, the transition corresponding to $^4G_{5/2} \rightarrow {}^6H_{7/2}$ (at 613 nm), exhibits a combination of both ED and MD characteristics [44]. ED transitions are not allowed in 4f-4f transitions. The ratio of ED to MD transitions provides information about the environment of the dopant ion within the host matrix. Asymmetric nature arises from a greater electric-dipole transition, and vice versa [45]. In the examined lattice, the magnetic-dipole transition corresponding to the $^4G_{5/2} \rightarrow {}^6H_{5/2}$ transition at a wavelength of 581 nm is more prominent than the other electric-dipole transitions, indicating the high symmetry lattice position of Sm^{3+}. Consequently, the luminescent color emitted by the Sm^{3+} ions is expected to fall within the deep orange-red spectral region.

The energy level diagram of Sm^{3+} ions in the synthesized sample is presented in Fig. 9, along with the processes for excitation, emission, energy transfer (ET), and cross-relaxation channels (CRC). The Sm^{3+} ions are gradually promoted to a higher energy state before decaying non-radiatively to a meta-stable state ($^4G_{5/2}$). Since a part of the energy is quickly released during non-radiative transitions that transfer Sm^{3+} ions in a $^4G_{5/2}$ metastable state. Through radiative decay, the Sm^{3+} ions transition from the $^4G_{5/2}$ energy level to the lower energy levels of $^6H_{5/2}$, $^6H_{7/2}$, and $^6H_{9/2}$. The reduction in intensity is attributed to resonant energy transfer (RET) interactions between the adjacent Sm^{3+} ions, which also contribute to processes like CRC and self-quenching. The probable RET channel ($^4F_{9/2} \rightarrow {}^6H_{15/2}$) and CRC are responsible for non-radiative ET mechanisms in the prepared phosphor materials and are represented below [30]:

CRC 1: $(^4G_{5/2} + {}^6H_{5/2} \rightarrow {}^6F_{5/2} + {}^6F_{11/2})$,
CRC 2: $(^4G_{5/2} + {}^6H_{5/2} \rightarrow {}^6F_{7/2} + {}^6F_{9/2})$,
CRC 3: $(^4G_{5/2} + {}^6H_{5/2} \rightarrow {}^6F_{9/2} + {}^6F_{7/2})$,
CRC 4: $(^4G_{5/2} + {}^6H_{5/2} \rightarrow {}^6F_{11/2} + {}^6F_{5/2})$

The PL spectral intensities of the peaks increase with Sm^{3+} ions concentration up to 2.0 mol% and after this, concentration quenching is observed due to the energy transfer between Sm^{3+}-Sm^{3+} ions. As the concentration of Sm^{3+} increases, the distance between the nearest Sm^{3+} ions decrease significantly. This proximity enhances interaction among neighbouring Sm^{3+} ions, leading to non-radiative energy transfer. The various interaction such as exchange contact, multipole-multipole interaction among the activator ions and radiation re-absorption etc. collectively contributes to concentration quenching.

To comprehend the nature of interaction among the Sm^{3+} ions, the critical distance (R_c) is estimated using the Blasse's equation [46]:

$$R_c \approx 2\left[\frac{3V}{4\pi X_c N}\right]^{1/3} \tag{6}$$

where $\mathbf{X_c}$ is the critical concentration of the dopant ion, **N** is the number of cations and **V** stands for the unit cell's volume. The Sm^{3+} ions doped LBW phosphor's critical distance was found to be 31.94 Å. The value of $\mathbf{R_c}$ defines the type of interaction responsible for energy transfer. If $\mathbf{R_c} \geq 5$ Å, then electric multipolar contact is responsible for energy transfer.

Huang's theory gives a relation between the luminous intensity (I) and mole fraction (x) of the activator ions as follows [47]:

$$log\left(\frac{I}{x}\right) = A - \frac{slog(x)}{d} \tag{7}$$

where I represents the emission intensity, d stands for the compound's dimensions (=3) [48] and A is independent of the dopant ions concentration. The value of parameter "s" defines the nature of interactions involving i.e., $s = 6$ for dipole–dipole (d-d), $s = 8$ for dipole-quadrupole (d-q), and $s = 10$ for quadrupole–quadrupole (q-q) interactions. The (s/d) value is calculated using the slope of the linear fitted curve between $log\left(\frac{I}{x}\right)$ and $log(x)$, as illustrated in Fig. 10. The predicted s value of 5.4, which is extremely near to 6,

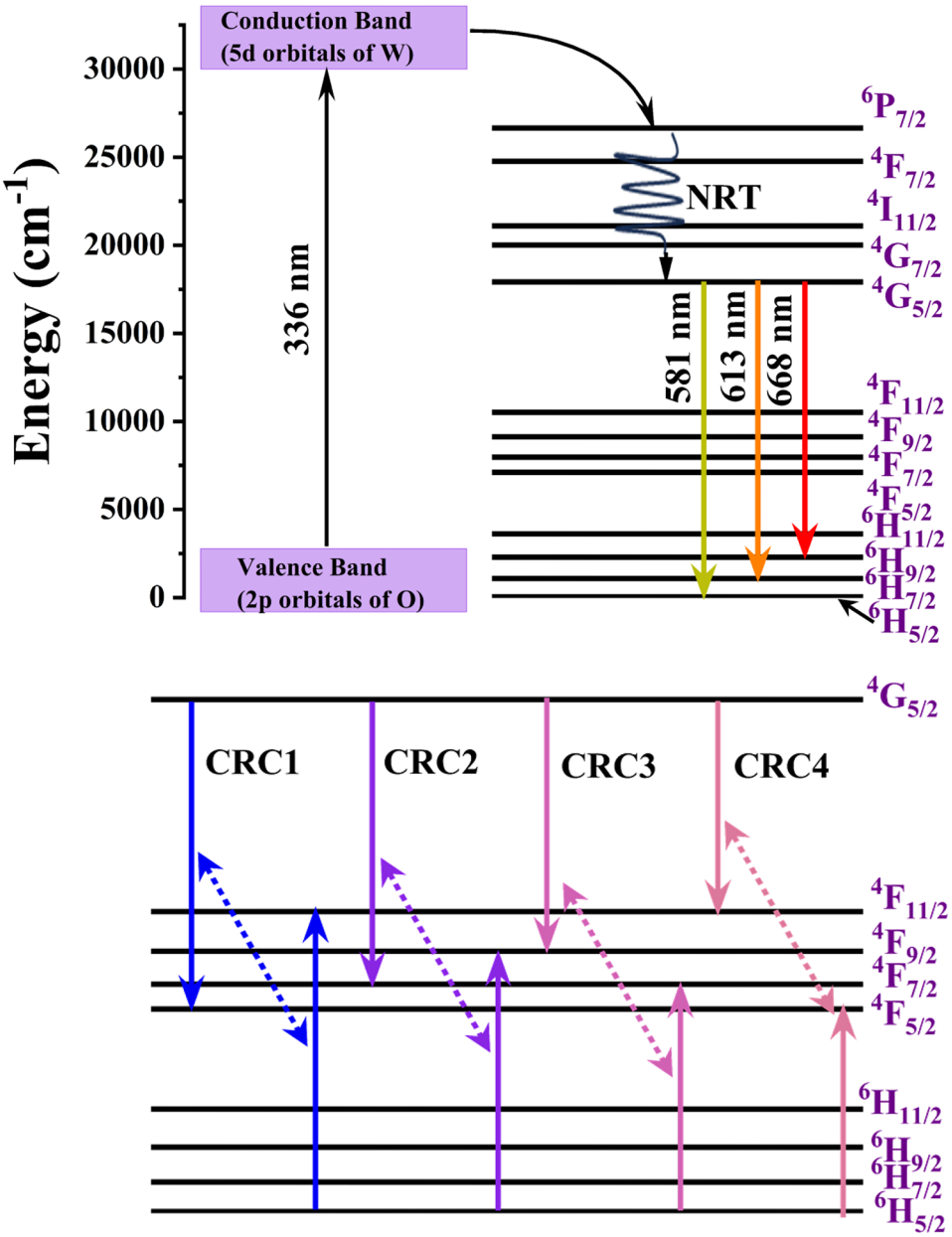


Fig. 9 The schematic energy level diagram of Sm^{3+} ions in the as-prepared LBW sample

suggests that the dipole–dipole character of the concentration quenching between the Sm^{3+} ions in LBW phosphors is the cause.

Colorimetric Analysis

Commission Internationale de l'Eclairclage CIE coordinates are commonly used to illustrate and define the color of a phosphor. The shape or location of the emission peaks are unaffected by variations in Sm^{3+} concentration, however, the PL emission intensity may alter. The CIE coordinates were evaluated using the formula given in literature [36, 49]. All the phosphor's CIE coordinates fall in the region of deep orange-red light in the CIE diagram, as illustrated in Fig. 11, making it suitable to produce w-LEDs.

A further significant measure for highlighting the role and importance of RE^{3+} doped phosphor is the correlated colour temperature (CCT). A formula for computing the CCT values is offered by McCamy shown by the following equation [50]:

$$CCT = a\eta^3 + b\eta^2 + c\eta + d \tag{8}$$

here, $a = -449$, $b = 3525$, $c = -6823.2$ and $d = 5520.3$ are the constants; and $\eta = \frac{x-x_e}{y-y_e}$; $x_e = 0.3320$, $y_e = 0.1858$.

A relatively high CCT values appear to be colder, while low CCT values suggest warmer light [51]. The CCT values for the

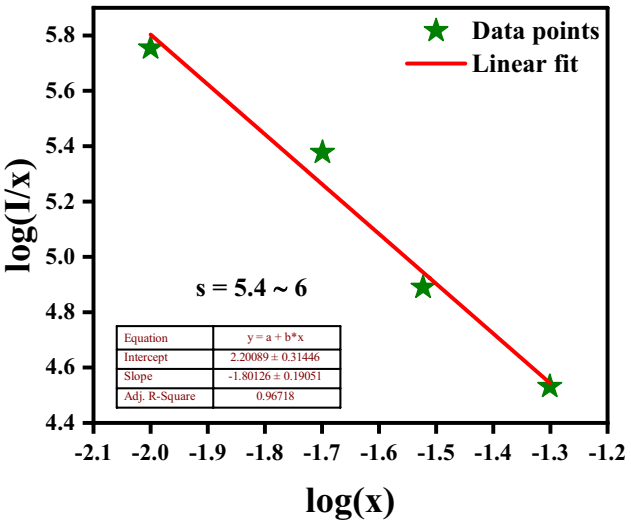


Fig. 10 The relation between log(x) and log(I/x), for Sm^{3+} ions doped LBW phosphors under $\lambda_{ex} = 336$ nm for $\lambda_{em} = 581$ nm

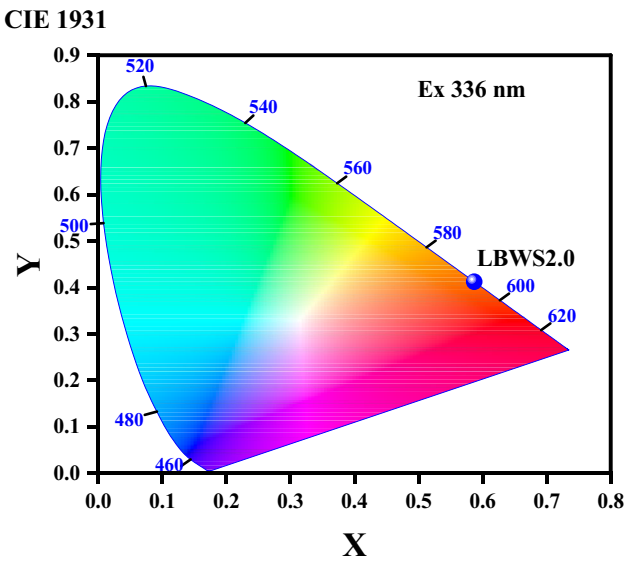


Fig. 11 CIE chromaticity coordinates for LBWS2.0 phosphor under Ex = 336 nm

as-prepared Sm^{3+} ions doped LBW phosphors were found to be less than 5000 K, indicating that it can be used to produce warm white light.

Another important element in colorimetric calculations used to examine the colour saturation is colour purity (CP). It is possible to calculate CP using the formula given below [52]:

$$CP = \frac{\sqrt{(x - x_{ee})^2 + (y - y_{ee})^2}}{\sqrt{(x_d - x_{ee})^2 + (y_d - y_{ee})^2}} \quad (9)$$

where (x_{ee}, y_{ee}) indicates an equal energy point, (x, y) is the LBW:Sm^{3+} phosphor's colour coordinates, and (x_d, y_d) are points of dominating wavelength. The colour purity for the optimized LBWS2.0 sample is found to be more than 80%. Table 3 presented the values of CIE coordinates, CCT (K) and color purity *CP* (%) for the as-prepared Sm^{3+} ions doped LBW phosphors. Therefore, the CIE color chromaticity coordinates, CCT and *CP* values demonstrate the suitability of the as-prepared Sm^{3+} ions doped LBW phosphor materials for plasma display panels and solid-state illumination as well as in n-UV or blue light excited w-LEDs.

Photoluminescence Decay Profile Analysis

The PL decay spectra for as-prepared Sm^{3+} ions doped LBW phosphors are shown in Fig. 12 under Ex = 336 nm and Em = 581 nm under normal conditions. The decay curves were analyzed to various exponential equations, but for all Sm^{3+} ion concentrations, a double exponential equation provided the best fit, signifying a substantial interaction between dopant ions and energy transfer from excited donor Sm^{3+} ions to unexcited acceptor Sm^{3+} ions. This exchange implies that the rare-earth (RE) ions within the host matrix probably exist in two separate crystallographic locations.

The following equation can explain the double exponential fitting [51]:

$$I = I_0 + A_1 * \exp\left(-\frac{t}{\tau_1}\right) + A_2 * \exp(-\frac{t}{\tau_2}) \quad (10)$$

where, I_0 is the intensity at t = 0 & I is the intensity at time t, A_1 & A_2 are the amplitudes of decay constants and τ_1 & τ_2 are the luminescence decay lifetimes.

The τ_{exp} values were evaluated using the following equation [53]:

$$\tau_{exp} = \frac{A_1\tau_1^2 + A_2\tau_2^2}{A_1\tau_1 + A_2\tau_2} \quad (11)$$

The τ_{exp} values of the as-synthesized LBW: x Sm^{3+} (x = 0.5, 1.0, 2.0, 3.0 and 5.0 mol%) phosphors were 1.73,

Table 3 CIE chromaticity coordinates, CCT (K) and color purity (CP%) Sm^{3+} ions doped LBW phosphors

System Name	CIE Coordinates		CCT (K)	Color Purity (CP %)
	X	Y		
LBWS0.5	0.5845	0.4141	1749	82
LBWS1.0	0.5848	0.4138	1748	81
LBWS2.0	0.5851	0.4136	1747	88
LBWS3.0	0.5848	0.4138	1748	83
LBWD5.0	0.5849	0.4137	1747	85

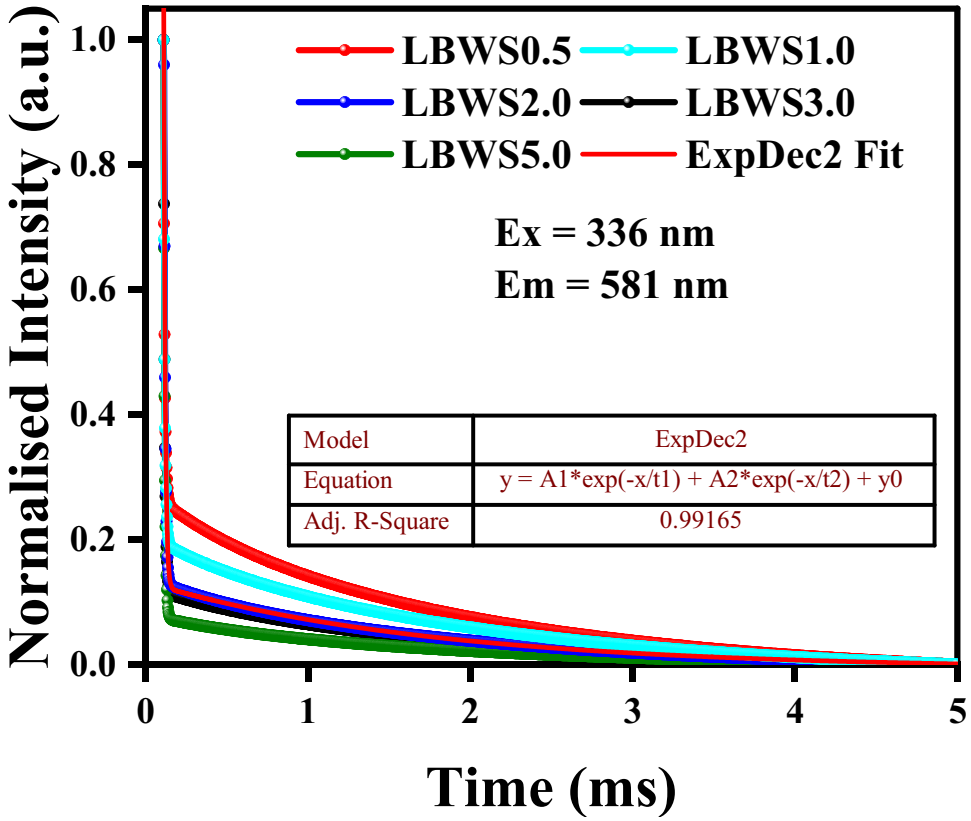


Fig. 12 PL decay curves of Sm^{3+} ions doped LBW phosphors under Ex = 336 nm and Em = 581 nm

1.63, 1.53, 1.52 and 1.24 ms, respectively. The observed τ_{exp} values exhibited a decline as the concentration of Sm^{3+} ions increased, possibly indicating an energy transfer between adjacent Sm^{3+}-Sm^{3+} ions over shorter distances.

Temperature Dependent Photoluminescence Spectral Analysis

TDPL spectra of LBWS2.0 phosphor were measured over a temperature range of 30 ºC to 170 ºC using an $\lambda_{ex} = 336$ nm to study the thermal quenching characteristics of the titled phosphor materials. Analysis of the TDPL spectra of LBWS2.0 phosphor revealed that as the temperature increased, the intensities of the photoluminescent bands decreased while maintaining their peak positions, as illustrated in Fig. 13(a).

The decrease in phosphor luminescence with increasing temperature can be explained by a non-radiative transition process. The inset in Fig. 13(a) illustrates the normalized intensity variation of the $^4G_{5/2} \rightarrow {}^6H_{5/2}$ transition concerning temperature, where the emission intensity is set as 100% at 30 ºC. The emission intensity reduces to 74% at 150 ºC and 65% at 170 ºC, indicating the relatively high thermal stability of LBWS2.0 phosphor.

Furthermore, the assessment of thermal characteristics for the optimized phosphor sample LBWS2.0 included the determination of activation energy (ΔE). The value of ΔE was calculated using the Arrhenius equation, employing the following formula [52]:

$$I_T = \frac{I_0}{1 + C\exp\left(-\frac{\Delta E}{KT}\right)} \tag{12}$$

where I_0 represents the photoluminescence (PL) intensity at 303 K (30 ºC), I_T is the PL intensity at T (in Kelvin) temperature, C is a constant and K denotes the Boltzmann constant ($= 8.617 \times 10^{-5}$ eVK^{-1}). The value of ΔE was evaluated using the slope of a linearly fitted graph that correlates $\ln\left[\left(\frac{I_o}{I_T}\right) - 1\right](along\ y - axis)$ with $\left(\frac{1}{KT}\right)(along\ x - axis)$, depicted in Fig. 13(b). The estimated ΔE value for LBWS2.0 phosphor is 0.249 eV, relatively higher than the values reported in prior research [54–56]. A relatively higher activation energy (ΔE) value indicates that the as-prepared LBWS2.0 phosphor possesses comparatively better thermal stability. The minimal variation in the CIE color chromaticity coordinates as temperature increases, illustrated in Fig. 14(a) and (b), further confirms the thermal stability of the LBWS2.0 phosphor sample. Consequently, the comprehensive TDPL analysis confirms the enhanced thermal stability of the LBWS2.0 phosphor, endorsing its suitability for the fabrication of visible photonic devices.

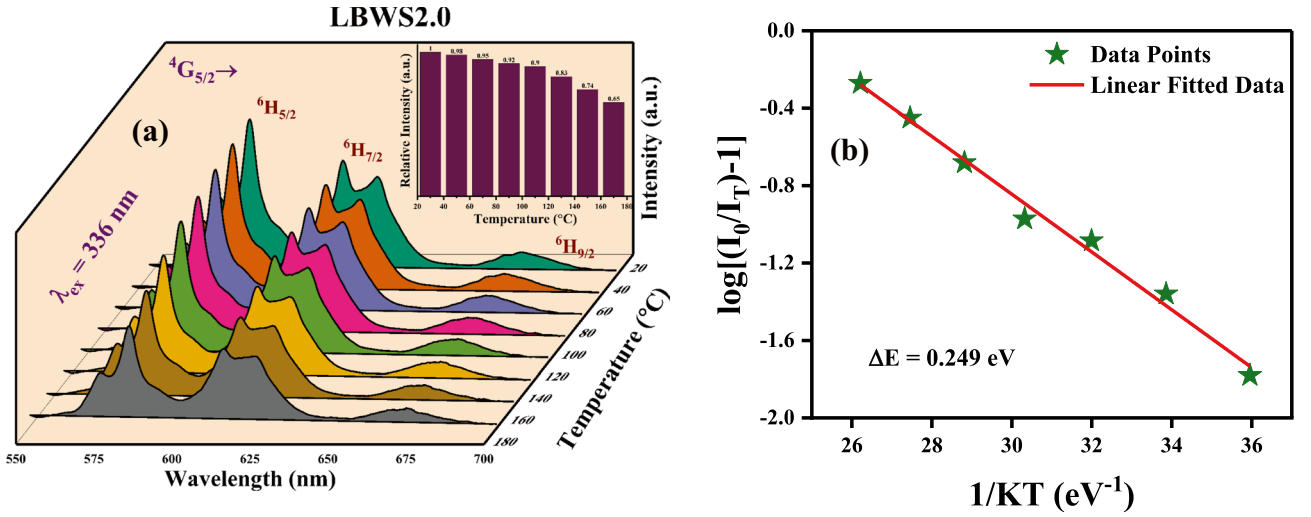


Fig. 13 **a** TDPL spectra of LBWS2.0 phosphor at 30º to 170 ºC under 336 nm UV light, (Inset shows the variation of the relative PL intensities of the emission peak at 581 nm with temperature); **b** Plot between $\log[(I_0/I_T)-1]$ versus 1/KT

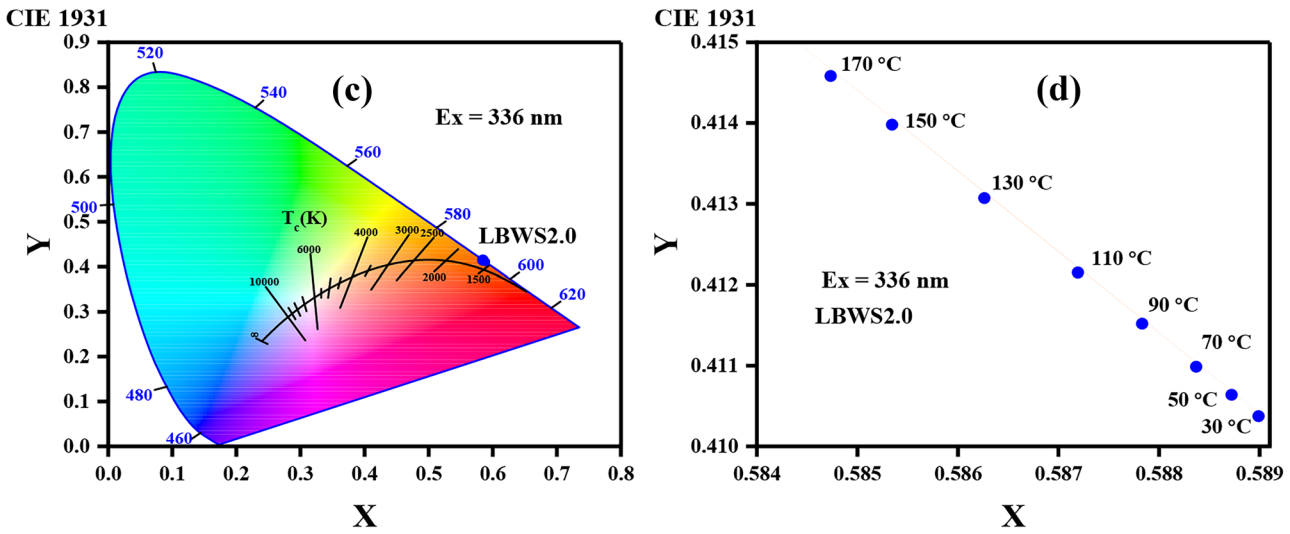


Fig. 14 **a**, **b** The variation in CIE color chromaticity coordinates with temperature

Conclusions

In brief, the present study involved the preparation of Sm^{3+} ions doped LBW phosphors using the high-temperature solid-state reaction technique. Various aspects of their structural, morphological, and photoluminescent properties were examined. The XRD analysis confirmed the crystallinity and purity of the synthesized phosphors alongwith an average crystallite size of approximately 82.87 nm. The SEM images depicted irregular polyhedron shapes with mild clustering, and the particle size lies within the micro-range. The FT-IR spectrum reveals a diverse array of vibrational and molecular bonding present within the host matrix. The optical bandgap (E_g) of 3.74–3.63 eV indicates the potential utility of the as-prepared phosphor samples in photonic devices. From PL analyses, the synthesized phosphors exhibited three emission bands, with the most prominent at 581 nm attributed to the $^4G_{5/2} \rightarrow {}^6H_{5/2}$ transition upon excitation at 336 nm. The influence of Sm^{3+} ion concentration on the PL properties of the LBW phosphor highlighted concentration quenching at a Sm^{3+} concentration of 2.0 mol%. The interaction between Sm^{3+} ions through dipole–dipole interaction is elucidated by employing the Dexter theory. The optimized sample exhibits CIE coordinates of (0.586, 0.412) at an

excitation wavelength of 336 nm, lies in a deep orange-red region. The PL decay pattern displays a double exponential fitting with decay lifetimes (τ_{exp}) of millisecond (ms) range. The thermal stability of the phosphors was confirmed through the comparative high activation energy evaluated using temperature-dependent photoluminescence (TDPL) analysis. Collectively, these investigations emphasize the superior potential of Sm^{3+} ion-doped LBW phosphors for the creation of photonic devices such as white light-emitting diodes (w-LEDs).

Acknowledgements Ms. Anu is very much thankful to CSIR/UGC for the award of a Junior Research Fellowship (JRF) (UGC Ref.1352: CSIR-UGC NET DEC. 2018). The authors are also thankful to the authorities of Delhi Technological University, New Delhi, INDIA for providing the necessary infrastructure facilities to carry out the present research work.

Authors' Contributions All authors have contributed to the study methodology and experiment. Anu: Investigation, Methodology, Formal analysis, Software, Validation, Visualization, Writing– original draft, Writing– review & editing; A.S. Rao: Resources, Visualization, Supervision, Writing- review & editing. All authors read and approved the final manuscript.

Funding None.

Availability of Data and Materials The data that support the findings of this study are not openly available due to reasons of sensitivity and are available from the corresponding author upon reasonable request.

Declarations

Ethical Approval The authors declare that no animal or animal data was used in any study or experiment mentioned in this manuscript. We approve all the ethics.

Competing Interests 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. Sumalatha C, Doddoji R, Venkateswarlu M, Rani PR, Swapna K, Mahamuda S, Rao AS (2020) White light emission from Dy^{3+}-doped ZnO + Bi_2O_3 + BaF_2 + B_2O_3 + TeO_2 glasses: Structural and spectroscopic properties. Spectrochim Acta Part A Mol Biomol Spectrosc 240:118568
2. Deopa N, Saini S, Kaur S, Prasad A, Rao AS (2019) Spectroscopic investigations on Dy^{3+} ions doped zinc lead alumino borate glasses for photonic device applications. J Rare Earths 37:52–59
3. Tang Y, He X, Zhang Y, Yuan H, Xin Y, Ren X, Chen Q, Yin H (2022) Anchoring of red perovskite nanocrystals on YAG: Ce phosphor for high color rendering index WLEDs. J Alloys Compd 899:163347
4. Kaur S, Deopa N, Prasad A, Bajaj R, Rao AS (2018) Intense green emission from Tb^{3+} ions doped zinc lead alumino borate glasses for laser and w-LEDs applications. Opt Mater (Amst) 84:318–323
5. Nair GB, Swart HC, Dhoble SJ (2020) A review on the advancements in phosphor-converted light emitting diodes (pc-LEDs): Phosphor synthesis, device fabrication and characterization. Prog Mater Sci 109:100622
6. Balaram V (2019) Rare earth elements: A review of applications, occurrence, exploration, analysis, recycling, and environmental impact. Geosci Front 10:1285–1303
7. Halmurat D, Yusufu T, Wang QL, He J, Sidike A (2019) Rare earth ion Tb^{3+} doped natural sodium feldspar (NaAlSi3O8) Luminescent properties and energy transfer. Sci Rep 9:1–8
8. Reddy KSRK, Swapna K, Mahamuda S, Venkateswarlu M, Rao AS, Prakash DV (2018) Investigation on structural and luminescence features of Dy^{3+} ions doped alkaline-earth boro tellurite glasses for optoelectronic devices. Opt Mater (Amst) 85:200–210
9. Jin X, Xie Y, Tang R, Geng X, Che J, Lou L, Chen W, Deng B, Yu R (2022) Novel double perovskite Sr_3WO_6:Sm^{3+}, Na^+ orange-red emitting phosphors with high thermal stability for white LEDs. J Alloys Compd 899:162739
10. Zhang L, Che J, Ma Y, Wang J, Kang R, Deng B, Yu R, Geng H (2020) Luminescent and thermal properties of novel orange-red emitting Ca_2MgTeO_6:Sm^{3+} phosphors for white LEDs. J Lumin 225:117374
11. Qin Z, Dong L, Zhang G, Liu Y, Zhao G, Fang Y, Hou J (2022) Hetero-valent substitution design of high thermal stability reddish-orange $Sr_3Ga_2Sn_{1.5}Si_{2.5}O_{14}$:$Sm^{3+}$ phosphor for healthy lighting white-light-emitting-diodes applications. Opt Mater (Amst) 131:112640
12. Seeta Rama Raju G, Yu JS, Park JY, Jung HC, Moon BK (2012) photoluminescence and cathodoluminescence properties of nanocrystalline $Ca_2Gd_8Si_6O_{26}$: Sm^{3+} phosphors. J Am Ceram Soc 95:238–242
13. Zhou W, Gu M, Ou Y, Zhang C, Zhang X, Zhou L, Liang H (2017) Concentration-driven selectivity of energy transfer channels and color tunability in $Ba_3La(PO_4)_3$:Tb^{3+}, Sm^{3+} for warm white LEDs. Inorg Chem 56:7433–7442
14. Devi S, Khatkar A, Taxak VB, Dalal M, Chahar S, Dalal J, Khatkar SP (2018) Optical properties of trivalent samarium-doped $Ba_5Zn_4Y_8O_{21}$ nanodiametric rods excitable by NUV light. J Alloys Compd 767:409–418
15. Ekmekçi MK, İlhan M, Başak AS, Deniz S (2015) Structural and luminescence properties of Sm^{3+} doped TTB -type $BaTa_2O_6$ ceramic phosphors. J Fluoresc 25:1757–1762
16. Devi P, Sehrawat P, Sheoran M, Dalal H, Sehrawat N, Malik RK (2023) Crystal chemistry and photoluminescent aspects of bright orange light emitting Sm^{3+}- activated $Ba_2BiV_3O_{11}$ nanomaterials for advanced illuminating applications. J Fluoresc.
17. Zhou L, Su H, Li C, Yang R, Shang Y, Wang C, Lian J, Xie S, Ji X, Yu R (2023) Synthesis and photoluminescence properties of a novel double perovskite $NaGdMgTeO_6$:Sm^{3+} red-emitting phosphor for plant growth LEDs and w-LEDs. Ceram Int 49:28246–28255
18. Yang Y, Pan H, Guan L, Wang D, Zhao J, Yang J, Yang Z, Li X (2020) Electronic structure and luminescent properties of $Sr_3Al_2O_6$:Sm^{3+} orange phosphor prepared by hydrothermal method. J Mater Res Technol 9:3847–3855
19. Li J, Liu J, Ni Q, Zhu Q, Zeng Z, Huo J, Long C, Wang Q (2022) Key role effect of samarium in realizing zero thermal quenching and achieving a moisture-resistant reddish-orange emission in $Ba_3LaNb_3O_{12}$:$Sm3^+$. Inorg Chem 61:17883–17892
20. Nair GB, Dhoble SJ (2016) Photoluminescence properties of Eu^{3+}/Sm^{3+} activated $CaZr_4(PO_4)_6$ phosphors. J Fluoresc 26:1865–1873
21. Du P, Wang L, Yu JS (2016) Luminescence properties and energy transfer behavior of single-component $NaY(WO_4)_2$:Tm^{3+}/Dy^{3+}/Eu^{3+} phosphors for ultraviolet-excited white light-emitting diodes. J Alloys Compd 673:426–432
22. Kaewnuam E, Kaewkhao J (2018) Synthesis-temperature effect on the luminescence under light and UV excitation of Eu^{3+}

doped lithium lanthanum borate phosphor. Mater Today Proc 5:15086–15091
23. Xie J, Cheng L, Tang H, Yu X, Wang Y, Wang C, Mi X, Liu Q, Zhang X (2020) Synthesis and photoluminescence properties of $NaBi(WO_4)_2$:Eu^{3+} red-emitting phosphor for N-UV-based w-LEDs. J Lumin 219:116841
24. Shi L, Han YJ, Wang HX, Shi DC, Geng XY, Zhang ZW (2019) High-efficiency and thermally stable far-red emission of Mn^{4+} in double cubic perovskite $Sr_9Y_2W_4O_{24}$ for plant cultivation. J Lumin 208:307–312
25. Zeng Q, He P, Zhang X, Hongbin L, Menglian G, Qiang S (2009) $M_9Gd_2W_4O_{24}$:Eu^{3+} (M^{2+} = Ca^{2+} and Sr^{2+}) red phosphors for NUV and blue InGaN-based WLEDs. Chem Lett 38:1172–1173
26. Qin C, Huang Y, Seo HJ (2012) The thermal stability and structural site-distribution of Eu^{3+} ions in the red-emitting phosphors $Ca_9Eu_2W_4O_{24}$ and $Sr_9Eu_2W_4O_{24}$. J Alloys Compd 534:86–92
27. Long S, Hou J, Zhang G, Huang F, Zeng Y (2013) High quantum efficiency red-emission tungstate based phosphor $Sr(La_{1-x}Eu_x)_2Mg_2W_2O_{12}$ for WLEDs application. Ceram Int 39:6013–6017
28. Zeng Q, Liang H, Gong M (2013) Preparation of tetra-tungstate red phosphors and corresponding white light-emitting diodes. Asian J Chem 25:5971–5974
29. Parauha YR, Halwar DK, Dhoble SJ (2022) Photoluminescence properties of Eu^{3+}-doped Na_2CaSiO_4 phosphor prepared by wet-chemical synthesis route. Displays 75:102304
30. Kaur S, Rao AS, Jayasimhadri M (2017) Spectroscopic and photoluminescence characteristics of Sm^{3+} doped calcium aluminozincate phosphor for applications in w-LED. Ceram Int 43:7401–7407
31. Sreeja E, Vidyadharan V, Jose SK, George A, Joseph C, Unnikrishnan NV, Biju PR (2018) A single-phase white light emitting Pr^{3+} doped Ba_2CaWO_6 phosphor: Synthesis, photoluminescence and optical properties. Opt Mater (Amst) 78:52–62
32. Chaudhary V, Kaur A (2015) Enhanced room temperature sulfur dioxide sensing behaviour of in situ polymerized polyaniline-tungsten oxide nanocomposite possessing honeycomb morphology. RSC Adv 5:73535–73544
33. Carnall WT, Fields PR, Rajnak K (1968) Electronic energy levels of the trivalent lanthanide aquo ions. IV. Eu^{3+}. J Chem Phys 49:4450–4455
34. Xia Z, Zhang Y, Molokeev MS, Atuchin VV (2013) Structural and luminescence properties of yellow-emitting $NaScSi_2O_6$:Eu^{2+} phosphors: Eu^{2+} site preference analysis and generation of red emission by codoping Mn^{2+} for white-light-emitting diode applications. J Phys Chem C 117:20847–20854
35. Taru Chanu TT, Singh NR (2020) Influence of Sm^{3+} concentration on structural and spectroscopic properties of orange-red emitting $PbWO_4$ phosphor: An energy transfer study. J Solid State Chem 284:121190
36. Anu KA, Ravina DN, Rao AS (2023) Efficient tunable photoluminescence of Dy^{3+}/Eu^{3+} co-doped OFSZBS glasses for warm white LEDs. J Non Cryst Solids 616:122421
37. Ghosh A, Kumari N, Tewari S, Bhattacharjee A (2015) Structural, electrical and optical studies on ruthenium doped ZnO pellets for device applications. Mater Sci Eng B 196:7–14
38. López R, Gómez R (2012) Band-gap energy estimation from diffuse reflectance measurements on sol–gel and commercial TiO_2: a comparative study. J Sol-Gel Sci Technol 61:1–7
39. Köferstein R, Jäger L, Ebbinghaus SG (2013) Magnetic and optical investigations on $LaFeO_3$ powders with different particle sizes and corresponding ceramics. Solid State Ionics 249–250:1–5
40. Zhou L, Liu L, Yang R, Kong J, Li Y, Zhao Y, Du Y, Li C, Zou Z, Deng B, Yu R (2023) A novel yellow-emitting $CaGdSbWO_8$:Dy^{3+} single-phase phosphor with high thermal stability and application in w-LEDs. Inorg Chem Commun 148:110316
41. Zhang A, Sun Z, Jia M, Fu Z, Choi BC, Jeong JH, Park SH (2021) Sm^{3+}-doped niobate orange-red phosphors with a double-perovskite structure for plant cultivation and temperature sensing. J Alloys Compd 889:161671
42. Wu J, Shi S, Wang X, Li J, Zong R, Chenc W (2014) Controlled synthesis and optimum luminescence of Sm^{3+}-activated nano/submicroscale ceria particles by a facile approach. J Mater Chem C 2:2786–2792
43. Nanda K, Kundu RS, Pal I, Punia R, Kishore N (2016) Concentration dependence of intensity parameters and radiative properties of Sm^{3+} ions doped in BaO-ZnO-B_2O_3 glasses. J Alloys Compd 676:521–526
44. Kumar I, Gathania AK (2022) Photoluminescence and quenching study of the Sm^{3+}-doped $LiBaPO_4$ phosphor. J Mater Sci Mater Electron 33:328–341
45. Yu R, Guo Y, Wang L, Noh HM, Moon BK, Choi BC, Jeong JH (2014) Characterizations and optical properties of orange-red emitting Sm^{3+}-doped Y_6WO_{12} phosphors. J Lumin 155:317–321
46. Hua Y, Hussain SK, Yu JS (2019) Eu^{3+}-activated double perovskite Sr_3MoO_6 phosphors with excellent color purity for high CRI WLEDs and flexible display film. Ceram Int 45:18604–18613
47. Il JY, Krishna Bharat L, Yu JS (2015) White-light emission of $Ca_2La_8(GeO_4)_6O_2$: Tb^{3+}/Sm^{3+} nanocrystalline phosphors for solid-state lighting applications. J Lumin 166:93–100
48. Du P, Huang X, Yu JS (2018) Facile synthesis of bifunctional Eu^{3+}-activated $NaBiF_4$ red-emitting nanoparticles for simultaneous white light-emitting diodes and field emission displays. Chem Eng J 337:91–100
49. Deopa N, Kaur S, Prasad A, Joshi V, Rao AS (2018) Spectral studies of Eu^{3+} doped lithium lead alumino borate glasses for visible photonic applications. Opt Laser Technol 108:434–440
50. McCamy CS (1992) Correlated color temperature as an explicit function of chromaticity coordinates. Color Res Appl 17:142–144
51. Anu DN, Rao AS (2022) Structural and luminescence characteristics of thermally stable Dy^{3+} doped oxyfluoride strontium zinc borosilicate glasses for photonic device applications. Opt Laser Technol 154:108328
52. Anu, Deopa N, Rao AS (2023) Structural and luminescent properties of dysprosium ions-doped tungstate phosphor for w-LEDs. In: Rendina I, Petti L, Sagnelli D, Nenna G (eds) Smart Materials for Opto-Electronic Applications. SPIE, p 43
53. Uma V, Vijayakumar M, Marimuthu K, Muralidharan G (2018) Luminescence and energy transfer studies on Sm^{3+}/Tb^{3+} codoped telluroborate glasses for WLED applications. J Mol Struct 1151:266–276
54. Zhu G, Li Z, Zhou F, Wang C, Xin S (2018) A novel temperature sensitive Sm^{3+} doped niobate orange-red phosphor : The synthesis and characteristic luminescent property investigation. J Lumin 196:32–35
55. Li Q, Zhang S, Lin W, Li W, Li Y, Mu Y, Wu F (2019) Spectrochimica acta part A: Molecular and biomolecular spectroscopy phosphors by energy transfer of Bi^{3+} -sensitized Eu^{3+}/ Sm^{3+}. Spectrochim Acta Part A Mol Biomol Spectrosc 117755
56. Wu L, Bai Y, Wu L, Yi H, Kong Y, Zhang Y, Xu J (2017) Sm^{3+} and Eu^{3+} codoped $SrBi_2B_2O_7$: A red-emitting phosphor with improved thermal stability. RSC Adv 7:1146–1153

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

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

Journal of Low Temperature Physics (2023) 213:251–263
https://doi.org/10.1007/s10909-023-03003-y

Thermodynamic Properties of Conical Quantum Dot Modulated by External Fields and Rashba Spin–Orbit Interaction

Yash Gupta[1] · Pranay Khosla[1] · Sakshi Arora[1] · Priyanka[1] · Rinku Sharma[1]

Received: 7 August 2023 / Accepted: 11 September 2023 / Published online: 13 October 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

In this paper, we present a thorough examination of several critical thermodynamic properties of a cylindrical cone-shaped GaAs quantum dot exposed to electric and magnetic fields, as well as the Rashba spin–orbit interaction. Mean energy, heat capacity, free energy, entropy, and magnetocaloric effect are among the thermodynamic characteristics investigated. To derive the eigenenergies of the system, we solve the one-particle time-independent Schrödinger equation. Following that, we assess the thermodynamic properties of the quantum dot and identify a relationship between these properties and various parameters such as temperature, magnetic field, and Rashba spin–orbit interaction parameter. We observe the mean energy first rises with temperature before stabilizing. Free energy diminishes while the heat capacity fluctuates before falling. The system's entropy stabilizes with high temperature with the maximum possible disorder. Different effects are produced by varying magnetic field strengths. Reduced intensity increases free energy and entropy while improving mean energy, heat capacity, and the magnetocaloric effect. Free energy, entropy, and entropy change are all enhanced by the Rashba spin–orbit interaction parameter, whereas heat capacity and mean energy are reduced. Our research aims to explain the behavior of these thermodynamic quantities in the system and provide a better understanding of the underlying physics of quantum systems.

Keywords Thermodynamic properties · Quantum dot · Rashba spin–orbit interaction · Confinement length

✉ Rinku Sharma
rinkusharma@dtu.ac.in

[1] Department of Applied Physics, Delhi Technological University, Delhi 110042, India

1 Introduction

There has been a great interest in the study of nanostructures for the past decade which has led to massive nanoscale advancement in the fields of engineering, health care, the energy sector, and the environment. Nanostructures have piqued the interest of many researchers due to their novel physical, thermodynamical, magnetic, chemical, and optical properties [1]. So naturally, there has been extensive research to explore the electronic as well as optical properties of quantum systems under multiple external factors like electron–phonon interaction, temperature, impurity, etc. [2–7]. However, the immense potential of spin-dependent orbit systems and their applications in spin electronic devices and the presence of a multitude of physical observable phenomena make the study of spin–orbit interactions (SOI) of great value and interest. SOIs are relativistic effects where the charged particle travels under the influence of an electric field and experiences a magnetic field that couples with the existing spin of the charged particle [8]. Even though there are multiple types of spin–orbit coupling, Rashba spin–orbit coupling caused by structural inversion asymmetry is more prominent while working on nanostructures [9] since it dominates over any other SOIs [10] for narrow gap semiconductors and can be precisely tuned [11] with the gate voltage which gives precise control over the final spin orientations of electrons [12] and opens it to a variety of practical applications.

To comprehend SOIs in low-dimensional structures better, numerous studies and experiments have been conducted [13]. In their paper, Chen et al. described the SOI for double ring-shaped oscillators [10], Flindt et al. presented a strategy for using spin–orbit coupling to control electron spins in nanostructures [11], and Khordad et al. investigated the impact of changing the SOI on intersubband optical coefficients and refractive index [12]. Similarly, there has been a lot of active research theoretically [14–16] as well as experimentally [17–19] on the electronic and optical properties of nanostructures. Sukirti et al. and Ibragimov, respectively, examined the thermodynamic properties of quantum dots in their studies of the impacts of Rashba SOI and magnetic field [14] and the thermodynamic properties of asymmetric parabolic quantum dots [15, 20]. Still, the impact of Rashba spin–orbit coupling under an external electric field and magnetic field on the thermodynamic properties of a quantum dot (Q-dot) has not yet been fully investigated.

Because of its unusual geometrical features, which give rise to fascinating electrical and optical properties, the conical quantum dot shape is frequently researched in theoretical and experimental research. The conical shape gives a variable aspect ratio, allowing the researchers to alter the confinement potential and customize the electron and hole energy levels. Furthermore, the conical shape provides a larger surface area, allowing for more efficient interaction between the exciton and external fields such as electric and magnetic fields. As a result, the conical quantum dot shape is a good choice for studying the effects of external forces on exciton states and their possible applications in optoelectronics and quantum information processing. And hence it is the focus of this study [21].

This paper explores the effect of Rashba SOI under external fields on thermodynamic quantities, viz. mean energy, free energy, entropy, heat capacity, and the

magnetocaloric effect of a cylindrical cone-shaped GaAs quantum dot. Moreover, we also propose an analytical solution for one particle time-independent Schrödinger equation under the effects of external magnetic and electric fields, as well as Rashba spin–orbit coupling on a cylindrical cone-shaped quantum dot made up of gallium arsenide (GaAs). In the subsequent section, a model describing Rashba spin–orbit coupling under a parabolic confined Q-dot is discussed. Furthermore, we present the numerical results and inferences of the obtained results and finally, the conclusion of the study is presented.

2 Theory and Model

A cylindrical cone-shaped GaAs quantum dot structure is considered with an external electric field.

$\vec{E} = (E, 0, 0)$ and magnetic field $\vec{B} = (0, 0, B)$. It also experiences Rashba spin–orbit coupling.

Ineffective mass framework approximation, the Hamiltonian of a single electron in such a setup can be written as [22]:

$$H = \frac{1}{2m^*}(i\hbar\vec{\nabla} + e\vec{A})^2 + eEz + V(x, y, z) + H_{\text{zeeman}} + H_{\text{Rashba}} \tag{1}$$

$$H_{\text{zeeman}} = \frac{1}{2}g^*\mu_B\vec{\sigma}\cdot\vec{B} \tag{2}$$

$$H_{\text{Rashba}} = -\alpha(\vec{\sigma}\times i\hbar\vec{\nabla}) \tag{3}$$

where $\hbar$ is the reduced Planck's constant, $\vec{\nabla}$ is the gradient operator, e is the electronic charge $\vec{A} = Bxe_y$ is the vector potential in the Landau gauge, m^* is the effective mass of an electron, g^* represents the Lande factor, μ_{B} is the Bohr magneton, $\vec{\sigma}$ is the set of Pauli spin matrices, and α is the Rashba coefficient, which controls the strength of the Rashba effect.

Here, the confinement potential $V(x, y, z)$ is

$$\begin{aligned} V(x, y, z) &= 0, \quad \text{for inside the cone} \\ &= V_0, \quad \text{for other regions} \end{aligned} \tag{4}$$

Upon expanding, the final Hamiltonian becomes:

$$\begin{aligned} H = {} & \frac{-\hbar^2}{2m^*}\nabla^2 + \frac{ie\hbar B}{2m^*}\left(y\nabla_{x-x}\nabla_y\right) + \frac{e^2B^2}{8m^*}\left(x^2 + y^2\right) \\ & + eEz + V(x, y, z) + \frac{1}{2}g^*\mu_B\vec{\sigma}_z - i\hbar\alpha(\sigma_x\nabla_x - \sigma_y\nabla_y \end{aligned} \tag{5}$$

The Schrödinger equation is given by:

$$\hat{H}\Psi(x,y,z) = E\Psi(x,y,z), \tag{6}$$

The Schrödinger Eq. (6) is solved for the given Hamiltonian (5) to derive the corresponding eigenenergies. It is solved using the COMSOL MULTIPHYSICS software. The software solves equations numerically using the finite element method. The Schrödinger equation interface has been used to obtain the eigenenergies for the system. The Hamiltonian obtained in Eq. (5) is fed into the model to obtain the eigenenergies, eigenvalues, and normalized wave functions. Dirichlet boundary condition is also applied to the wave function as it is finite, i.e., its value reduces to zero for places away from the conical dot structure.

Once we have obtained the eigenenergies through the COMSOL model, we calculate the partition function, which is essential for calculating the thermodynamic quantities of the system:

$$Q = \sum_{n=0} e^{-\beta E_n} \tag{7}$$

where β can be found using relation $\beta = 1/k_B T$ and k_B represents the Boltzmann constant and E_n are the eigenvalues found from the Eq. (6).

Using the partition function thermodynamic quantities can now be figured using these relations [23, 24]:

1. Mean energy, $U = -\frac{\partial \ln Q}{\partial \beta}$
2. Heat capacity, $C_v = \frac{\partial U}{\partial T} = k_B \beta^2 \frac{\partial^2 \ln Q}{\partial \beta^2}$
3. Free energy, $F = -k_\beta T \ln Q$
4. Entropy, $S = K_B \ln Q - k_B \beta \frac{\partial \ln Q}{\partial \beta}$
5. Magnetocaloric effect, $\Delta S = S(B \neq 0, T) - S(B = 0, T)$

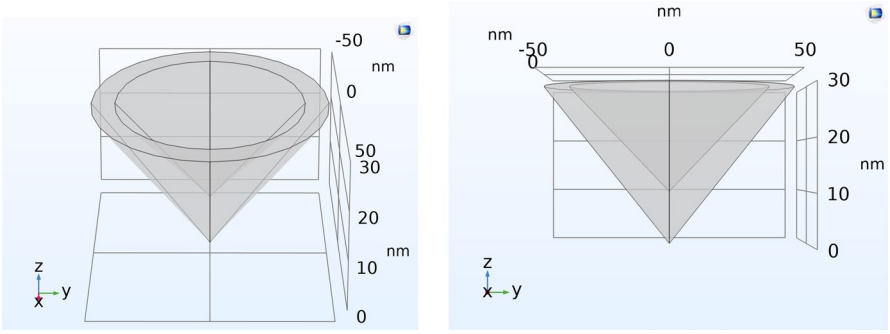


Fig. 1: 3-D representation of the conical quantum dot under consideration

3 Results and Discussion

We achieve our goal in this study for a cylindrical cone-shaped GaAs quantum dot with dimensions as shown in Fig. 1. The system is subjected to a continuous electric field of 0.6×10^{12} V/m. Confinement potential (l_o) has been defined for our calculations where $l_o = \sqrt{\frac{\hbar}{m^* w_0}}$. In the presence of these, we want to analyze thermodynamic quantities like the mean energy (in Sect. 3.1), heat capacity (in Sect. 3.2), free energy (in Sect. 3.3), entropy (in Sect. 3.4), and magnetic entropy (in Sect. 3.5).

3.1 Mean Energy

The impact of temperature and magnetic field on the mean energy of Q-dot is shown in Fig. 2. It is clear from the graph (see Fig. 2) that when the temperature rises and the magnetic field falls, the system's mean energy first increases and then stabilizes. This can be explained by stating that energy level splitting grows along the magnetic field strength which affects the distribution of energy levels and its occupation probability. Since the Q-dot is constrained and the energy levels are discrete, the mean energy of the system depends on the distribution of energy levels and its occupation probability therefore as the temperature rises and the magnetic field weakens [25] occupation probability of higher energy levels increases which leads to increase in mean energy. However, as the temperature rises more, the system's mean energy stabilizes.

Figure 3 illustrates the effect of Rashba SOI on the mean energy at a particular magnetic field and varying temperature. It is observed that at low temperatures (see Fig. 3) mean energy increases linearly with an increase in temperature but after a

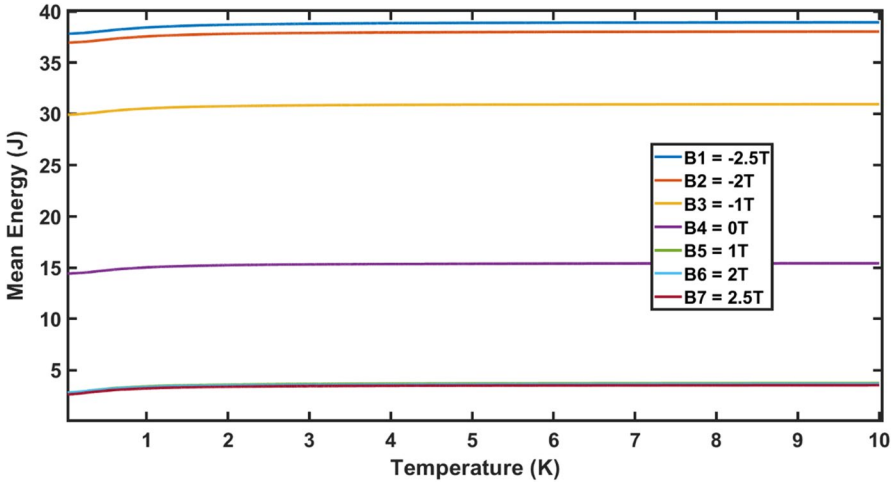


Fig. 2 Mean energy versus temperature for different magnetic field strength with electric field = 0.6 MV/m, $l_0 = 0.2$ nm and $\alpha_0 = 2.5 \times 10^{-11}$ eVm

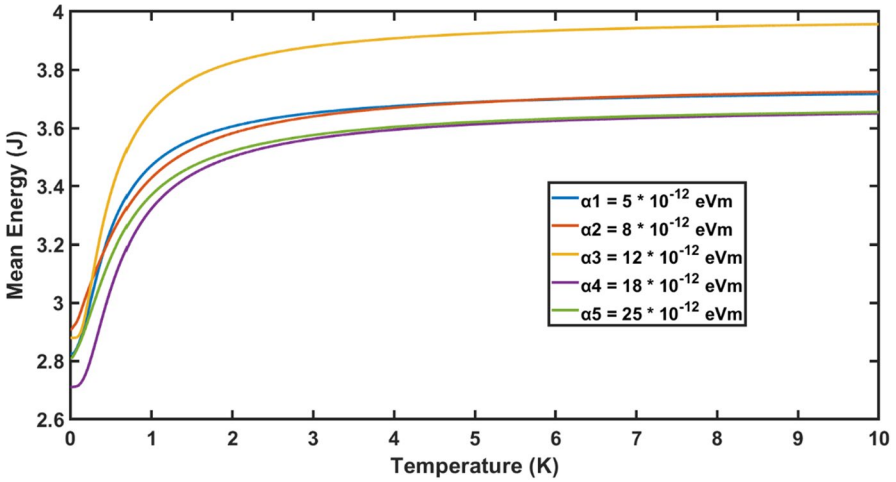


Fig. 3 Mean energy versus temperature for different Rashba parameters with electric field = 0.6 MV/m, $l_0 = 0.2$ nm, magnetic field = 2 T

point mean energy reaches a peak and becomes constant. The Rashba SOI improves energy level spacing, which decreases the probability of energy level occupation. However, the relationship between the Rashba effect and the mean energy of the electrons in the conical dot is more complex than a simple inverse correlation. While the occupation probability of the energy levels decreases as Rashba coupling increases, the mean energy of the electrons can either increase or decrease depending on the specific conditions [26].

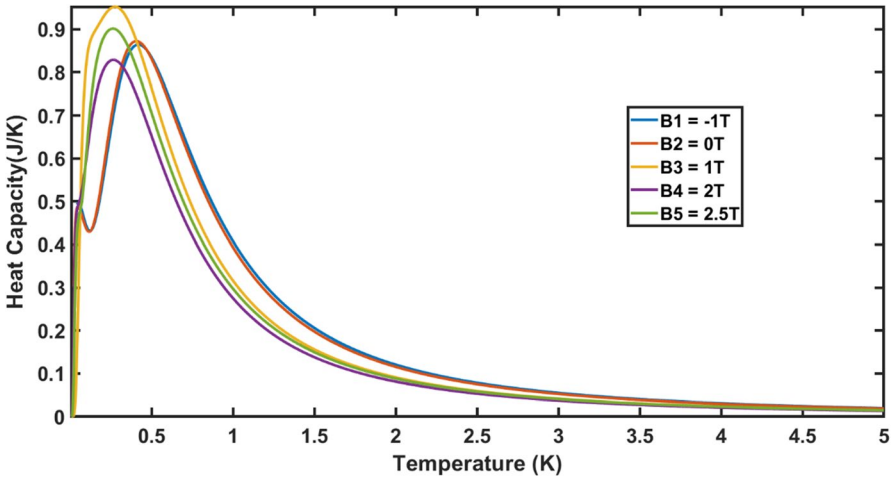


Fig. 4 Heat capacity versus temperature for different magnetic field strength with electric field = 0.6 MV/m, $l_0 = 0.2$ nm and $\alpha_0 = 2.5 \times 10^{-11}$ eVm

3.2 Heat Capacity

The impact of temperature and magnetic field on a Q-dot's heat capacity is seen in Fig. 4. The system's heat capacity diminishes with an increasing magnetic field [27]. However, the system's heat capacity initially rises exponentially with temperature but degrades at high temperatures. This kind of behavior is known as the Schottky anomaly [28]. Only the lowest energy levels are occupied at low temperatures; therefore, increasing temperature triggers a transition to higher energy levels that sharply increase the system's heat capacity. As temperature rises further, all energy levels are occupied, which causes heat capacity values to degrade [29].

The impact of Rashba SOI on the heat capacity at a constant magnetic field and changing temperature is shown in Fig. 5. It can be observed from the figure (see Fig. 5) that increasing the Rashba parameter heat capacity value reduces. The amount of thermal energy required to raise a system's temperature by one unit is known as its heat capacity. Rashba SOI's existence alters the system's potential for excitation. Because the excitation possibility increases with an increase in the Rashba parameter value, less thermal energy is required to raise the temperature by one unit, and the heat capacity value decreases [30].

3.3 Free Energy

The impact of temperature and magnetic field on a Q-dot's free energy is depicted in Fig. 6. As can be seen, the absolute value of the system's free energy grows with the increasing magnetic field, and it sharply decreases with increasing temperature (see Fig. 6).

Figure 7 illustrates the effect of Rashba SOI on the free energy of the system with varying temperature values. With increasing temperature (see Fig. 7), there is

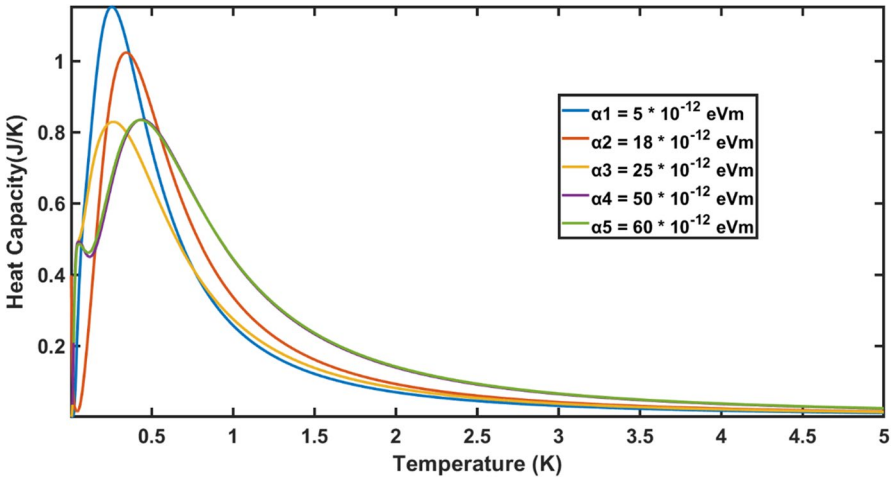


Fig. 5 Heat capacity versus temperature for different Rashba parameters with electric field = 0.6 MV/m, $l_0 = 0.2$ nm, magnetic field = $2T$

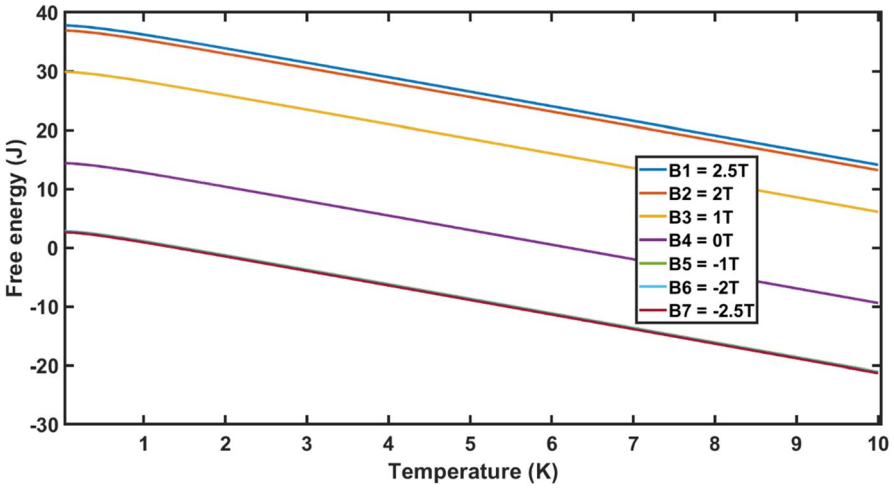


Fig. 6 Free energy versus temperature for different magnetic field strength with electric field = 0.6 MV/m, $l_0 = 0.2$ nm and $\alpha_0 = 2.5 \times 10^{-11}$ eVm

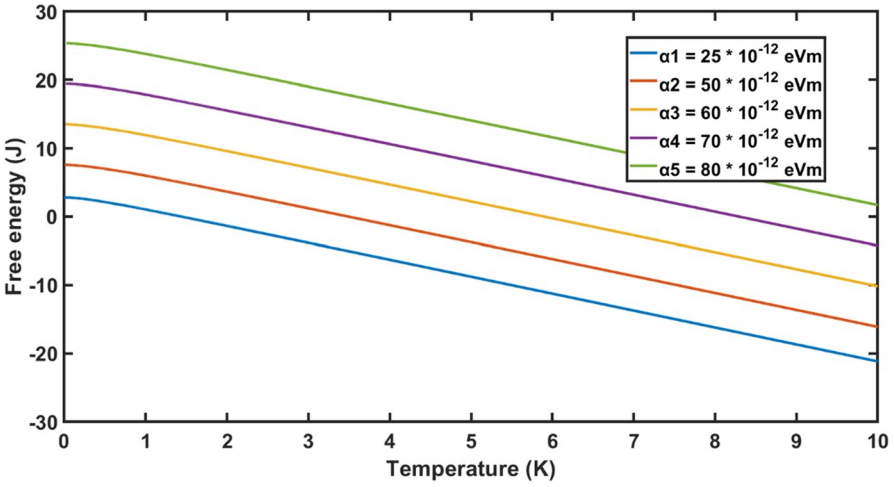


Fig. 7 Free energy versus temperature for different Rashba parameters with electric field = 0.6 MV/m, $l_0 = 0.2$ nm, magnetic field $2T$

a linear decline in the free energy of the system. It can be observed from the figure that for a particular value of temperature (see Fig. 7) the free energy of the system is higher for samples with higher Rashba parameters and this is in accordance with [26, 31].

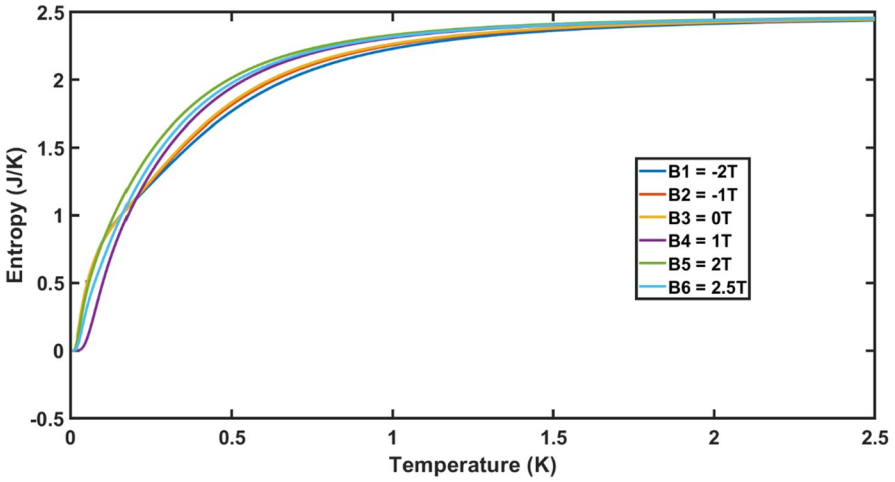


Fig. 8 Entropy versus temperature for different magnetic field strength with electric field = 0.6 MV/m, $l_0 = 0.2$ nm and $\alpha_0 = 2.5 \times 10^{-11}$ eVm

3.4 Entropy

Figure 8 illustrates the effect of magnetic field and temperature on the entropy of a Q-dot. It can be observed from the graph that by increasing the magnetic field, the entropy of the system increases. As the magnetic field increases so does the energy splitting which enhances system disorder. Similarly, on increasing temperature, entropy initially increases but soon stabilizes at the peak. At a

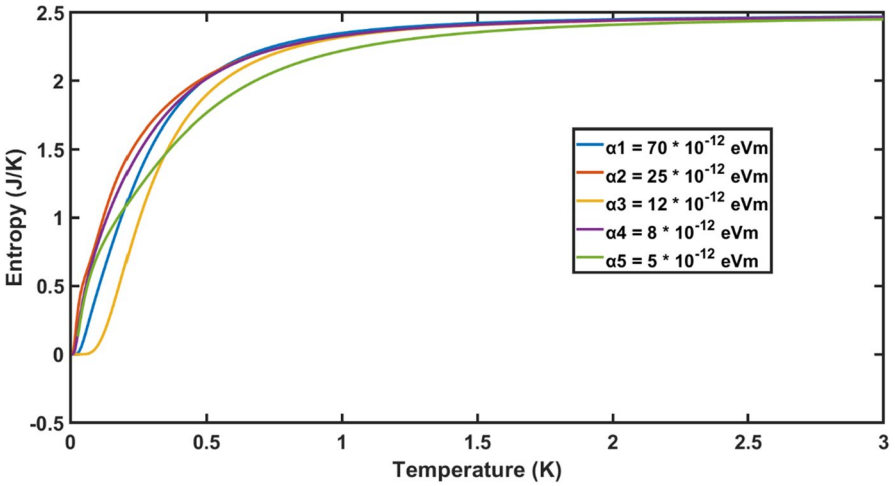


Fig. 9 Entropy versus temperature for different Rashba parameters with electric field = 0.6 MV/m, $l_0 = 0.2$ nm, magnetic field $2T$

high temperature, the system is at the maximum disorder and hence the consistently high entropy value [23, 26].

Figure 9 shows the entropy as a function of temperature with different Rashba SOI parameters. As the temperature increases, entropy also increases exponentially (see Fig. 9) and becomes stable at high temperatures (see Fig. 9). However, entropy increases with an increase in the Rashba parameter value for a particular temperature. The presence of Rashba SOI in the system increases the number of accessible energy states due to which the entropy value increases [29].

3.5 Magnetocaloric Effect

Figure 10 illustrates the effect of temperature and magnetic field on the magnetocaloric effect on a Q-dot. It can be observed from the figure that the magnetocaloric effect is largely dependent on the magnetic field than the temperature (see Fig. 10). The magnetocaloric effect is the change in entropy with respect to the magnetic field and by increasing the magnetic field the magnitude of change in entropy also amplifies (see Fig. 10) [32, 33].

Figure 11 illustrates the effect of Rashba SOI on the magnetocaloric effect on a Q-dot with increasing temperature. It can be observed from the figure that the magnitude of change in entropy is greater for samples with a higher Rashba SOI effect [28].

4 Conclusion

In this study, we obtained the energy eigenvalues of the quantum conical dot under external fields and Rashba SOI. Using them, we studied the effects of temperature, magnetic field, electric field, and Rashba SOI on the thermodynamic properties of

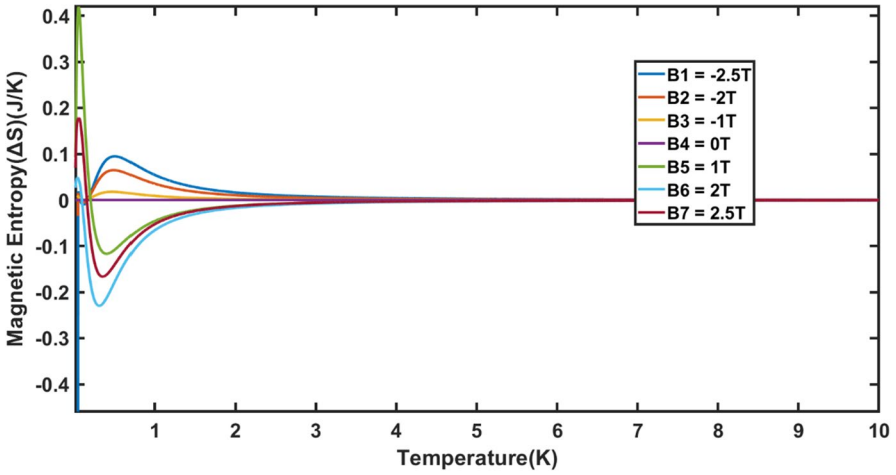


Fig. 10 Magnetocaloric effect versus temperature for different magnetic field strength with electric field = 0.6 MV/m, $l_0 = 0.2$ nm and $\alpha_0 = 2.5 \times 10^{-11}$ eVm

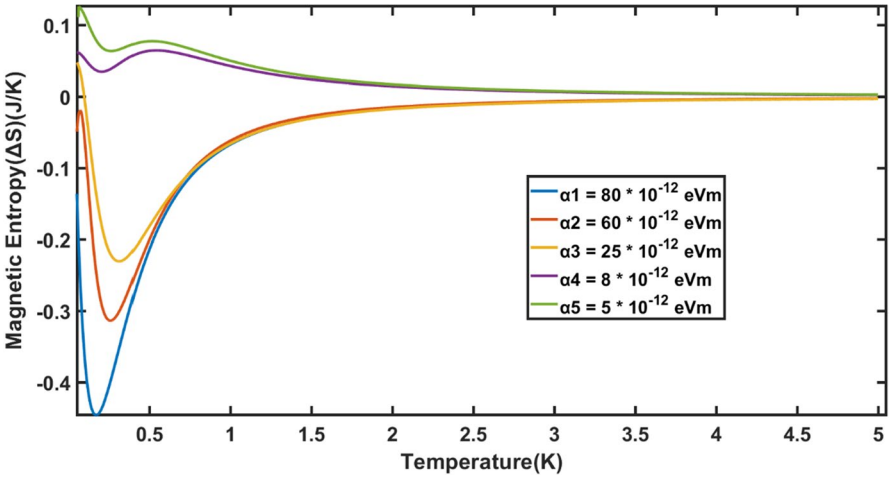


Fig. 11 Magnetocaloric effect versus temperature for different Rashba parameters with electric field = 0.6 MV/m, $l_0 = 0.2$ nm, magnetic field = 2 T

the quantum dot. It was observed that with the temperature rise, the mean energy of the system increases and gradually becomes constant, heat capacity spikes but soon reaches a peak and starts decaying whereas there is a sharp decline in free energy. The entropy of the system stabilizes with high temperature with the maximum possible disorder. Differing magnetic field impacts thermodynamical properties, i.e., reducing the magnetic field intensity increases mean energy, heat capacity, and magnitude of the magnetocaloric effect whereas reduces entropy and free energy of the system. At constant temperature, Rashba SOI plays significant roles where increasing Rashba SOI of the system increases free energy entropy and magnitude of change in entropy and decreases heat capacity and mean energy of the system.

These observations may produce some extremely intriguing findings and might be applied to determine the true nature of nanostructures in 1D system's thermodynamic properties and might lead to the development of future molecular and nanoscale electronic gadgets.

Acknowledgements Yash Gupta, Pranay Khosla, Sakshi Arora, Priyanka, and Rinku Sharma are grateful for the research facilities provided by Delhi Technological University. Priyanka also thanks the University Grants Commission for their financial assistance.

Author contributions Yash Gupta, Pranay Khosla, Sakshi Arora, Priyanka, and Rinku Sharma are grateful for the research facilities provided by Delhi Technological University. Priyanka also thanks the University Grants Commission for their financial assistance.

Declarations

Competing interests The authors declare no competing interests.

References

1. M. Nasrollahzadeh, Z. Issaabadi, M. Sajjadi, S.M. Sajadi, M. Atarod, Types of nanostructures. Interface Sci. Technol. **28**, 29 (2019)
2. J.L. Xiao, Magnetic field and anisotropic parabolic potential effects on ground state binding energy of impurity magnetopolaron in asymmetrical semi-exponential quantum wells. Phys. B Condens. Matter **639**, 413992 (2022)
3. H. Bahramiyan, R. Khordad, The effect of electron-phonon interaction on optical properties of a triangular quantum wire. Superlattices Microstruct. **63**, 267 (2013)
4. L. Van-Tan, T. Van Thang, N.D. Vy, H.T. Cao, Spin polarization and temperature dependence of electron effective mass in quantum wires. Phys. Lett. A **383**, 2110 (2019)
5. P. Hosseinpour, Effect of Gaussian impurity parameters on the valence and conduction subbands and thermodynamic quantities in a doped quantum wire. Solid State Commun. **322**, 114061 (2020)
6. R. Sharma, M. Kumar, Effects of impurity factor on the physical and transport properties for $Ga_{1-X}Al_xAs$ quantum wire in the presence of Rashba spin–orbit interaction. Phys. B Condens. Matter **629**, 413649 (2022)
7. A. Sayari, M. Servatkhah, R. Pourmand, Effects of electron–phonon interaction and pressure on optical properties of wedge-shaped quantum dots. Phys. B Condens. Matter **628**, 413631 (2022)
8. C.H.L. Quay, T.L. Hughes, J.A. Sulpizio, L.N. Pfeiffer, K.W. Baldwin, K.W. West, D. Goldhaber-Gordon, and R. de Picciotto, *Observation of a One-Dimensional Spin–Orbit Gap in a Quantum Wire* (2010)
9. M. Governale, U. Zülicke, Spin accumulation in quantum wires with strong Rashba spin–orbit coupling. Phys. Rev. B Condens. Matter Mater. Phys. **66**, 1 (2002)
10. M. Kumar, S. Lahon, P.K. Jha, M. Mohan, Energy dispersion and electron G-factor of quantum wire in external electric and magnetic fields with Rashba spin orbit interaction. Superlattices Microstruct. **57**, 11 (2013)
11. D. Liang, X.P.A. Gao, Strong tuning of Rashba spin–orbit interaction in single InAs nanowires. Nano Lett. **12**, 3263 (2012)
12. D. Grundler, Large Rashba splitting in InAs quantum wells due to electron wave function penetration into the barrier layers. Phys. Rev. Lett. **84**, 6074 (2000)
13. B. Donfack, F. Fotio, A.J. Fotue, L.C. Fai, Cumulative effects of temperature, magnetic field and spin orbit interaction (SOI) on the properties of magnetopolaron in RbCl quantum well. Chin. J. Phys. **66**, 573 (2020)
14. B. Donfack, A.J. Fotue, Effects of spin orbit interaction (SOI) on the thermodynamic properties of a quantum pseudodot. J. Low Temp. Phys. **204**, 206 (2021)
15. A. Rodríguez, *One-Dimensional Models of Disordered Quantum Wires: General Formalism* (2006)
16. C. Strunk, *Thermodynamics and the Quantum Transport of Particles and Entropy* (2012)
17. C. De Daniloff, M. Tharrault, C. Enesa, C. Salomon, F. Chevy, T. Reimann, and J. Struck, In situ thermometry of fermionic cold-atom quantum wires. Phys. Rev. Lett. **127** (2021)
18. A.S. Barnard, H. Xu, X. Li, N. Pradhan, X. Peng, Modelling the formation of high aspect CdSe quantum wires: axial-growth versus oriented-attachment mechanisms. Nanotechnology **17**, 5707 (2006)
19. I. Daruka, C. Grossauer, G. Springholz, and J. Tersoff, Equilibrium phase diagrams for the elongation of epitaxial quantum dots into hut-shaped clusters and quantum wires. Phys. Rev. B Condens. Matter Mater. Phys. **89** (2014)
20. B. Donfack, F.C.F. Mbognou, G.T. Tedondje, T.M. Cedric, A.J. Fotue, Cumulative effects of laser and spin–orbit interaction (SOI) on the thermal properties of quantum pseudo-dot. J. Low Temp. Phys. **206**, 63 (2022)
21. C. Heyn et al., Exciton states in conical quantum dots under applied electric and magnetic fields. Opt. Laser Technol. **139**, 106953 (2021)
22. R. Khordad, Optical properties of quantum wires: Rashba effect and external magnetic field. J. Lumin. **134**, 201 (2013)
23. R. Khordad, H.R.R. Sedehi, Thermodynamic properties of a double ring-shaped quantum dot at low and high temperatures. J. Low Temp. Phys. **190**, 200 (2018)
24. B. Donfack, A.J. Fotuea, and T. +237674756348, *Thermodynamic Properties of Asymmetric Semiconductor Quantum Wire under the Magnetic Eld Thermodynamic Properties of Asymmetric Semiconductor Quantum Wire under the Magnetic Field* (2022)

25. R. Khordad, H.R. Rastegar Sedehi, Low temperature behavior of thermodynamic properties of 1D quantum wire under the Rashba spin–orbit interaction and magnetic field. Solid State Commun. **269**, 118 (2018)
26. B. Boyacioglu, A. Chatterjee, Heat capacity and entropy of a GaAs quantum dot with gaussian confinement. J. Appl. Phys. **112** (2012)
27. R. Khordad, B. Mirhosseini, M.M. Mirhosseini, Thermodynamic properties of a GaAs quantum dot with an effective-parabolic potential: theory and simulation. J. Low Temp. Phys. **197**, 95 (2019)
28. J.D. Castaño-Yepes, D.A. Amor-Quiroz, C.F. Ramirez-Gutierrez, E.A. Gómez, Impact of a topological defect and rashba spin-orbit interaction on the thermo-magnetic and optical properties of a 2D semiconductor quantum dot with Gaussian confinement. Phys. E Low Dimens. Syst. Nanostruct. **109**, 59 (2019)
29. Y. Khoshbakht, R. Khordad, H.R. Rastegar Sedehi, Magnetic and thermodynamic properties of a nanowire with Rashba spin–orbit interaction. J. Low Temp. Phys. **202**, 59 (2021)
30. P. Hosseinpour, The role of Rashba spin–orbit interaction and external fields in the thermal properties of a doped quantum dot with gaussian impurity. Phys. B Condens. Matter **593**, 412259 (2020)
31. B. Donfack, A.J. Fotue, Effects of spin orbit interaction (SOI) on the thermodynamic properties of a quantum pseudodot. J. Low Temp. Phys. (2021). https://doi.org/10.1007/S10909-021-02604-9
32. H.R. Rastegar-Sedehi, *Magnetocaloric effect in Rashba spin–orbit coupling and zeeman splitting of a narrow nanowire quantum dot*, Eur. Phys. J. Plus **136** (2021)
33. H.R. Rastegar Sedehi, R. Khordad, Magnetocaloric effect magnetic susceptibility and specific heat of tuned quantum dot/ring systems. Phys. E Low Dimens. Syst. Nanostruct. **134**, 114886 (2021)

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

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

Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-15513-8

Unsupervised domain alignment of fingerprint denoising models using pseudo annotations

Indu Joshi[1,2] · Tushar Prakash[3] · Rohit Kumar[3] · Antitza Dantcheva[1] · Sumantra Dutta Roy[2] · Prem Kumar Kalra[2]

Received: 16 April 2022 / Revised: 27 September 2022 / Accepted: 19 April 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract
State-of-the-art fingerprint recognition systems perform far from satisfactory on noisy fingerprints. A fingerprint denoising algorithm is designed to eliminate noise from the input fingerprint and output a fingerprint image with improved clarity of ridges and valleys. To alleviate the unavailability of annotated data to train the fingerprint denoising model, state-of-the-art fingerprint denoising models generate synthetically distorted fingerprints and train the fingerprint denoising model on the synthetic data. However, a visible *domain shift* exists between synthetic training data and the real-world test data. Subsequently, state-of-the-art fingerprint denoising models suffer from poor generalization. To counter this drawback of state-of-the-art, this research proposes to align the synthetic and real fingerprint domains. Experiments conducted on publicly available rural Indian fingerprint demonstrate that after the proposed domain alignment, equal error rate improves from 7.30 to 6.10 on Bozorth matcher and 5.96 to 5.31 on minutiae cylinder code (MCC) matcher. Similar improved fingerprint recognition results are obtained for IIITD-MOLF and private rural fingerprints database as well.

Keywords Fingerprint enhancement · Unsupervised domain adaptation · Self-supervision · Latent fingerprints

1 Introduction

Fingerprint recognition systems are among the most popular biometrics trait. These are frequently exploited in applications including border security, access control, and law enforcement. However, the performance of fingerprint recognition systems on low quality fingerprints obtained for subjects involved in a high amount of manual labor and aged

✉ Indu Joshi
indu.joshi@inria.fr

1 Inria Sophia Antipolis, Valbonne, France

2 Indian Institute of Technology, Delhi, India

3 Delhi Technological University, Delhi, India

Published online: 03 October 2023

subjects is unsatisfactory [44, 56, 60]. Similarly, the fingerprint recognition performance on fingerprints obtained from a crime scene (latent fingerprints) is also poor. The challenge in poor quality live scan fingerprints arises due to the presence of scars, warts, dry or wet fingertips leading to unclear or thick ridges [22]. On the other hand, a latent fingerprint is characterized by smudged ridge patterns and complex backgrounds with lines, text, stains, and sometimes overlapping fingerprints (See Fig. 1(a)).

A fingerprint denoising model is designed to eliminate noise from a poor quality input fingerprint image and generate a higher quality fingerprint image with an enhanced ridge structure. Subsequently, improving the fingerprint comparison scores. State-of-the-art

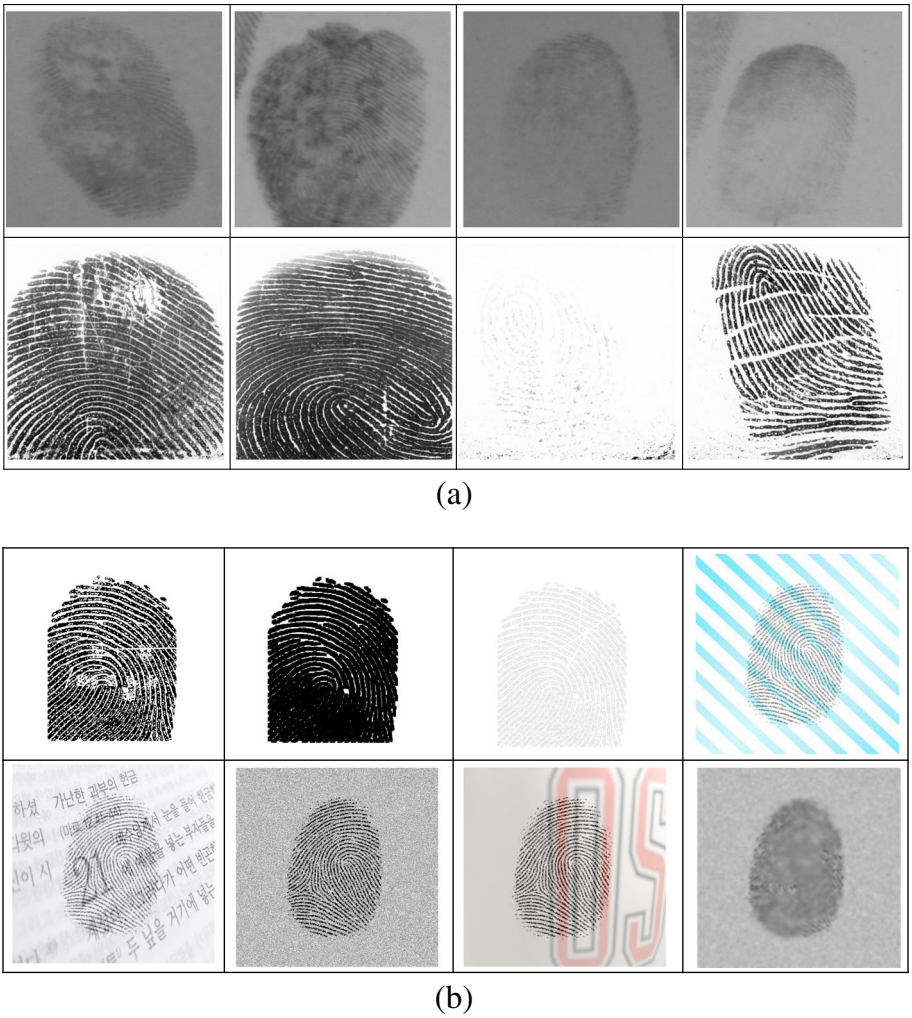


Fig. 1 (a) Sample examples of (real) fingerprints from the IIITD- MOLF Database [49] (top row), and the Rural Indian Fingerprint Database [44] (bottom row) (b) Sample training samples of synthetically distorted fingerprints. There is a visible *domain shift* between the synthetically distorted and real distorted fingerprints. Subsequently, existing fingerprint denoising models trained on the synthetic data lack generalization to real-world distorted fingerprints

fingerprint denoising models are supervised models that are trained on pixel-level annotated data of poor quality distorted and the corresponding good quality enhanced image. However, there is no publicly available annotated training set available to train a fingerprint denoising model. To address this limitation on the availability of a pixel-level annotated dataset, researchers train the model on the synthetically distorted fingerprints instead [45, 64]. The synthetically distorted fingerprints are generated by introducing different distortions such as Gaussian blurring, lines, blending with varying textures [23, 35] (See Fig. 1(b)). However, as also illustrated in Fig. 1, there is a visible *domain shift* in real low quality and synthetically distorted fingerprints. As a result, state-of-the-art fingerprint denoising models are limited in their ability to generalize to real low quality fingerprints.

1.1 Research contributions

To address this limitation of state-of-the-art fingerprint denoising models, this paper formulates fingerprint denoising as an *unsupervised domain adaptation* problem and proposes to align the source and target domains, i.e., synthetically distorted and real low quality fingerprints. In particular, we introduce two *auxiliary tasks* of predicting rotation and location for both annotated synthetically distorted fingerprints (source domain) and unannotated real poor quality fingerprints (target domain). The model jointly optimizes learning of fingerprint denoising as well the two auxiliary tasks. The introduction of these auxiliary tasks simultaneously on both the domains helps to align the representations learnt for both the domains and improve fingerprint denoising performance on the real fingerprints. The contributions of this paper are summarized as below:

- Impact of domain alignment of synthetic and real distorted fingerprints is studied. In particular, two auxiliary tasks of rotation and location prediction are introduced to bridge the domain gap between real and synthetic fingerprints.
- To the best of our knowledge, no study so far in the fingerprint denoising literature has exploited multi-task learning to align synthetic and real fingerprint domains.
- The proposed DA-GAN and the most recent fingerprint denoising models are thoroughly compared. Additionally, a comparison between the proposed DA-GAN and state-of-the-art fingerprint denoising models based on generative adversarial networks (GAN) is done.
- Ablation study illustrating the contribution of each auxiliary task in improving the fingerprint denoising performance is provided.
- A comprehensive analysis of successful and challenging cases for the proposed DA-GAN is presented.
- Improved fingerprint recognition performance on challenging latent fingerprints and fingerprints of the rural Indian population signifies improved generalization ability of the proposed DA-GAN.

1.2 Organization of paper

Given the wide applications of fingerprint recognition technology, the main contribution of this research is to enhance the quality of noisy fingerprints and subsequently improve its recognition performance. Towards this, Section 2 provides a summary of the most *recent developments* in the field. Section 3 introduces the *proposed domain alignment* technique for fingerprint denoising, while Section 4 elaborates *implementation details*. Section 5 provides

details on the *experimental protocols*. In Section 6, the effectiveness of the proposed DA-GAN is evaluated in *comparison* to the most recent *fingerprint denoising models*. Section 7 delves into providing insights on the proposed model through *ablation studies* and *discussions* on challenging cases. Finally, Section 8 *concludes* the research findings and discusses future scope.

2 Related work

2.1 Fingerprint denoising

2.1.1 Denoising using classical image processing methods

The most widely used classical image processing methods for fingerprint denoising propose to exploit contextual information in a fingerprint image either using ridge orientation or ridge continuity. These approaches typically employ *filtering* in spatial or frequency domain to predict ridge details in unclear or noisy regions of a fingerprint image [7, 15, 18, 53, 62]. Other approaches include using transformation such as wavelet decomposition [20] and discrete cosine transform [21]. Li et al. [36] propose denoising of noisy fingerprint regions using non-local Cahn-Hilliard equation. Gupta et al. [16] approximate fingerprint ridge orientation and minutiae density using higher order polynomial. Availability of reliable contextual information is a key factor to obtain satisfactory performance using these approaches. However, the extracted contextual information is often incorrect for highly noisy fingerprints. As a result, these approaches often do not generalize for heavily corrupted fingerprint regions observed in challenging real-world fingerprint samples (Table 1).

2.1.2 Denoising using learning based methods

One of the earliest learning based approaches for fingerprint denoising include *dictionary based methods* [5, 6, 10, 37, 54, 66]. These approaches generally create a dictionary of orientation fields of fingerprint ridges. Later, the constructed dictionary is exploited to approximate fingerprint ridge orientation in fingerprint regions. However, a key limitation of these approaches arises from the fact that the dictionary of fingerprint ridge orientation is created by computing fingerprint ridge orientations on good quality fingerprint regions. As a result, these orientation approximation methods do not generally work well on heavily distorted fingerprints. To address the limitations of dictionary based fingerprint denoising methods, *orientation field prediction models* are introduced [2, 46]. These models are trained to predict the orientation field for an input patch of a fingerprint image. However, in these approaches, although the orientation fields are learnt through an orientation field prediction model, however, Gabor filtering using the predicted orientations is applied to denoise input fingerprints.

Recent research direction for fingerprint denoising exploits learnable models that *directly output the denoised fingerprint image* [45, 47, 48, 51, 55] rather than predicting the orientations of fingerprint ridges and applying contextual filtering using a Gabor filter tuned at the predicted orientations. One of the early approaches in this direction includes the contributions of Schuch et al. [51]. The authors propose an autoencoder architecture (DeConvNet) for fingerprint denoising. Similar to [51], Svoboda et al. [55] also propose an autoencoder model for fingerprint denoising. However, the proposed autoencoder minimizes orientation and gradient between the output denoised fingerprints and ground truth. Horapong et al. [19]

Table 1 A table listing all the fingerprint denoising models investigated and contrasted with the proposed DA-GAN

Algorithm	Reference	Approach and Limitation
Classical Image Processing	Hong et al. [18]	Filtering using Gabor filters tuned on local ridge frequency and orientations. Correct contextual information cannot be extracted from noisy regions due to which the algorithm gives unsatisfactory performance in noisy regions.
	STFT [7]	Filtering in Fourier domain. Correct frequency information cannot be extracted from noisy regions due to which the algorithm gives unsatisfactory performance in noisy regions.
Deep learning based approaches	DeConvNet [51]	Exploits an auto-encoder network. Generates blurred fingerprints with poor ridge-valley clarity. Suffers from domain shift between training and test datasets.
	FP-E-GAN [24]	A paired image-to-image translation based generative adversarial network. Suffers from domain shift between training and test datasets.
	Svoboda et al. [55]	An auto-encoder based fingerprint denoising network that generates spurious ridge patterns in noisy regions. Suffers from domain shift between training and test datasets.
	Cycle-GAN [33]	An unpaired image-to-image translation based generative adversarial network. Fails to preserve ridge details in noisy regions. Suffers from domain shift between training and test datasets.
	DU-GAN [30]	Data uncertainty guided generative adversarial network for fingerprint denoising. Does not generalize on challenging latent fingerprints. Suffers from domain shift between training and test datasets.
	MU-GAN [27]	Explainable fingerprint denoising model that provides model uncertainty information. Does not generalize on challenging latent fingerprints. Suffers from domain shift between training and test datasets.
	DA-GAN (Proposed)	A paired image-to-image translation based generative model for fingerprint denoising. The model exploits auxiliary tasks to address the *domain shift* between training and test datasets.

suggest to identify artifacts in denoised fingerprint regions by utilizing a spectral autoencoder. Qian et al. [45] suggest a DenseUnet architecture for denoising fingerprint patches. Interested readers are referred to [52] for a comprehensive survey of fingerprint denoising methods.

Recently, GANs have emerged as state-of-the-art fingerprint denoising models. Joshi et al. suggest FP-E-GAN, a paired image to image translation GAN [23, 24] to denoise fingerprint images. Karabulut et al. [33] suggest Cycle-GAN [68], an unpaired image to image translation based solution to fingerprint denoising. Recently, model uncertainty [29], data uncertainty [30] and channel level attention [32] are introduced into GANs to design generalizable fingerprint denoising models.

To summarize, we note the drawback of state-of-the-art that all the recent fingerprint denoising approaches that directly output the denoised fingerprint images are trained on synthetically distorted fingerprints. However, there is a visible *domain shift* between the

synthetically distorted and real low quality fingerprints. As a result, existing fingerprint denoising models suffer from poor generalization on challenging real-world fingerprint images [25, 26, 28]. This research proposes to overcome the domain shift by training a fingerprint denoising model to learn auxiliary tasks.

2.2 Unsupervised domain adaptation

To save the effort in human annotations and to promote transferable learning across domains, *unsupervised domain adaptation* is introduced. Domain adaptation can be performed through a variety of techniques explained next. The earliest approaches for unsupervised domain adaptation include minimization of domain discrepancy. These approaches minimize the domain discrepancy loss between the target and source image distributions such that the model learns domain-invariant features. Deep Domain Confusion [59], deep adaptation network [39] and weighted domain adaptation network [65] are some popular unsupervised domain adaptation approaches in this category.

Some prominent studies in unsupervised domain adaptation propose to adapt the domains using a reconstruction based approach. These approaches utilize a secondary reconstruction task such that the shared representation across source and target domains can be learnt [1, 8, 12, 17, 38, 41, 50]. Another important direction of research in unsupervised domain adaptation advocates to utilize an adversarial discriminator. Recently, Joshi et al. [31] propose an adversarial learning based domain adaptation for learning sensor-invariant features for fingerprint segmentation. Adversarial learning based domain adaptation methods employ a discriminator network to discriminate between features of source and target domain. This encourages the model to learn domain invariant features [34, 40, 57, 58, 61, 67]. For a comprehensive survey on unsupervised domain adaptation, the interested readers may refer [63].

2.3 Learning with pseudo-annotations

Image recognition applications where no or only limited annotations are available, pseudo-annotations can be generated for the unannotated data. Several studies demonstrate that learning with pseudo annotations facilitates improved representation learning. Towards this, Doersch et al. [9] propose to learn context by extracting two patches and predicting the position of the second patch with respect to the first patch. Gidaris et al. [14] propose to rotate image at different angles and generate pseudo annotations. Pathak et al. [43] propose an image painting based approach to learn context in an image. Ghifary et al. [13] propose a denoising based approach to facilitate improved representation learning. Motivated by the usefulness of pseudo-annotations for improved representation learning, this paper exploits pseudo-annotations to train the fingerprint denoising model to learn two auxiliary tasks, namely rotation and patch location prediction.

3 Proposed method

3.1 Problem formulation

Let X_{src} denotes the distribution of synthetically distorted fingereprints and the corresponding pixel-level annotated denoised binarized fingerprints. X_{tgt} denotes the distributions of

unannotated real fingerprints. $X_{src} = \{(x_{src}, y_{src})\}$ and $X_{tgt} = \{x_{tgt}\}$. The objective is to learn a fingerprint model that achieves improved fingerprint denoising performance on real fingerprints (X_{tgt}) by aligning features of synthetic (X_{src}) and real (X_{tgt}) fingerprints.

3.2 Overview

We formulate fingerprint denoising as a multi-tasking learning based unsupervised alignment problem such that the fingerprint denoising model is trained to not only minimize the fingerprint denoising loss but two more additional losses corresponding to the two tasks auxiliary tasks, namely rotation and patch location prediction. The joint training of the fingerprint denoising task with two auxiliary tasks common to both the synthetically distorted (source) and real fingerprints (target) enforces domain alignment on the features learnt by a backbone fingerprint denoising model. This, in turn, facilitates better generalization on real noisy fingerprints.

3.3 Data preparation for learning auxiliary tasks

The auxiliary tasks are introduced to align the domains of synthetically distorted and real low quality fingerprints. However, as the annotations of real fingerprints are unavailable, the learning of auxiliary tasks is conducted using *pseudo-annotations* (see Fig. 2). For learning the *rotation prediction task*, the input fingerprint samples are rotated in increments of 90°, i.e. {0°, 90°, 180°, 270°}. Let the transformed source dataset for learning to predict rotation is represented by $T_{rot}(X_{src}) = \{(t_{rot}(x_{src}), y_{rot})\}$. Similarly, the transformed target dataset for learning to predict rotation is represented by $T_{rot}(X_{tgt}) = \{(t_{rot}(x_{tgt}), y_{rot})\}$.

Next, for learning the *patch location prediction task*, random patches are cropped from the input fingerprint image. The location prediction task is defined to predict the quadrant from which the patch is cropped. It is a two-dimensional regression problem in with

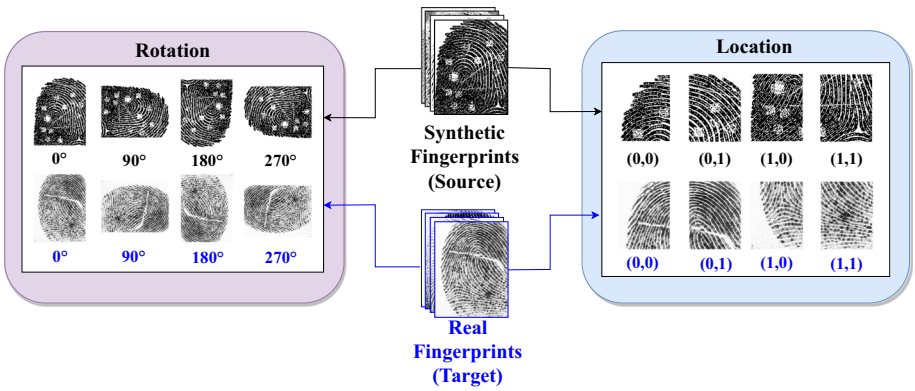


Fig. 2 Pseudo-annotations are generated for both source and target datasets for training the auxiliary tasks. To generate pseudo-annotated data for training the rotation prediction network, input fingerprint image is rotated at different angles. Likewise, to generate pseudo-annotated data for training the patch location prediction network, random patches are cropped from the input fingerprint image

pseudo-annotations defined as {(0,0), (1,0), (0,1), (1,1)}. Let the transformed source dataset for learning to predict patch location is represented by $T_{loc}(X_{src}) = \{(t_{loc}(x_{src}), y_{loc})\}$. Similarly, the transformed target dataset for learning to predict rotation is represented by $T_{loc}(X_{tgt}) = \{(t_{loc}(x_{tgt}), y_{loc})\}$.

3.4 Multi-task fingerprint denoising using auxiliary tasks

To align domains of synthetic and real fingerprints, the backbone fingerprint denoising model is modified as a multi-task learning model (see Fig. 3).

Let the model parameters of the backbone fingerprint denoising model be represented by $\{\theta, \alpha\}$, where θ represents the model parameters till the penultimate layer and α denotes the parameters of the last layer. Model parameters of the rotation and patch location prediction networks are represented by β and γ, respectively. For an input fingerprint x, the corresponding denoised fingerprint, predicted rotation and location are represented by $\psi(\theta(x); \alpha)$,

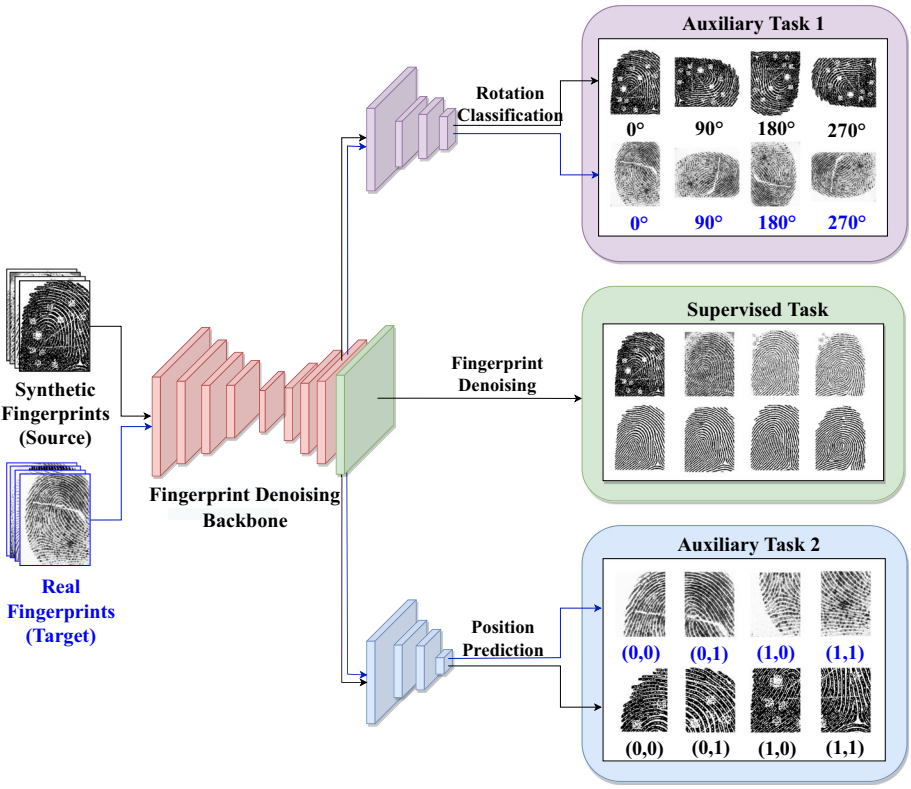


Fig. 3 Flowchart depicting proposed domain alignment using two auxiliary tasks. A multi-task learning mechanism is exploited such that fingerprint denoising is learnt using a synthetic annotated training set (source). Additionally, the model also learns to predict rotation and location on unlabelled real (target) and synthetic (source) fingerprints using pseudo-annotations. As a common fingerprint denoising backbone model is used for auxiliary tasks on both the domains, the model learns to align features of synthetic and real fingerprints

$\psi(\theta(x);\beta)$ and $\psi(\theta(x);\gamma)$ respectively. Let L_{bs} represents the denoising loss of the baseline fingerprint denoising model. Fingerprint denoising task is learnt using only the *source* dataset by minimizing $\mathcal{L}_{bs}$ (Table 2).

$$\mathcal{L}_{bs}(X_{src};\theta,\alpha) = \sum_{(x_{src},y_{src})\in X_{src}} L_{bs}(\psi(\theta(x_{src});\alpha), y_{src}) \tag{1}$$

Let L_{rot} represents the cross-entropy loss function used to train the auxiliary rotation task. Rotation task is learnt using the *transformed source and target* dataset by minimizing $\mathcal{L}_{rot}$.

$$\begin{aligned}\mathcal{L}_{rot}(X_{src}, X_{trgt};\theta,\beta) = &\sum_{(x_{rot},y_{rot})\in T_{rot}(X_{src})} L_{rot}(\psi(\theta(x_{rot});\beta), y_{rot}) \\ &+ \sum_{(x_{rot},y_{rot})\in T_{rot}(X_{tgt})} L_{rot}(\psi(\theta(x_{rot});\beta), y_{rot})\end{aligned} \tag{2}$$

Similarly, Let L_{loc} represents the regression based loss function used to train the auxiliary patch location prediction task. Location prediction task is learnt using the *transformed source*

Table 2 Table summarizing the mathematical symbols used in the paper

Notation	Description
X_{src}	Source data distribution
X_{tgt}	Target data distribution
T_{rot}	Distribution of rotation transformation
y_{rot}	Ground truth rotation label
T_{loc}	Distribution of location transformation (patch extraction)
y_{loc}	Ground truth location label
t_{rot}	A given rotation transformation
t_{loc}	A given location transformation
(x_{src}, y_{src})	An annotated sample from the source dataset
x_{tgt}	An unannotated sample from the target dataset
(x_{rot}, y_{rot})	Input x rotated at angle y_{rot}
(x_{loc}, y_{loc})	Image patch extracted from input x from pixel location y_{loc}
θ	Model parameters till the penultimate layer
α	Model parameters of the last layer of the backbone model
β	Model parameter of the rotation prediction branch
γ	Model parameter of the location prediction branch
$\psi(\theta(x);\alpha)$	Denoised fingerprint output for an input x
$\psi(\theta(x);\beta)$	Predicted rotation for an input x
$\psi(\theta(x);\gamma)$	Predicted location for an input x
γ	Model parameter of the location prediction branch
$\mathcal{L}_{bs}$	Baseline fingerprint denoising loss
$\mathcal{L}_{rot}$	Rotation prediction loss
$\mathcal{L}_{loc}$	Location prediction loss

and target dataset by minimizing $\mathcal{L}_{loc}$.

$$\begin{aligned}\mathcal{L}_{loc}(X_{src}, X_{trgt}; \theta, \gamma) = & \sum_{(x_{loc}, y_{loc}) \in T_{loc}(X_{src})} L_{loc}(\psi(\theta(x_{loc}); \gamma), y_{loc}) \\ & + \sum_{(x_{loc}, y_{loc}) \in T_{loc}(X_{tgt})} L_{loc}(\psi(\theta(x_{loc}); \gamma), y_{loc})\end{aligned} \tag{3}$$

Note that the baseline fingerprint denoising task is trained using $\mathcal{L}_{bs}(X_{src}; \theta, \alpha)$ and takes only the *annotated source dataset* as the input. On the other hand, both the auxiliary tasks of rotation and location prediction trained using $\mathcal{L}_{lrot}(X_{src}, X_{trgt}; \theta, \beta)$ and $\mathcal{L}_{loc}(X_{src}, X_{trgt}; \theta, \gamma)$ respectively take pseudo-annotated data corresponding to *both the source and target domains* as the input. Joint training of the three loss functions enforces domain alignment. The total loss minimized by the fingerprint denoising model is formalized as:

$$\arg\min [\mathcal{L}_{bs}(X_{src}; \theta, \alpha) + \mathcal{L}_{rot}(X_{src}, X_{trgt}; \theta, \beta) + \mathcal{L}_{loc}(X_{src}, X_{trgt}; \theta, \gamma)] \tag{4}$$

For an input fingerprint x, for inference during test-time, the sub-networks pertaining to the auxiliary tasks are discarded, and only the output of the fingerprint denoising task, i.e., $\psi(\theta(x); \alpha)$ is used. In this paper, as suggested in Section 3, we modify the backbone fingerprint denoising model, FP-E-GAN [24] and propose *DA-GAN* (Domain Aligned GAN).

4 Implementation details

We proceed to elaborate the network design of DA-GAN. Figure 4 illustrates different convolution blocks constituting the fingerprint denoising model as well as branches for auxiliary tasks, rotation and location prediction. Details on the convolution layer characteristics and activation functions is provided in Tables 3 and 4. The synthetic noisy and corresponding

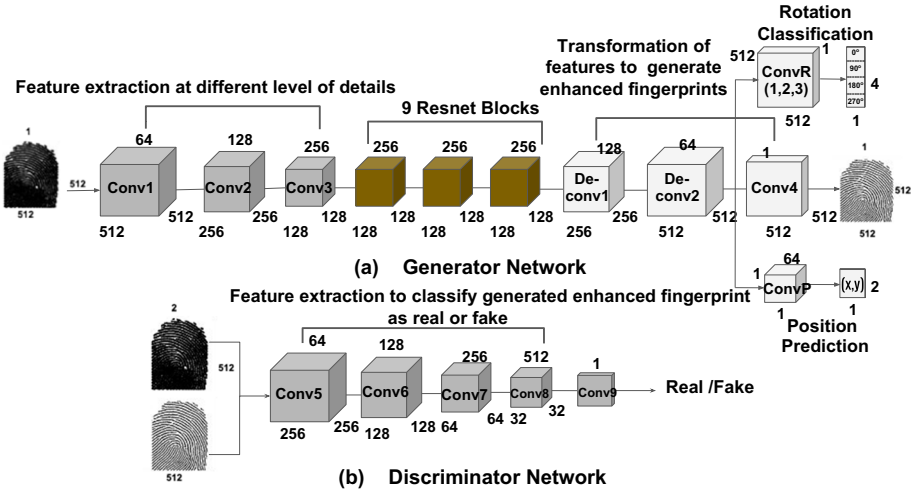


Fig. 4 Network architecture of DA-GAN

Table 3 Network architecture of the generator module of DA-GAN

Block	Kernels	Size	Stride	Padding	Layers
Conv1	64	7	1	3	Conv. Layer + Batch Norm + ReLu
Conv2	128	3	2	1	Conv. Layer + Batch Norm + ReLu + Conv. Layer + Batch Norm
Conv3	256	3	2	1	Conv. Layer + Batch Norm + ReLu + Conv. Layer + Batch Norm
ResNet Block	256	3	2	1	Conv. Layer + Batch Norm + ReLu + Conv. Layer + Batch Norm
Deconv1	128	3	2	1	Conv. Layer + Batch Norm + ReLu + Conv. Layer + Batch Norm
Deconv2	64	3	2	1	Conv. Layer + Batch Norm + ReLu + Conv. Layer + Batch Norm
Conv4	1	7	1	3	Conv. Layer + Tanh
ConvR-1	64	3	1	1	Conv. Layer + Batch Norm + ReLu
ConvR-2	128	3	1	1	Conv. Layer + Batch Norm + ReLu
ConvR-3	64	1	1	0	Conv. Layer + Batch Norm + ReLu
ConvP	64	1	1	0	Adaptive Pooling + Conv. Layer

good quality fingerprints are obtained as per guidelines provided in [23, 24]. The source code of DA-GAN is implemented in PyTorch, v1.11.0. Adam optimizer with a learning rate of 0.0002 is used to optimize the loss function of DA-GAN. Training of DA-GAN is conducted on four NVIDIA GTX 1080 Ti GPUs and a E5-2620v4 CPU. Each GPU has a 11 GB RAM.

5 Experimental evaluation

5.1 Databases

Noise in fingerprint images typically originates from structured noise in the background or poor skin quality of fingertips. To assess the fingerprint denoising performance under complex background, the fingerprint denoising performance on latent fingerprints is studied. Furthermore, to study the fingerprint denoising performance under poor skin quality co-variates, the

Table 4 Network architecture of the discriminator module of DA-GAN

Block	Kernels	Size	Stride	Padding	Layers
Conv5	64	4	2	1	Conv. Layer + LeakyReLu
Conv6	128	4	2	1	Conv. Layer + Batch Norm + LeakyReLu
Conv7	256	4	2	1	Conv. Layer + Batch Norm + LeakyReLu
Conv8	512	4	1	1	Conv. Layer + Batch Norm + LeakyReLu
Conv9	1	4	1	1	Conv. Layer

proposed fingerprint denoising model is evaluated on two poor quality fingerprints databases obtained from rural Indian subjects. Next, we share details on these databases.

1. *IIITD-MOLF* [49]: IIITD-MOLF is the largest latent fingerprint database in the public domain with 4400 latent fingerprint images. Latent fingerprints in the dataset are used as the probe set, while the fingerprints acquired through the multi-spectral sensor is used as the gallery set.
2. *Rural Indian Fingerprint database* [44]: This database contains fingerprint samples of the rural Indian population involved in excessive manual work. The volunteers for database collection comprise farmers, carpenters, and villagers. This database constitutes ten impressions per finger with a total of 1631 fingerprint images.
3. A private rural fingerprint database acquired from elderly rural Indian population and population with severely poor fingertip quality. This database constitutes two impressions per finger with a total of 1000 fingerprint images.

5.2 Evaluation metrics

1. *Fingerprint Quality Scores:* A fingerprint denoising model is introduced in a automated fingerprint recognition pipeline to improve the clarity of ridges and valleys. Therefore, one of the evaluation metrics we use in this study is to evaluate the fingerprint quality of the denoised fingerprint image. The fingerprint quality score is obtained using NBIS [42], an open source tool provided by the national institute of standards and technology (NIST).
2. *Fingerprint Matching Performance:* A fingerprint denoising model is required to generate fingerprint images that obtain higher match scores on genuine matches, thus better matching performance. Towards evaluating the improvement in matching performance, we compare cumulative matching characteristics (CMC) curves for latent fingerprints and average equal error rate (EER) for both the rural Indian fingerprint databases. Standard fingerprint matching tools, Bozorth [42] and MCC [3, 4, 11] are used to perform fingerprint matching.

6 Comparison with state-of-the-art

6.1 Performance on rural indian fingerprints

Figure 5 contrasts state-of-the-art fingerprint denoising algorithms with DA-GAN. These results demonstrate that the proposed DA-GAN is able to predict unclear and missing ridge details due to dry and wet fingertips or creases and scars. Thus, facilitating improved fingerprint matching performance. We now present a quantitative analysis of the findings from the rural Indian fingerprint database. Table 5 reports that the average fingerprint quality scores has improved from *1.31* to *1.28* after domain alignment (lower quality scores indicates better fingerprint quality scores). These results demonstrate that the proposed domain alignment framework helps to obtain better fingerprint denoising performance on challenging rural Indian fingerprints.

Next, to quantify the improvement in matching performance, Table 6 reports that the average EER has improved from *7.30* to *6.10* for Bozorth matcher and *5.96* to *5.31* for MCC matcher. Corresponding histogram of fingerprint quality scores and detection error tradeoff (DET) curves are plotted in Fig. 6(a) and (b)-(c) respectively. Similar trends are reported for

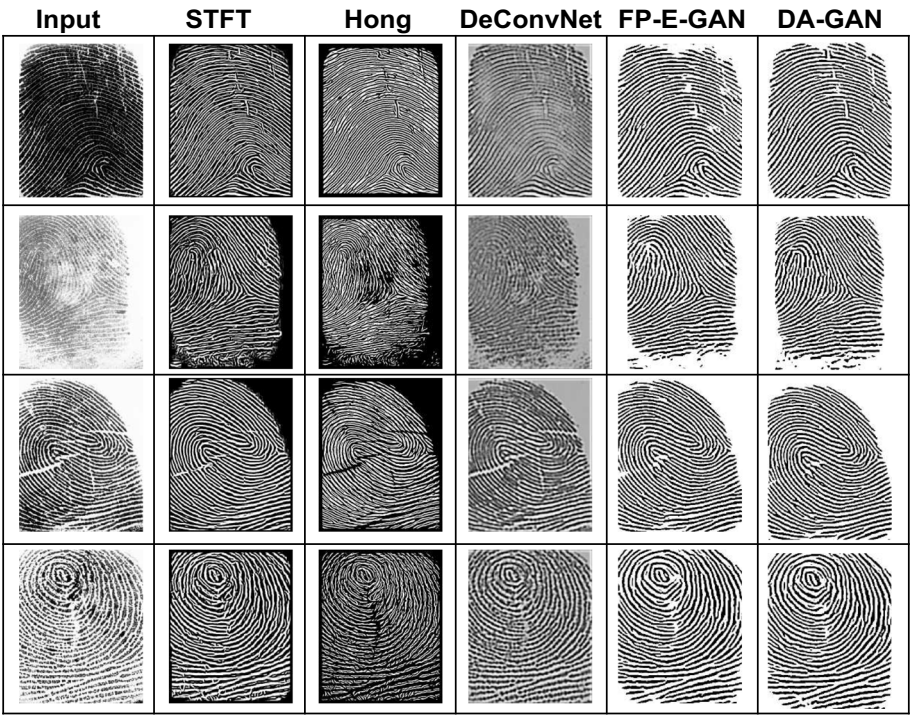


Fig. 5 Existing fingerprint denoising models perform worse than DA-GAN in predicting unclear and missing ridge details due to dry and wet fingertips or creases and scars

the private rural Indian fingerprint database in which the average EER has improved from *17.06* to *11.53* for Bozorth matcher and *15.85* to *9.70* for MCC matcher. Corresponding results are reported in Tables 7, 8 and Fig. 7.

6.2 Performance on latent fingerprints

We demonstrate that the domain alignment of fingerprint denoising models generalizes on latent fingerprints. Sample denoised fingerprints obtained by the proposed DA-GAN are compared to the state-of-the-art in Fig. 8. DA-GAN generates fewer spurious ridge patterns

Table 5 Average fingerprint quality scores obtained on the rural Indian fingerprint database [44]

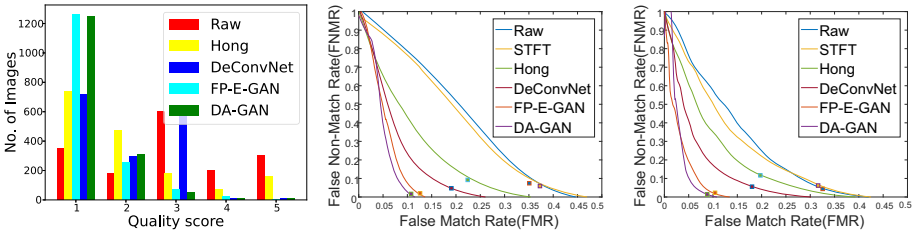
Denoising Algorithm	Avg. Quality Score (↓)
Raw Image	2.94
STFT [7]	2.86
Hong et al. [18]	2.05
DeconvNet [51]	1.95
FP-E-GAN [24]	1.31
DA-GAN	1.28

The results indicate that improved quality denoised fingerprints are obtained after introducing the proposed domain alignment framework

Table 6 Average EER obtained on the rural Indian fingerprint database [44] using Bozorth and MCC matchers

Denoising Algorithm	Bozorth (↓)	MCC (↓)
Raw Image	16.36	13.23
STFT [7]	18.13	14.52
Hong et al. [18]	11.01	11.46
DeConvNet [51]	10.93	10.86
FP-E-GAN [24]	7.30	5.96
DA-GAN	6.10	5.31

The results indicate that improved fingerprint matching performance is obtained after introducing the proposed domain alignment framework



(a) (b) (c)

Fig. 6 (a) Histogram of fingerprint quality scores and DET curves for (b) Bozorth (c) MCC. On the rural Indian fingerprint database [44], DA-GAN greatly outperforms contemporary fingerprint denoising techniques

Table 7 Average fingerprint quality scores scores obtained on the private rural fingerprint database

Denoising Algorithm	Avg. Quality Score (↓)
DeConvNet [51]	4.12
FP-E-GAN [24]	2.28
DA-GAN	1.94

in the background and unclear fingerprint image regions compared to the state-of-the-art. Next, we quantify the improvement in fingerprint denoising performance. We find that the

Table 8 Average EER obtained on the private rural Indian fingerprint database

Denoising Algorithm	Bozorth (↓)	MCC (↓)
DeConvNet [51]	28.75	26.80
FP-E-GAN [24]	17.06	15.85
DA-GAN	11.53	9.70

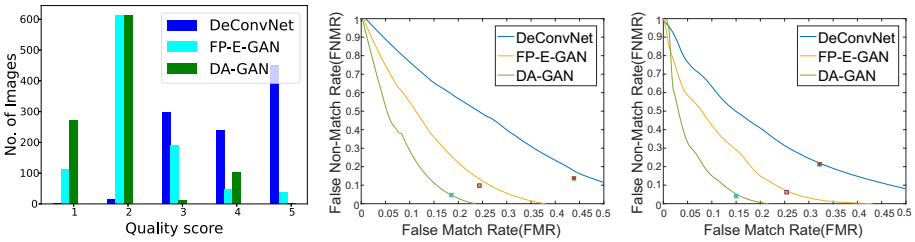


Fig. 7 (a) Histogram of fingerprint quality scores and DET curves for (b) Bozorth (c) MCC. On the private rural Indian fingerprint database, DA-GAN greatly outperforms contemporary fingerprint denoising techniques

rank-50 identification accuracy improves from *28.52* to *29.60* on the Bozorth fingerprint matcher and from *34.43* to *36.06* on the MCC matcher, after proposed domain alignment. We find that the fingerprint quality is also competitive to the baseline. Fingerprint quality scores and rank-50 accuracy are reported in Tables 9 and 10 respectively. Figure 9 shows the respective histograms of fingerprint quality scores and CMC curves. These results clearly establish the fact that domain alignment helps to improve the performance of a fingerprint denoising model.

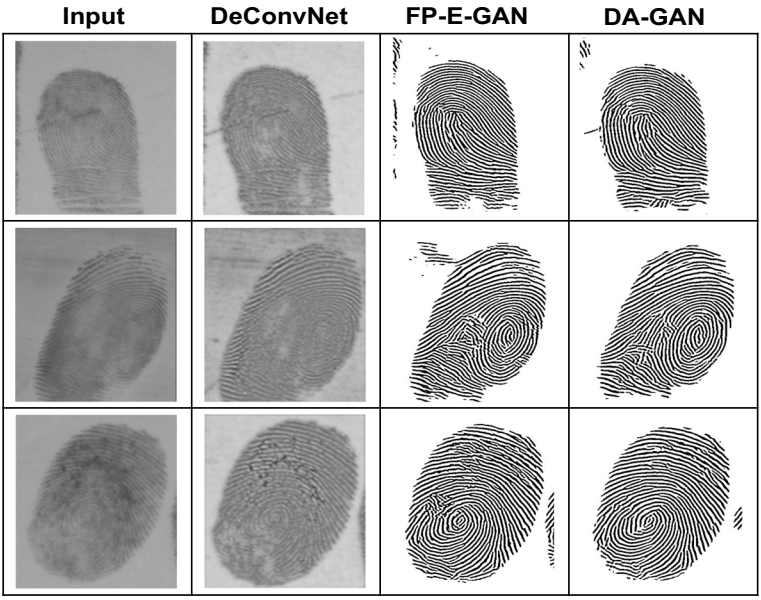


Fig. 8 Sample cases demonstrating improved fingerprint denoising ability of DA-GAN in contrast to state-of-the-art on the IIITD-MOLF database

Table 9 Average fingerprint quality scores obtained on the IIITD-MOLF database

Denoising Algorithm	Avg. Quality Score (↓)
Raw Image	4.96
DeConvNet [51]	4.09
FP-E-GAN [24]	1.91
DA-GAN	2.03

6.3 Comparison with recent GAN based fingerprint denoising models

Proposed DA-GAN significantly outperforms recently proposed generative adversarial network (GAN) based state-of-the-art fingerprint denoising models. Denoised fingerprints generated by the proposed DA-GAN obtain the highest rank-50 accuracy on both the fingerprint matchers. Although MU-GAN [27] obtains better average fingerprint quality scores as compared to the proposed DA-GAN; however, this is due to the limitation of the fingerprint quality assessment tool provided by NBIS [42]. Figure 10 compares the samples generated by the proposed DA-GAN with the state-of-the-art. These samples demonstrate that MU-GAN sometimes does not reconstruct the fingerprint ridge details on poor quality fingerprint regions, as opposed to the proposed DA-GAN. Poor ridge smoothness on such reconstructed ridges attributes to the second-best fingerprint quality scores obtained by the proposed DA-GAN. However, since the minutiae details are preserved while denoising such poor quality fingerprint regions, the proposed DA-GAN achieves the highest rank-50 accuracy (see Tables 11, 12 and Fig. 11).

6.4 Comparison of computational complexity

We begin the comparison of computational complexity of the proposed DA-GAN with state-of-the-art (hyperparameters reported in Table 13) by comparing the model capacity characterized by number of model parameters. As reported in Table 14, Cycle-GAN has the highest number of model parameters which makes it prone to overfitting and poor generalization. MU-GAN hs the least number of parameters, however the models parameters of the proposed DA-GAN is competitive to MU-GAN with far better denoising ability, which confirms the superiority of the proposed DA-GAN.

Next, we compare the computational head of the proposed DA-GAN with state-of-the-art by comparing the number of GFLOPS. As reported in Table 15, Cycle-GAN has the lowest computational overhead, however the least denoising capability. DA-GAN has the highest computational complexity. However, the computational overhead of DA-GAN is competitive

Table 10 Rank-50 accuracy obtained on the IIITD-MOLF database

Denoising Algorithm	Bozorth (↑)	MCC (↑)
Raw Image	5.45	6.06
DeConvNet [51]	14.02	14.27
Svoboda et al. [55]	NA	22.36
FP-E-GAN [24]	28.52	34.43
DA-GAN	29.61	36.06

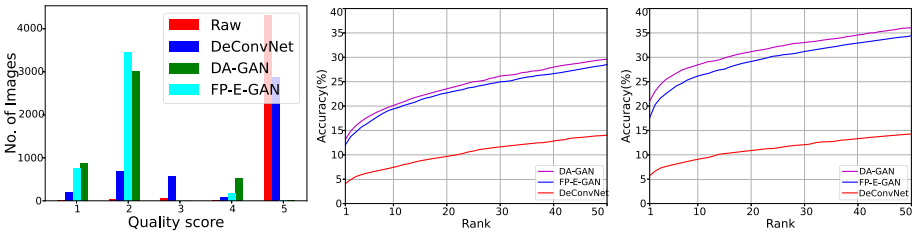


Fig. 9 Comparison of fingerprint denoising performance obtained by the proposed DA-GAN and state-of-the-art on the IIITD-MOLF database: (a) histogram of fingerprint quality scores; identification performance characterized by CMC curves using fingerprint matching tools (b) Bozorth (c) MCC

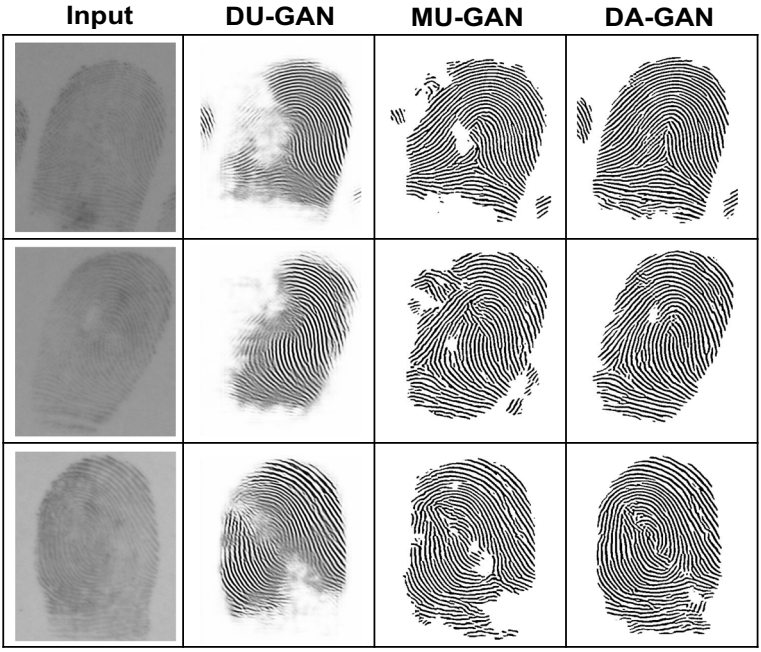


Fig. 10 Sample denoised fingerprints from the IIITD-MOLF database comparing the proposed DA-GAN with recent GAN based fingerprint denoising models

Table 11 Comparison of average fingerprint quality scores obtained by the proposed DA-GAN and recent GAN based state-of-the-art fingerprint denoising models on the IIITD-MOLF database

Denoising Algorithm	Avg. Quality Score (↓)
Cycle-GAN [33]	4.90
DU-GAN [30]	3.01
MU-GAN [27]	1.48
DA-GAN	*2.03*

Table 12 Rank-50 accuracy obtained by the proposed DA-GAN and recent GAN based state-of-the-art fingerprint denoising models on the IIITD-MOLF database

Denoising Algorithm	Bozorth (↑)	MCC (↑)
Cycle-GAN [33]	6.29	4.65
DU-GAN [30]	23.16	27.21
MU-GAN [27]	25.09	28.61
DA-GAN	29.61	36.06

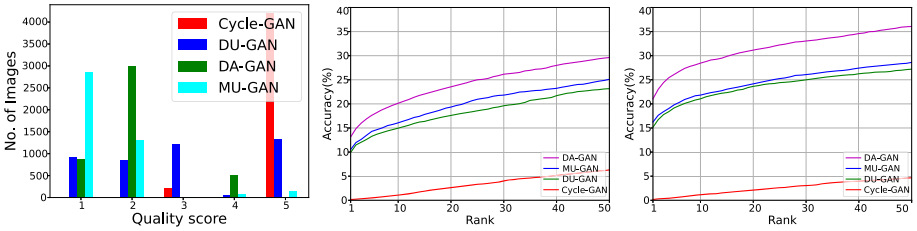


Fig. 11 Comparison of fingerprint denoising performance obtained by the proposed DA-GAN and the recent GAN based fingerprint denoising models on the IIITD-MOLF database: (a) histogram of fingerprint quality scores; identification performance characterized by CMC curves using fingerprint matching tools (b) Bozorth (c) MCC

Table 13 Hyperparameters for the proposed DA-GAN and recent GAN based state-of-the-art fingerprint denoising models

Denoising Algorithm	Learning Rate	Batch Size
Cycle-GAN [33]	0.0002	1
DU-GAN [30]	0.0002	2
MU-GAN [27]	0.0002	2
DA-GAN	0.0002	1

Table 14 Comparison of model parameters of the proposed DA-GAN and recent GAN based state-of-the-art fingerprint denoising models

Denoising Algorithm	Model Parameters (in millions)
Cycle-GAN [33]	114.32
DU-GAN [30]	14.16
MU-GAN [27]	14.14
DA-GAN	15.33

Table 15 Comparison of GFLOPS required by the proposed DA-GAN and recent GAN based state-of-the-art fingerprint denoising models

Denoising Algorithm	GFLOPS
Cycle-GAN [33]	168.26
DU-GAN [30]	237.54
MU-GAN [27]	238.28
DA-GAN	277.12

to DU-GAN and MU-GAN while providing the best denoising performance. These results demonstrate the efficacy of the proposed DA-GAN over state-of-the-art.

7 Model analysis

7.1 Ridge preservation ability

A low quality fingerprint image is characterized by the presence of structured noise in the background or blurred ridges. A fingerprint denoising model is required to eliminate noise while preserving ridge details. Figure 12 demonstrates that a high structural similarity index metric (SSIM) between the denoised fingerprint and the ground truth is obtained for DA-GAN. These results demonstrate that the proposed DA-GAN preserves ridge details of fingerprints including ridge orientation and minutiae while denoising them.

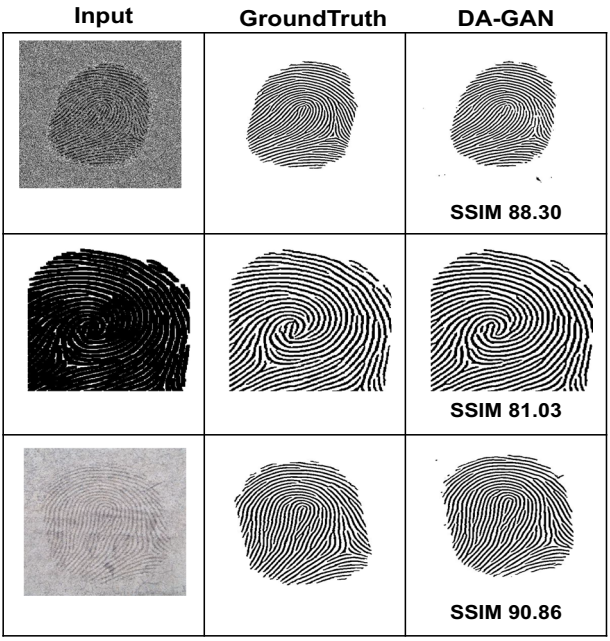


Fig. 12 Sample synthetic test cases illustrating the ridge preservation ability of DA-GAN

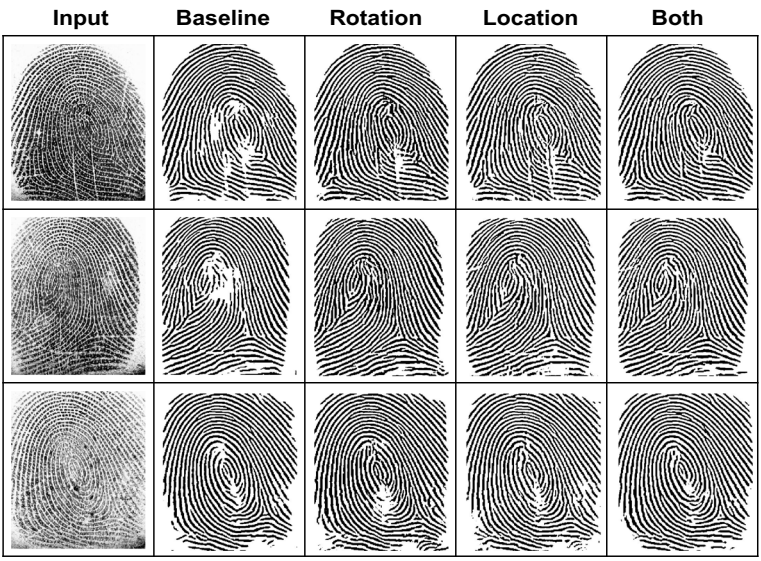


Fig. 13 Sample cases comparing the impact of each auxiliary task

7.2 Significance of each auxiliary task

Next, we investigate the individual contribution of each auxiliary task in facilitating domain alignment and, subsequently, improved fingerprint denoising performance. Samples presented in Fig. 13 demonstrate that the fingerprint denoising performance of the baseline (FP-E-GAN) fingerprint denoising model improves by introducing either of the auxiliary task, location, or rotation prediction. Location prediction turns out to be more significant compared to rotation prediction. However, the introduction of both the auxiliary tasks renders the best fingerprint denoising performance. Improved fingerprint quality scores and matching performance are obtained after the introduction of the auxiliary tasks, as reported in Tables 16 and 17. The corresponding histogram of fingerprint quality scores and DET curves are plotted in Fig. 14.

7.3 Successful cases

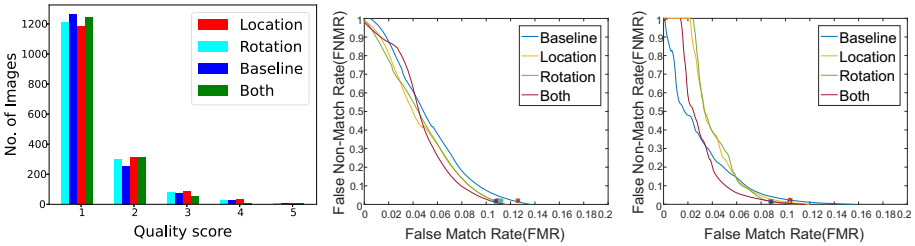
Figure 15 demonstrates successful sample cases of fingerprint denoising by DA-GAN. We observe that DA-GAN successfully predicts missing ridge details due to creases or smudged ridge patterns. Similarly, for fingerprint samples with unclear valley details due to thick

Table 16 Average fingerprint quality scores obtained on the rural Indian fingerprint database for the ablation study

Auxiliary Task	Avg. Quality Score (↓)
Baseline	1.31
Rotation	1.34
Location	1.37
Both	1.28

Table 17 Average EER obtained on the rural Indian fingerprint database for the ablation study

Auxiliary Task	Bozorth (↓)	MCC (↓)
Baseline	7.30	5.96
Rotation	6.67	5.82
Location	6.73	5.70
Both	6.10	5.31



(a) (b) (c)

Fig. 14 (a) Histogram of fingerprint quality scores and DET curves for (b) Bozorth (c) MCC. Fingerprint denoising performance is improved on the rural Indian fingerprint database [44] after introducing each auxiliary task: location and rotation prediction. As expected, best fingerprint denoising performance is obtained after introduction of both the auxiliary tasks

Fig. 15 Sample successful fingerprint denoising cases for the proposed DA-GAN

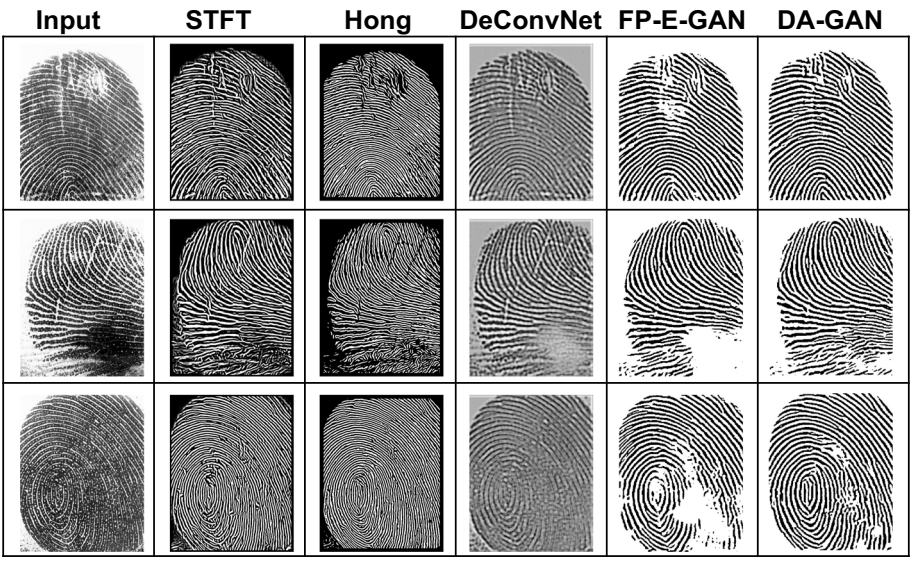


Fig. 16 Comparison of the proposed DA-GAN with state-of-the-art fingerprint denoising models on challenging samples

ridges or samples with non-uniform chemical powder, the proposed DA-GAN successfully executes fingerprint denoising. These outcomes illustrate the generalization of DA-GAN on challenging latent and rural Indian fingerprints.

7.4 Challenging cases

Figure 16 contrasts the denoised fingerprints output by DA-GAN with the ones obtained using existing fingerprint denoising models on challenging samples. The first row depicts that the proposed DA-GAN predicts spurious ridge patterns around the scarred fingerprint region. The second row depicts dark ridges in fingerprint regions with high pressure. Around high-pressure regions, DU-GAN predicts non-smooth and spurious ridges. However, in all the presented cases, denoising ability of DA-GAN is superior to state-of-the-art and the baseline FP-E-GAN. These results confirm that the proposed unsupervised domain alignment improves fingerprint denoising performance.

8 Conclusion

This research highlights the limitation in the generalization ability of state-of-the-art fingerprint denoising models due to training on synthetic data. To counter the domain shift that exists between synthetically distorted and real low quality fingerprints, this research proposes multi-task learning using two auxiliary tasks, rotation and location prediction. Results reveal that the proposed unsupervised domain alignment improves the performance of a fingerprint denoising model. However, we find that there is further scope to improve denoising performance in significantly noisy fingerprint regions (discussed in Section 7.4) by exploiting contextual information. In future, the proposed domain alignment technique can be exploited

to reduce the domain gap observed during other stages of an automated fingerprint recognition system, such as region of interest segmentation and presentation attack detection.

Acknowledgements The authors acknowledge the support of the HPC services of Inria Sophia Antipolis and IIT Delhi for the computational infrastructure. The authors sincerely thank Prof. Phalguni Gupta from IIT Kanpur and Prof. Kamlesh Tiwari from BITS Pilani for sharing the private rural Indian fingerprint used in this research. The authors acknowledge Naman Mukund from Tekie for his technical guidance.

Funding This work is partly supported by the French government, the National Research Agency (ANR) under grant ANR-18-CE92-0024, project RESPECT.

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

Declarations

Competing Interests A. Dantcheva, one of the co-authors of this manuscript is a member of the editorial board of this journal.

References

1. Bousmalis K, Silberman N, Dohan D, Erhan D, Krishnan D (2017) Unsupervised pixel-level domain adaptation with generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp 3722–3731
2. Cao K, Jain AK (2015) Latent orientation field estimation via convolutional neural network. In: Proceedings of the International Conference on Biometrics (ICB), pp 349–356
3. Cappelli R, Ferrara M, Maltoni D (2010) Fingerprint Indexing Based on Minutia Cylinder-Code. IEEE Trans Pattern Anal Mach Intell 33(5):1051–1057
4. Cappelli R, Ferrara M, Maltoni D (2010) Minutia Cylinder-Code: A New Representation and Matching Technique for Fingerprint Recognition. IEEE Trans Pattern Anal Mach Intell 32(12):2128–2141
5. Chaidee W, Horapong K, Areekul V (2018) Filter design based on spectral dictionary for latent fingerprint pre-enhancement. In: Proceedings of the International Conference on Biometrics (ICB), pp 23–30
6. Chen C, Feng J, Zhou J (2016) Multi-scale dictionaries based fingerprint orientation field estimation. In: Proceedings of the International Conference on Biometrics (ICB), pp 1–8
7. Chikkerur S, Cartwright AN, Govindaraju V (2007) Fingerprint enhancement using STFT analysis. Pattern Recogn 40(1):198–211
8. Choi Y, Choi M, Kim M, Ha JW, Kim S, Choo J (2018) Stargan: unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8789–8797
9. Doersch C, Gupta A, Efros AA (2015) Unsupervised visual representation learning by context prediction. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp 1422–1430
10. Feng J, Zhou J, Jain AK (2013) Orientation Field Estimation for Latent Fingerprint Enhancement. IEEE Trans Pattern Anal Mach Intell 35(4):925–940
11. Ferrara M, Maltoni D, Cappelli R (2012) Noninvertible Minutia Cylinder-Code Representation. IEEE Trans Inform Forensics Secur 7(6):1727–1737
12. Ghifary M, Kleijn WB, Zhang M, Balduzzi D (2015) Domain generalization for object recognition with multi-task autoencoders. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp 2551–2559
13. Ghifary M, Kleijn WB, Zhang M, Balduzzi D, Li W (2016) Deep reconstruction-classification networks for unsupervised domain adaptation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 597–613
14. Gidaris S, Singh P, Komodakis N (2018) Unsupervised representation learning by predicting image rotations. In: Proceedings of the International Conference on Learning Representations (ICLR)
15. Gottschlich C (2011) Curved-Region-Based Ridge Frequency Estimation and Curved Gabor Filters for Fingerprint Image Enhancement. IEEE Trans Image Process 21(4):2220–2227

16. Gupta R, Khari M, Gupta D, Crespo RG (2020) Fingerprint Image Enhancement and Reconstruction using the Orientation and Phase Reconstruction. Inform Sci 530:201–218
17. Hoffman J, Tzeng E, Park T, Zhu JY, Isola P, Saenko K, Efros A, Darrell T (2018) Cycada: cycle-consistent adversarial domain adaptation. In: Proceedings of the International Conference on Machine Learning (ICML), pp 1989–1998
18. Hong L, Wan Y, Jain A (1998) Fingerprint Image Enhancement: Algorithm and Performance Evaluation. IEEE Trans Pattern Anal Mach Intell 20(8):777–789
19. Horapong K, Srisutheenon K, Areekul V (2021) Progressive and Corrective Feedback for Latent Fingerprint Enhancement using Boosted Spectral Filtering and Spectral Autoencoder. IEEE Access 9:96288–96308
20. Hsieh CT, Lai E, Wang YC (2003) An Effective Algorithm for Fingerprint Image Enhancement Based on Wavelet Transform. Pattern Recogn 36(2):303–312
21. Jirachaweng S, Areekul V (2007) Fingerprint enhancement based on discrete cosine transform. In: Proceedings of the International Conference on Biometrics (ICB), pp 96–105
22. Joshi, I.: Advanced Deep Learning Techniques for Fingerprint Preprocessing. Ph.D. thesis, IIT Delhi (2021)
23. Joshi I, Anand A, Dutta Roy S, Kalra PK (2021) On training generative adversarial network for enhancement of latent fingerprints. In: AI and deep learning in biometric security, pp 51–79
24. Joshi I, Anand A, Vatsa M, Singh R, Dutta Roy S, Kalra P (2019) Latent fingerprint enhancement using generative adversarial networks. In: IEEE Winter Conference on Applications of Computer Vision (WACV), pp 895–903
25. Joshi I, Dhamija T, Kumar R, Dantcheva A, Dutta Roy S, Kalra PK (2022) Cross-domain consistent fingerprint denoising. IEEE Sensors Letters
26. Joshi I, Grimmer M, Rathgeb C, Busch C, Bremond F, Dantcheva A (2022) Synthetic data in human analysis: a survey. arXiv:2208.09191
27. Joshi I, Kothari R, Utkarsh A, Kurmi VK, Dantcheva A, Dutta Roy S, Kalra PK (2021) Explainable fingerprint ROI segmentation using Monte Carlo dropout. In: IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW), pp 60–69
28. Joshi I, Prakash T, Jaiswal B, Kumar R, Dantcheva A, Dutta Roy S, Kalra PK (2022) Context-aware restoration of noisy fingerprints. IEEE Sensors Letters
29. Joshi I, Utkarsh A, Kothari R, Kurmi VK, Dantcheva A, Dutta Roy S, Kalra PK (2021 (accepted)) On estimating uncertainty of fingerprint enhancement models. In: Digital image enhancement and reconstruction ((accepted))
30. Joshi I, Utkarsh A, Kothari R, Kurmi VK, Dantcheva A, Dutta Roy S, Kalra PK (2021) Data uncertainty guided noise-aware preprocessing of fingerprints. In: International Joint Conference on Neural Networks (IJCNN), pp 1–8
31. Joshi I, Utkarsh A, Kothari R, Kurmi VK, Dantcheva A, Dutta Roy S, Kalra PK (2021) Sensor-invariant fingerprint ROI segmentation using recurrent adversarial Learning. In: International Joint Conference on Neural Networks (IJCNN), pp 1–8
32. Joshi I, Utkarsh A, Singh P, Dantcheva A, Dutta Roy S, Kalra PK (2022) On restoration of degraded fingerprints. Multimed Tool Appl 1–29
33. Karabulut D, Tertychnyi P, Arslan HS, Ozcinar C, Nasrollahi K, Valls J, Vilaseca J, Moeslund TB, Anbarjafari G (2020) Cycle-Consistent Generative Adversarial Neural Networks based Low Quality Fingerprint Enhancement. Multimed Tool Appl 79(25):18569–18589
34. Li H, Pan SJ, Wang S, Kot AC (2018) Domain generalization with adversarial feature learning. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp 5400–5409
35. Li J, Feng J, Kuo CCJ (2018) Deep Convolutional Neural Network for Latent Fingerprint Enhancement. Signal Process: Image Commun 60:52–63
36. Li Y, Xia Q, Lee C, Kim S, Kim J (2022) A robust and efficient fingerprint image restoration method based on a phase-field model. Pattern Recogn 123:108405
37. Liu S, Liu M, Yang Z (2017) Sparse Coding Based Orientation Estimation for Latent Fingerprints. Pattern Recogn 67:164–176
38. Liu YC, Yeh YY, Fu TC, Wang SD, Chiu WC, Wang YCF (2018) Detach and adapt: learning cross-domain disentangled deep representation. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp 8867–8876
39. Long M, Cao Y, Wang J, Jordan M (2015) Learning transferable features with deep adaptation networks. In: Proceedings of the International Conference on Machine Learning (ICML), pp 97–105
40. Long M, Cao Z, Wang J, Jordan MI (2017) Conditional adversarial domain adaptation pp 1647–1657

41. Murez Z, Kolouri S, Kriegman D, Ramamoorthi R, Kim K (2018) Image to image translation for domain adaptation. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp 4500–4509
42. NIST: NBIS- NIST Biometric Image Software. http://biometrics.idealtest.org/
43. Pathak D, Krahenbuhl P, Donahue J, Darrell T, Efros AA (2016) Context encoders: feature learning by inpainting. In: Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp 2536–2544
44. Puri C, Narang K, Tiwari A, Vatsa M, Singh R (2010) On analysis of rural and urban Indian fingerprint images. In: Proceedings of the international conference on ethics and policy of biometrics, pp 55–61
45. Qian P, Li A, Liu M (2019) Latent fingerprint enhancement based on denseUNet. In: Proceedings of the International Conference on Biometrics (ICB), pp 1–6
46. Qu Z, Liu J, Liu Y, Guan Q, Yang C, Zhang Y (2018) Orienet: a regression system for latent fingerprint orientation field extraction. In: Proceedings of the international conference on artificial neural networks, pp 436–446
47. Rama RK, Namboodiri AM (2011) Fingerprint enhancement using hierarchical Markov random fields. In: Proceedings of the IEEE International Joint Conference on Biometrics (IJCB), pp 1–8
48. Sahasrabudhe M, Namboodiri AM (2014) Fingerprint enhancement using unsupervised hierarchical feature learning. In: Proceedings of the IAPR- and ACM-sponsored Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP), pp 1–8
49. Sankaran A, Vatsa M, Singh R (2015) Multisensor Optical and Latent Fingerprint Database. IEEE Access 3:653–665
50. Sankaranarayanan S, Balaji Y, Castillo CD, Chellappa R (2018) Generate to adapt: aligning domains using generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp 8503–8512
51. Schuch P, Schulz S, Busch C (2016) De-convolutional auto-encoder for enhancement of fingerprint samples. In: Proceedings of the International Conference on Image Processing Theory, Tools and Applications (IPTA), pp 1–7
52. Schuch P, Schulz S, Busch, C (2017) Survey on the impact of fingerprint image enhancement. IET Biometrics pp 102–115
53. Sharma RP, Dey S (2019) Two-stage quality adaptive fingerprint image enhancement using fuzzy c-means clustering based fingerprint quality analysis. Image Vis Comput 1–16
54. Singh K, Kapoor R, Nayar R (2015) Fingerprint denoising using ridge orientation based clustered dictionaries. Neurocomputing 167:418–423
55. Svoboda J, Monti F, Bronstein MM (2017) Generative convolutional networks for latent fingerprint reconstruction. In: Proceedings of the IEEE International Joint Conference on Biometrics (IJCB), pp 429–436
56. Tiwari K, Gupta P (2014) Fingerprint quality of rural population and impact of multiple scanners on recognition. In: Chinese conference on biometric recognition, pp 199–207
57. Tzeng E, Hoffman J, Darrell T, Saenko K (2015) Simultaneous deep transfer across domains and tasks. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp 4068–4076
58. Tzeng E, Hoffman J, Saenko K, Darrell T (2017) Adversarial discriminative domain adaptation. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp 7167–7176
59. Tzeng E, Hoffman J, Zhang N, Saenko K, Darrell T (2014) Deep domain confusion: maximizing for domain invariance. arXiv:1412.3474
60. Vatsa M, Singh R, Bharadwaj S, Bhatt H, Mashruwala R (2010) Analyzing fingerprints of Indian population using image quality: a UIDAI case study. In: Proceedings of the international workshop on emerging techniques and challenges for hand-based biometrics, pp 1–5
61. Volpi R, Morerio P, Savarese S, Murino V. (2018) Adversarial feature augmentation for unsupervised domain adaptation. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp 5495–5504
62. Wang W, Li J, Huang F, Feng H (2008) Design and implementation of Log-gabor filter in fingerprint image enhancement. Pattern Recogn Lett 29(3):301–308
63. Wilson G, Cook DJ (2020) A Survey of Unsupervised Deep Domain Adaptation. ACM Trans Intell Syst Technol 11(5):1–46
64. Wong WJ, Lai SH (2020) Multi-Task CNN for Restoring Corrupted Fingerprint Images. Pattern Recogn 101:107203–107213

65. Yan H, Ding Y, Li P, Wang Q, Xu Y, Zuo W (2017) Mind the class weight bias: weighted maximum mean discrepancy for unsupervised domain adaptation. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp 2272–2281
66. Yang X, Feng J, Zhou J (2014) Localized Dictionaries Based Orientation Field Estimation for Latent Fingerprints. IEEE Trans Pattern Anal Mach Intell 36(5):955–969
67. Zhang W, Xu D, Ouyang W, Li W (2019) Self-Paced Collaborative and Adversarial Network for Unsupervised Domain Adaptation. IEEE Trans Pattern Anal Mach Intell 43(6):2047–2061
68. Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp 2223–2232

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

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

Check for updates

View Article Online
View Journal

Materials Advances

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: N. K. Chourasia, A. Rawat, R. K. Chourasia, H. Singh, R. K. Kulriya, V. Singh and P. K. Kulriya, *Mater. Adv.*, 2023, DOI: 10.1039/D3MA00631J.

This is an Accepted Manuscript, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this Accepted Manuscript with the edited and formatted Advance Article as soon as it is available.

You can find more information about Accepted Manuscripts in the Information for Authors.

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard Terms & Conditions and the Ethical guidelines still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this Accepted Manuscript or any consequences arising from the use of any information it contains.

View Article Online
DOI: 10.1039/D3MA00631J

Unveiling the Potential of $Ti_3C_2T_x$ MXene for Gas Sensing: Recent Developments and Future Perspectives

Nitesh K. Chourasia[a], Ankita Rawat[a], Ritesh Kumar Chourasia[b], Hemant Singh[a], Ramesh Kumar Kulriya[c], Vinod Singh[c], Pawan Kumar Kulriya[a*]

[a]School of Physical Sciences, Jawaharlal Nehru University, New Delhi -110067, India

[b]Department of Physics, Samastipur College, Samastipur-848134 (A Constituent College of L.N.M.U. Darbhanga-846004, India)

[c]Department of Applied Physics, Delhi Technological University, New Delhi-110042, India

Email: pkkulriya@mail.jnu.ac.in

Abstract: $Ti_3C_2T_x$ MXene, a two-dimensional transition metal carbide, has garnered significant attention as a promising material for gas sensing applications due to its exceptional properties. In this comprehensive review article, we provide an in-depth overview of the recent advancements in the field of $Ti_3C_2T_x$ -based gas sensors. We summarize the various synthesis methods employed for $Ti_3C_2T_x$ MXene production, including selective etching of MAX phase precursor materials and controlled annealing processes to customize its surface chemistry and properties. We thoroughly review the gas sensing properties of $Ti_3C_2T_x$ MXene towards a wide range of gases, including volatile organic compounds (VOCs), toxic gases, and other target gases. We provide detailed insights into the gas sensing performance of $Ti_3C_2T_x$ -based gas sensors, including sensitivity, selectivity, response and recovery times, and stability, under various temperature and gas concentration conditions. The gas sensing mechanisms involving charge transfer and surface chemisorption processes between gas molecules and the $Ti_3C_2T_x$ MXene surface are also discussed based on experimental and theoretical studies. Furthermore, we highlight the strategies employed to enhance the gas sensing performance of $Ti_3C_2T_x$ -based gas sensors, such as surface functionalization, doping, and hybridization with other materials. We discuss the influence of various factors, such as annealing temperature, surface modification, and device configurations, on the gas sensing properties of $Ti_3C_2T_x$ MXene-based gas sensors. Finally, we address the challenges and future perspectives of $Ti_3C_2T_x$ -based gas sensors, including the need for standardized testing protocols, long-term stability, and scalability of fabrication processes. Our comprehensive review aims to inspire further research efforts and promote the development of $Ti_3C_2T_x$ MXene-based gas sensors with enhanced performance and broader applications, highlighting the promising potential of $Ti_3C_2T_x$ MXene as a cutting-edge material for gas sensing applications.

Keywords: 2D Materials, MXene, $Ti_3C_2T_x$, Gas Sensor, Response Time

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

1. Introduction:

In our contemporary world, marked by swift industrial growth and urban development, the significance of sensors capable of precisely identifying a wide range of hazardous gases has grown considerably. Consequently, sensitive materials, serving as pivotal elements in gas sensor technology, have garnered significant attention in ongoing research endeavors. This heightened focus is driven by their crucial roles in enhancing industrial safety, safeguarding the environment, and preserving personal well-being.[1, 2] At present, several materials have been investigated and established for gas sensors, such as metal oxides (MO), carbon-based materials, metal-organic frameworks (MOF), conductive polymers, rare-earth oxides, etc.[3-12] Metal oxide semiconductor (MOS) can be classified into two types, n-type (e.g., ZnO, SnO_2, TiO_2, etc.) and p-type (e.g., SnO, CuO, WO_3, etc.). MO is the preferred choice for researchers as a transistor, photodetector, and gas sensor owing to its high sensitivity, high electron mobility, good transparency, low off-state current, chemical stability, compact dimensions, facile fabrication, and affordability.[13-24] For instance, Saini et al. used nanowires of Zn(II) 2,3,9,10,16,17,23,24-octakis(octyloxy)-29*H*,31*H*-phthalocyanine ($ZnPcOC_8$) and nanoflowers of Cu(II) 2,3,9,10,16,17,23,24-octakis(octyloxy)-29*H*,31*H*-phthalocyanine ($CuPcOC_8$) for Cl_2 sensor at room temperature.[25] The fabricated sensor has a detection limit as low as 5 ppb and with responses as high as 715% and 550% within 10 s and 15 s by using simple and cost-effective self-assembly techniques. Chaloeipote et al. employed 3D printing technology for the synthesis of CuO, and the resulting CuO sensor demonstrated a favorable response to the detection of NO_2, registering a response of 14.17% when detecting 200 ppm NO_2 at room temperature. Nonetheless, conventional gas sensors constructed with MO as gas-sensitive materials often attain higher response levels within the temperature range of 200°C to 500°C.[26] This requirement for elevated temperatures not only escalates the cost of gas detection but also introduces potential safety hazards due to the elevated operating temperatures. Wang et al. employed KIT-6 as the rigid template for the synthesis of WO_3 with incorporated Au particles. The resulting Au-WO_3 sensor exhibited remarkable sensitivity in detecting trimethylamine, with a rapid response time of just 1 second and a response rate of 42.56% when exposed to 100 ppm trimethylamine at 268°C.[27] Sun et al. reported H_2 gas sensor using ZnO nanofiber film (average diameter of ~60 nm). Within the temperature range of 210–330 °C, upon exposure to 20 ppm H_2 gas, the resistance of the ZnO nanofiber film increased, and the sensitivity ~ 28.84% at 270 °C.[28] Kumar et al. reported room temperature chemiresistive sensors using hexadecafluorinated copper phthalocyanine

View Article Online
DOI: 10.1039/D3MA00631J

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

(F16CuPc) and graphene oxide (rGO) hybrid materials. The fabricated sensor is highly sensitive to Cl_2 gas with a detection limit upto ppb level.[29] Generally, gas sensors utilizing MOSs typically demand elevated operating temperatures to attain their peak gas-sensing capabilities. This not only results in heightened energy consumption but also raises the potential danger of explosions when detecting flammable and explosive gases. Moreover, the requirement for high operating temperatures restricts their suitability for use in wearable gas sensors. Hence, there is a need to innovate and create new materials for gas sensing that can operate at ambient room temperature.

Recently, two-dimensional (2D) materials have attracted much attention to research society in different scientific applications. Since the discovery of $Ti_3C_2T_x$ MXene in 2011, 2D transition metal carbides/nitrides (together referred to as MXenes) have involved significant research consideration.[30, 31] Owing to their outstanding mechanical, biological and electrical properties such as efficient electromagnetic interference (EMI) shielding, large surface-to-volume ratio, extraordinary electrical conductivity, water dispersibility, high energy capacity, and excellent antibacterial activity, this newly discovered category of materials are important in a wide range of different areas of research.[32-37] In the case of sensors, the replacement of traditional sensor materials (e.g., Metal alloys and carbon) has been crucial in light of the fact that unadventurous technology is progressively reaching its limits in terms of potential improvements to sensor performance. As a result, significant exertions have been made to investigate the use of MXenes in a variety of sensor technologies for biological, chemical, mechanical, optical, and gas sensors.[38-42] MXene's 2D structure, having diverse terminal functionalization groups, provides a large number of active surface sites that can help as quick-to-respond sensor platforms for numerous exterior boosts. Apart from this, the high electrical conductivity of MXene is advantageous to attaining low noise.[43] Because of these unique characteristics, MXenes appear to be a promising alternative sensor material that can provide high sensitivity, extraordinarily low-level detectivity (ppb level), high selectivity, and low power consumption as well as can operate at ambient conditions. Last but not least, the fact that MXenes are easily dispersed in water enables environmentally responsible manufacture as well as amendment treatments, which further increases their utility in the processing industry. $M_{n+1}X_nT_X$ is the general formula of MXenes, where n demonstrates structures. The structure of MXene is known to be influenced by its MAX ($M_{n+1}X_nT_X$) phase precursor, which has previously constrained the MXene family to materials containing 2, 3, or 4 atomic layers of transition metals.[44, 45] Recently introducing a new MXene variant, M_5X_4, which contains

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

five atomic layers of transition metals.[46] M represents an early transition metal (e.g., Ti, Mo, and V), and X refers to carbon/nitrogen or their mixture. At the same time, T_X stands for numerous surface terminations, e.g., oxygen (=O), hydroxyl (−OH), and fluorine (−F).[47] MXene research has sparked interest due to its superior features and ease of processing; hence till now over 30 compositions of MXene have already been reported.[48, 49] Furthermore, the remarkable electrical properties and elevated electron density at the Fermi level exhibited by $Ti_3C_2T_x$ make it highly amenable to surface modification through the introduction of specific terminating species. As a consequence, these distinctive characteristics of $Ti_3C_2T_x$ MXene make it particularly promising for sensitive interactions with toxic gas molecules, rendering it a viable candidate for gas-sensing applications. One effective approach involves the integration of $Ti_3C_2T_x$ nanomaterials with other substances, such as transition metal dichalcogenides (TMDs), metal oxide semiconductors (MOS), and graphene, resulting in potential hybrid materials well-suited for gas sensing applications.[50] In brief, $Ti_3C_2T_x$ MXene materials have emerged as a promising advancement in the realm of gas sensors, offering numerous advantages over conventional metal oxides. Notably, $Ti_3C_2T_x$ MXenes display highest sensitivity to various gases, enabling the detection of even trace concentrations with greater accuracy. Moreover, $Ti_3C_2T_x$ MXene-based sensors exhibit faster response times, crucial for rapid gas detection in safety-critical applications. Unlike many metal oxide sensors that require elevated operating temperatures, $Ti_3C_2T_x$ MXenes can function effectively at room temperature, reducing energy consumption and eliminating the safety risks associated with high temperatures. Their superior selectivity, durability, and flexibility, along with the ability to engineer their properties, make $Ti_3C_2T_x$ MXene materials a compelling choice for gas sensors, especially in applications demanding precision, reliability, and versatility.

Herein, we systematically review the synthesis process, physical properties, and sensing applications in different areas of $Ti_3C_2T_x$ MXenes. Figure 1 shows the structure of this review article. At first, we will start with the synthesis of MXenes using different approaches (e.g. top down and bottom up approach)**,** further our focus will be the properties of MXenes that will be useful of sensing applications. At last, this review article emphasizes the different types of $Ti_3C_2T_x$ MXenes (e.g. pristine, functionalized and composite) for gas sensing applications. Although we will discuss different types of possible MXenes but our goal is to cover the state-of-the-art research accomplishments of using $Ti_3C_2T_x$ MXene and their application in gas sensors, including novel device architecture fundamental sensing mechanisms, and offer wise prospects for upcoming endeavours. This review article offers a fresh perspective on the

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

utilization of $Ti_3C_2T_x$ MXene in gas sensing, surpassing earlier review articles in terms of its comprehensive coverage of recent developments and its forward-looking insights. Unlike previous reviews, this article provides a more up-to-date and detailed examination of the latest advancements in MXene-based gas sensors. It highlights the cutting-edge techniques and applications that have emerged, shedding light on how $Ti_3C_2T_x$ MXene materials can be tailored for enhanced gas-sensing performance. This review articles also summarize theoretical studies of MXene-based gas sensors using Density Functional Theory (DFT) and Molecular Dynamics (MD) simulations, which provide insights into the adsorption mechanisms and dynamic behavior of gas molecules on MXene surfaces, allowing for a comprehensive understanding of sensor performance, selectivity, and stability at the atomic level. By offering future perspectives, it goes beyond summarizing existing knowledge and sets a forward-looking agenda for researchers and engineers, helping to shape the future of $Ti_3C_2T_x$ MXene-based gas sensing technology. This focus on recent developments and future directions positions the article as a valuable resource in the field of gas sensing, providing a more current and forward-thinking perspective compared to earlier review articles.

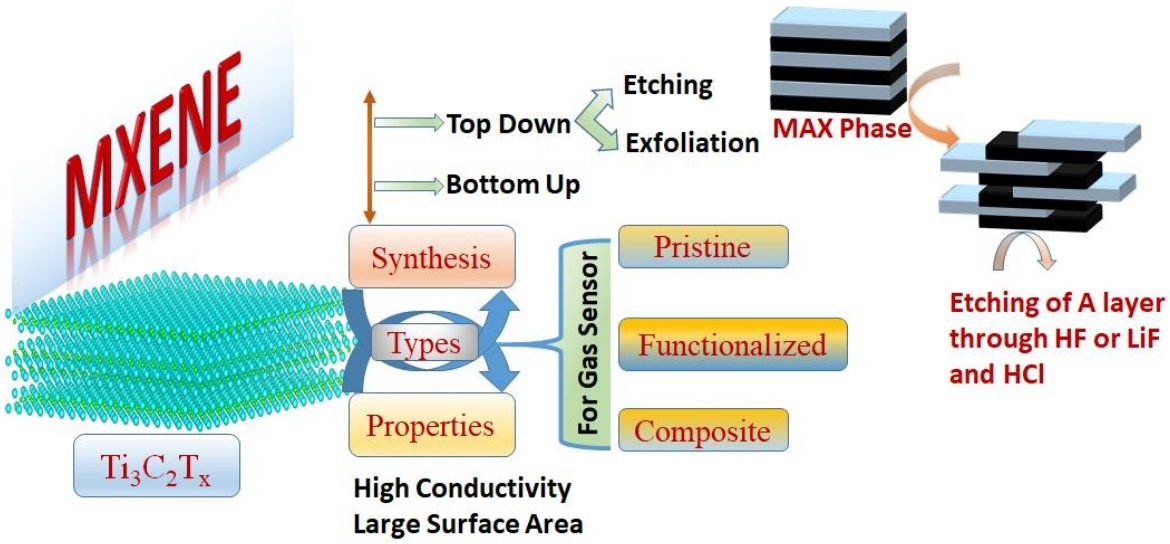


Figure 1: Structure of the review article

2. Technique for Synthesis of MXenes

The techniques used to synthesize the MXene significantly affect its structure and surface functional groups, which in turn impacts the sensor's performance. The top-down etching from MAX phases is now the most widely utilized method for making MXene nanosheets. The layered structure of the MAX phase is made up of densely packed M layers and X atoms that

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

occupy the octahedral sites and are bonded together with A atoms. In general, the M-A bond is typically thought of as a metallic bond, while the M-X bond can exhibit ionic, metallic, or covalent properties.[51] There are mainly two steps involved in the synthesis of MXene, i.e., 1) selective etching of A layer from MAX phase and 2) delamination of MXenes from MAX phase[52]; both steps will discuss one by one.

2.1 Top-down approach:

2.1.1 Etching

In the very first synthesis of MXene, it was suggested that an aqueous solution of hydrofluoric acid (HF) could etch the MAX phase of Ti_3AlC_2, and $Ti_3C_2T_x$ -based MXene can be formed after selective removal of the Al layer. This synthesis method is known as top-down synthesis and is also referred to as the main synthesis route nowadays. The etching of Ti_3AlC_2 involves three reaction steps, as mentioned below

$$2Ti_3AlC_2 + 6HF \rightarrow 2AlF_3 + 3H_2 + 2Ti_3C_2 \quad \ldots\ldots\ldots\ldots\ldots (1)$$

$$Ti_3C_2 + 2H_2O \rightarrow Ti_3C_2(OH)_2 + H_2 \quad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

$$Ti_3C_2 + 2HF \rightarrow Ti_3C_2F_2 + H_2 \quad \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

As one can see from equation 1, during etching, AlF_3 will be a byproduct that is water-insoluble; it's vital to pick the right temperature and time to avoid precipitation of AlF_3. For chemical etching, one needs to consider different reaction parameters like HF concentration, reaction time, optimum temperature, etc. According to the literature, $M_{n+1}AX_n$, with a large n number, necessitates a long etching time, a low pH value, and high etching temperature.[53] Therefore, a high concentration of HF is always beneficial for preferential etching of the Al layer, which can be confirmed from the structural investigation. XRD pattern exhibits a clear downshift of the (002) peak and displays the successful Al etching in Ti_3AlC_2.[54] Furthermore, the experimental results showed that the HF concentration can be used to rectify vacancy defects caused by the removal of surface Ti atoms.

As we know that handling HF is quite difficult, and apart from that, HF is highly corrosive in nature and not environment friendly. Furthermore, HF has a proclivity for over-corroding, resulting in defects in MXene. In 2014, NH_4HF_2 was used as an etchant to avoid using toxic HF directly.[55] This is a significantly safer and milder alternative. In the same year, researchers also utilized LiF and HCl mixture for the etching of the Al layer in Ti_3AlC_2 to form

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

$Ti_3C_2T_x$.[56] The reaction time was 45 hours at a constant temperature of 40 °C, and the results showed that the presence of protons and fluorine ions is essential for effective etching. Later on, researchers also put more effort into refining this method by varying the feeding ratio of LiF and MAX (LiF:MAX = 5:1 or 7.5:1).[57] Excess LiF was found to promote Al etching and Li^+ intercalation, resulting in $Ti_3C_2T_x$ MXene nanosheets with identical thickness, larger size, and fewer defects. In most cases, the etching process transforms solid dense MAX into structures resembling accordions with loosely collected layers, and these formations are also referred to as multilayer MXenes. As a result, an exfoliation step is compulsory for multilayer MXenes into monolayers.

2.1.2 Exfoliation:

Etching and delamination are the two primary processes that make up the majority of the MXene exfoliation process from the MAX phase. In the first step, a chemical etching process is used to remove the "A" layer from the middle of the MAX layers. Each MXene layer must be delaminated from the etched MAX phase in the second step. This procedure for the synthesis of MXene was carried out in a number of different ways, as shown in figure 2.[58] Each technique either incorporates HF into the MAX solution directly or introduces it indirectly through the use of LiF. The chemical connection that holds the M and X layers together can be broken apart if they are subjected to sonication, intercalants like dimethyl sulfoxide (DMSO), or ionic solutions. In any of these three scenarios, the layers are more likely to delaminate. Apart from this, LiF and HF are also favorable for non-Ti-based MXene (e.g., niobium (Nb), vanadium (V), etc.)[59, 60]

The nanosheets of MXene were able to be exfoliated into a single layer by first inserting large organic molecules into the interlayers of the accordion-like structure and then subjecting the structure to ultrasonication or mechanical vibration. [52] This allowed the nanosheets to be rearranged into a single layer. The frequently used intercalants comprise NH_4^+, dimethyl sulfoxide, hydrazine, urea, and tetrabutylammonium hydroxide (TBAOH).[61-63] Recently, Chia et al. reported that by using TBAOH as the intercalant, one could exfoliate $Ti_3C_2T_x$ MXene from HF etching products.[64] After removing the A layer, expanding MAX powder that contained micron-sized flakes led to the formation of Ti_3C_2-HF, which was then delaminated into MXenes with one or more than one layers by reducing the strength of the interlayer contact with TBAOH.

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

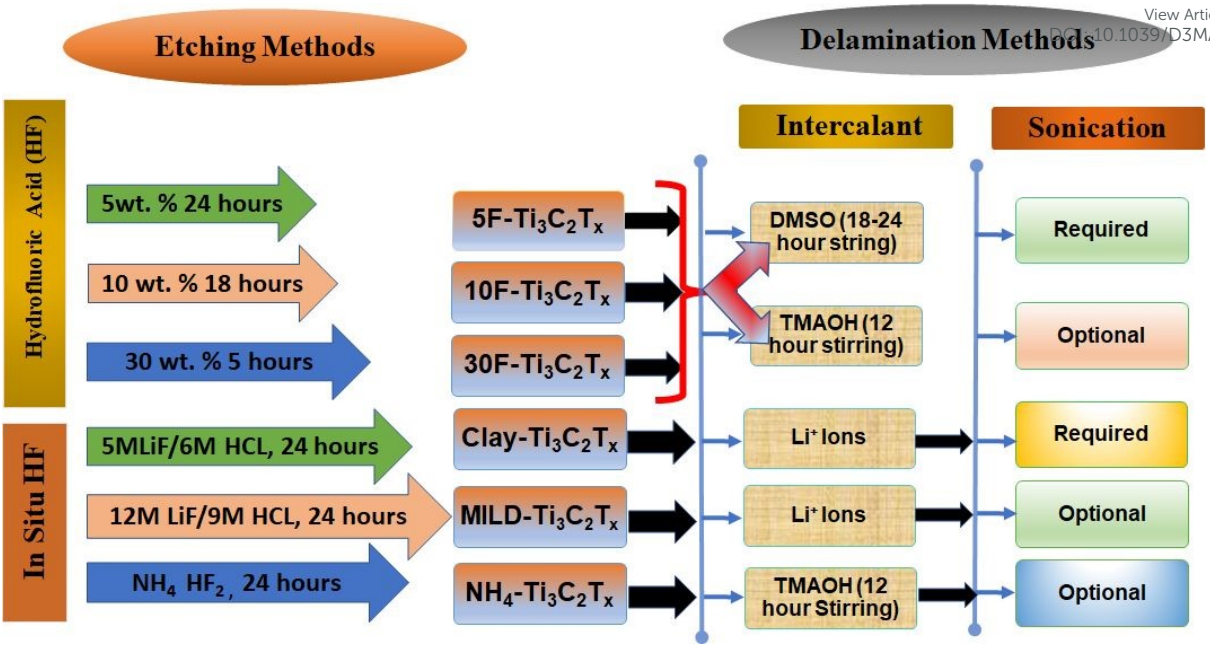


Figure 2: Schematic diagram of the synthesis route of $Ti_3C_2T_x$ MXene, Reproduced with permission from reference[58]. Copyright 2020. MDPI (open Access)

2.2 Bottom-Up Approach:

In addition to the top-down tactics that were discussed above, a multifunctional bottom-up approach known as chemical vapor deposition (CVD) was developed for the purpose of synthesizing high-quality and large-area MXenes. In contrast to the top-down method, which has a complex combination of groups and low controllability, the bottom-up approach has the advantage of accurately manipulating the morphology and termination groups of MXenes. The synthesis process is crucial for understanding the inherent characteristics of MXenes. Researchers reported ultrathin Mo_2C, WC, and TaC MXene crystals on Cu/Mo foil using CVD, and all of them have fewer defects with larger lateral dimensions.[65] The bottom-up synthesis technique of MXenes is still a big problem due to the costly equipment and tedious procedures.

In addition, if the MXene was subjected to mechanical vibration or sonication, it would contract and develop further defects. The inherent properties of 2D TMCs and TMNs have been hindered due to the chemical modification and mechanical damage that has been done to MXenes. The constituent parts of the mother 3D MAX phase have a direct influence on the structure and species of MXenes and vice versa. All of the MXenes that have been synthesized up to this point contain structures of the M_2X, M_3X_2, and M_4X_3 types. New synthesis methods are urgently required to improve the quality of 2D TMCs and TMNs as well as to expand the

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

family of these materials because these MX-structured materials make up a large class of TMCs and TMNs and have a variety of fascinating properties

The CVD production of high-quality ultrathin TMCs crystals utilizing a bilayer metal foil as a growth substrate was initially reported in 2015 by Xu et al., who was also the first to reveal the technique.[65] The stack of Cu/Mo foils was annealed in hydrogen (H_2) to a temperature that was higher than 1085 ℃, which is the melting point of copper. Following this step, methane was injected, which led to the formation of Mo_2C crystals on the surface of the liquid copper. During the course of development, the top layer of liquid Cu plays several crucial tasks that are essential to the process. Meanwhile, it participates in the process of converting methane into carbon atoms by acting as a catalyst, and it serves as a pathway for controlling the diffusion of molecular molybdenum from the molybdenum foil to the surface of the liquid copper. On the surface of the liquid Cu, ultrathin Mo_2C crystals form due to a process involving carbon and molybdenum atoms. In addition, due to the significant differences in the chemical reactivity of copper and molybdenum carbide (Mo_2C), copper acts as a sacrificial layer after Mo_2C is developed. This allows the transfer of Mo_2C to desired target substrate through etching. A bottom-up method, such as CVD, can produce high-quality films on various substrates; however, this method is not typically utilized in the production of MXenes because the films produced using this method are not mono-layers but relatively thinner films. The CVD was the method Xu and colleagues used to fabricate thin films of Mo_2C, WC, and TaC, for example.[66] In spite of this, even the most imperceptibly thin Mo_2C films consisted of at least six layers of Mo_2C, as opposed to a single layer of MXenes. To synthesize ultrathin TaN, Wang et al. used NH_3 as a nitrogen source and Cu/Ta foil as a growth substrate at 1077 ℃. Copper foil maintains a solid state throughout CVD development. According to the images obtained from the TEM, both TaN crystals are single crystals. In addition, ultrathin TaB crystals were produced by employing boron powder as the source of element B. Wang et al. synthesised zirconium based MXenes (e.g. Zr_2CCl_2 and Zr_2CBr_2) by utilizing direct CVD techniques.[67] These two zirconium MXenes seemed in the same common morphology as that of the Ti-MXenes. In summary, the choice between bottom-up and top-down synthesis methods for MXene can influence its gas sensing properties. Bottom-up synthesis offers precise control over composition and structure, while top-down synthesis provides scalability but may yield less controlled surfaces. Both approaches have their advantages and considerations depending on the specific gas sensing requirements and desired performance characteristics.

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

3. Properties of MXene

High electrical conductivity, unique layered structure, wide specific surface area, and excellent dispersity distinguish MXenes from other functional materials, paving the way for further research and development in the sensor industry.[68] In addition to its superior conductivity, the most well-liked $Ti_3C_2T_x$ MXene also has a large surface area and an abundance of surface functional groups.[69]

3.1 Stability

The key parameter which decides the stability of MXene is; reaction conditions, solutions, and composition. Since, $Ti_3C_2T_x$ MXene is easily oxidized in the water and air, its environmental stability is an important study topic, as is its stability and service life during the sensor fabrication and utilization. Because the oxidation of MXene often begins at the edge of the molecule and then spreads throughout the molecule, it is impossible for the entire molecule to be oxidized at the same time. Oxidation proceeds most slowly in the solid state and most quickly in the liquid state. The reaction of oxidation can be understood from the below reaction-

$$Ti_3C_2O_2 + 4H_2O \rightarrow 3TiO_2 + 2C + 4H_2$$

MXene's fragile oxidation resistance is one of the challenges it faces in the process of usage, and it's worth looking into ways to alleviate the performance loss affected by oxidation**.** $Ti_3C_2T_x$ MXene can be effortlessly transformed into TiO_2 through annealing at either 1150 ℃ for thirty seconds or 400 ℃ for a longer time.[70, 71] Some of MXene's benefits, like conductivity and capacitance, are often sacrificed during oxidation. As reported earlier, the ionic group of polyphenylene has the ability to effectively occupy the edge of the MXene sheet, which protects it from being oxidized by the water molecules.[72]

3.2 Electrical Properties

The electrical properties of a single layer of $Ti_3C_2T_x$ MXene are worth researching in order to better understand its electrical performance and potential application in sensors. The band gap in MXene arises due to surface functionalization; otherwise, all MXenes exhibit metallic behaviour. The metallic $Ti_3C_2T_x$ material really does have the capacity to serve as an electrode for super capacitors. The performance of electrochemical capacitor devices is significantly influenced by the electronic conductivity of MXene. After being functionalized, some MXenes transform into semiconductors with energy gaps ranging from 0.25 to 2.0 eV.[73]

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

3.3 Mechanical Properties

Because of their exceptional in-plane tensile characteristics, bending tolerance, and flexibility, MXenes are perfect for use in the fabrication of bendable electrodes and other electrical devices.[56] The surface terminations of MXenes have a significant impact on the mechanical characteristics. MXenes with an O termination are expected to have extremely high stiffness; however those with F and OH terminations exhibit lower elastic stiffness than O-terminated MXenes. That is why bonding strength between Ti-O being stronger than in the situations of Ti-OH and Ti-F. The flexibility of the surface-functionalized MXenes is greater than that of bare MXenes. The $Ti_3C_2T_x$ materials also have a remarkable degree of flexibility and tensile resistance. The 5µm thick $Ti_3C_2T_x$ sheet can tolerate 1.3MPa or roughly 4000 times of own weight, with no any observable distortion or fracture.[74]

4. Important characteristics of gas sensors

The main characteristic of a gas sensor is determined by a few critical parameters such as sensitivity, selectivity, stability, resolution, and response/retrace time shown in figure 3.

- **Sensitivity or Response:** It is defined as a relative change in sensor signal per unit concentration of the given analyte. Sensitivity (S) is often expressed as Ra/Rg for reducing gases or Rg/Ra for oxidizing gases, where Ra denotes the resistance of the gas sensors in the reference gas (typically air) and Rg denotes the resistance in the reference gas containing the target gases.
- **Selectivity:** The ability to selectively identify the particular analyte in the presence of other analytes.
- **Stability:** The time period over which the sensor retains its original characteristics.
- **Detection limit:** The lowermost concentration difference can be eminent by the sensors.
- **Response time and recovery time**: The response and recovery time was defined as the time required for 90 % of the total resistance change to occur while introducing and removing the analyte.

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

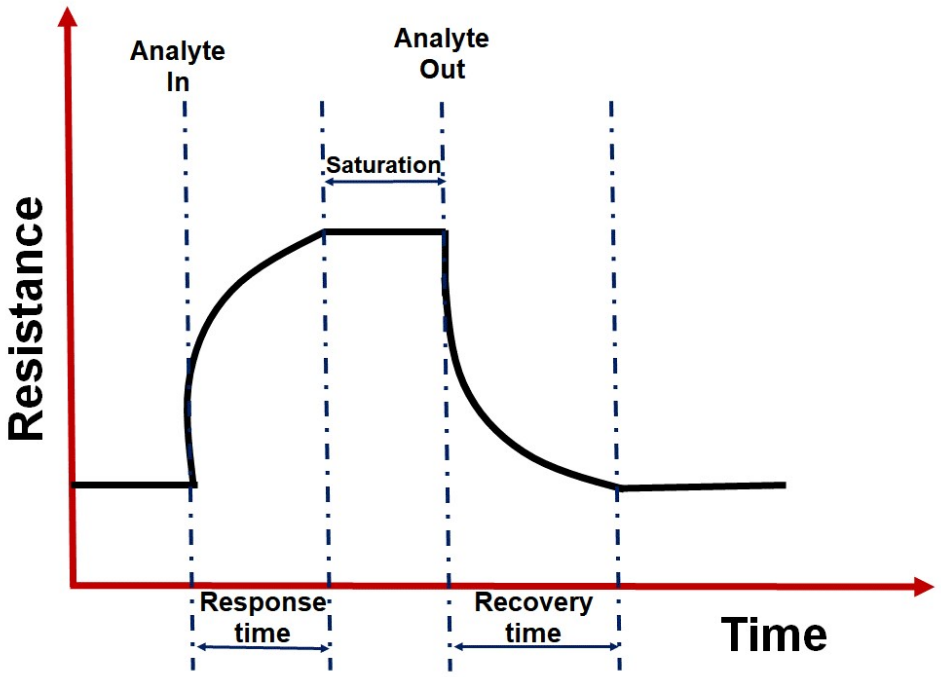


Figure 3: Important characteristics of a typical metal oxide-based gas sensor

4.1 Working Principle and Mechanism of Chemiresistive Gas Sensor

A chemiresistive gas sensor is a remarkable device designed to detect the presence of specific gases by measuring changes in the electrical resistance or conductivity of a sensing material when it interacts with the target gas. The best example of chemiresistive gas sensor are metal oxide (MO) gas sensor. The heart of any chemiresistive gas sensor is the sensing material. The choice of sensing material is pivotal to the sensor's performance because different materials exhibit varying levels of sensitivity to different gases. Common sensing materials include metal oxides, conducting polymers, and organic materials, each with its unique gas-sensing mechanisms. The gas sensing mechanism of a chemiresistive gas sensor primarily involves two fundamental processes: adsorption and chemisorption.

4.1.1 Absorption Mechanism

Adsorption is the process by which gas molecules adhere to the surface of the sensing material without forming strong chemical bonds. It occurs due to weak van der Waals forces, electrostatic interactions, or physisorption. When a gas molecule interacts with the surface of the sensing material, it may physically adsorb onto the surface or get trapped in surface defects or pores. This adsorption process can lead to a change in the electronic structure of the sensing

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

material. In the case of semiconducting materials like MOs or carbon nanotubes, gas adsorption can alter the concentration of charge carriers (electrons or holes) in the material's crystal lattice.

4.1.2 Chemisorption Mechanism

Chemisorption is a more specific and stronger form of gas-surface interaction than adsorption. It involves the formation of chemical bonds between gas molecules and surface atoms of the sensor material. In chemisorption, the gas molecules undergo a chemical reaction with the surface atoms, resulting in the creation of new chemical entities on the sensor's surface. The formation of these chemical bonds can lead to significant changes in the electronic structure of the sensing material, including changes in the number of charge carriers and their energy levels. The strength of chemisorption depends on the nature of the gas and the surface chemistry of the sensing material. Some gases chemisorb readily, while others may only adsorb weakly or not at all.

4.2 Factors Affecting Gas Sensing Performance

Several factors can influence the gas sensing mechanism and the performance of sensor:

1. Sensing Material: The choice of sensing material is crucial and depends on the target gas or gases. Different materials have varying degrees of sensitivity to specific gases.

2. Target Gas: The type of gas to be detected plays a significant role. Different gases have distinct chemical properties, and their interactions with the sensing material vary accordingly.

3. Temperature: Temperature can influence the gas-sensing mechanism by altering the kinetics of adsorption and desorption processes. Many chemiresistive sensors operate at elevated temperatures to enhance sensitivity.

4. Humidity: Humidity levels can affect the adsorption and chemisorption processes. Some sensors are sensitive to changes in humidity, which can lead to false readings.

5. Interfering Gases: The presence of other gases in the environment can interfere with the sensor's response to the target gas. Selectivity is a critical consideration in sensor design.

In conclusion, chemiresistive gas sensors operate on the fundamental principles of adsorption and chemisorption. These mechanisms lead to changes in the electrical conductivity or resistance of a sensing material when it interacts with a specific target gas. The choice of sensing material, target gas, environmental conditions, and the presence of interfering gases all

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

influence the sensor's performance. Understanding the gas sensing mechanism is essential for designing and optimizing chemiresistive gas sensors for various applications.

4.3 Sensing Mechanism of $Ti_3C_2T_x$ MXene based Gas Sensor

The gas-sensing mechanisms in typical MO gas sensors are closely tied to the exchange of electrons, which occurs as a result of electron donation or acceptance from the target gas. Nevertheless, the gas-sensing mechanisms of MXenes exhibit greater complexity compared to the conventional charge-transfer model. In the previous report on $Ti_3C_2T_x$ MXene based gas sensor, p-type sensing behavior was observed for ethanol, methanol, acetone, and ammonia gas at room temperature.[75] However, in subsequent investigations, it was observed that $Ti_3C_2T_x$ MXene consistently exhibited an increase in resistance when exposed to either oxidizing or reducing gases.[74, 76] Based on the above observations, researchers proposed two possible mechanism. **(1)** MXene exhibits metallic characteristics rather than semiconductor behavior, and the adsorption of gas can diminish the number of carriers while raising the channel resistance of MXene[77], and **(2)** The expansion of MXene interlayers upon exposure to gases obstructs its electron transport in the out-of-plane direction.[78] Both of these mechanisms can clarify why MXenes consistently display a positive response to all gases. Though, according to Yang et al. report, the response signal of $Ti_3C_2T_x$ MXene had a shift from positive to negative followed by NaOH treatment.[79] The change in response direction is ascribed to the rise in the O/F ratio, suggesting that surface terminations also play a noteworthy role in gas-sensing behavior. However, the fundamental mechanism behind this transformation remains unclear.

In sensors, employing hybrid structures combining MXenes and other materials, the sensor's response direction frequently hinges on the MXene's concentration within the composite. When the composite features a notably high MXenes content in conjunction with MOs, the gas response of the composite sensor aligns with that of a pure MXenes sensor, and the composite sensor exhibits significantly lower resistance compared to a pure MO sensor.[80] In this scenario, the presence of MOs enhances the specific surface area and facilitates the adsorption of gases.[81] Conversely, in situations where the MXenes content is relatively low, the gas response of the composite sensor aligns with that of the MO sensor, and the composite sensor exhibits considerably higher resistance compared to a pure MXene sensor.[82] In this instance, the MO take on a central role in facilitating carrier conduction during the gas-sensing process.

5. Why MXene for Sensing Applications?

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

The integration of sophisticated nanomaterials has ushered in a new era in the realm of sensors. When it comes to the process of signal transduction, selecting the appropriate transducer (material) is of the utmost importance. Traditional transducers suffer from the necessity of a high over-potential in the case of electrochemical sensors because of their extremely small surface area. Additionally, the target analyte's slow charge kinetics and the mass transfer constraint make it difficult for the transducer to transfer enough mass. The materials for strain sensors, instead, must preserve conductivity and mechanical strength. As a result, any viable sensor device's transducer material should have significant properties such as high surface area and high conductivity, superior mechanical strength, and a simple synthesis process. To accomplish this, a wide range of materials, including metals, polymers, metal oxides, and carbonaceous materials, as well as their combinations, have been investigated. The race to identify the ideal combination with a high signal throughput is still ongoing despite this. In 2011, Drexel University was credited with discovering MXenes, a novel type of 2D material.[30] Although their primary application is in the field of energy materials, such as batteries and supercapacitors, the community of sensor researchers is showing an unusual interest in these 2D materials.[83] The diversity of their surface and chemical properties is directly tied to the growing application of sensor technology, which is directly related to their increasing reach. MXenes have relatively high hydrophilicity and are simple to functionalize on the surface. This is in contrast to other 2D materials like graphene, which, despite having a high conductivity, is challenging to integrate with electrode systems due to its low hydrophilicity and the fact that its sheets can be folded or re-stacked.[84] In addition, because of their exceptionally high conductivity, MXenes are an excellent substrate for the research and development of electrochemical sensors.[33] Because of their superior mechanical qualities and conductivity, MXenes provide a great conductive filler that may be used in strain and pressure sensors. It is an attractive transducing material for various sensors due to MXenes' excellent surface characteristics, metal-like conductivity, and flexibility with just the right amount of stiffness. These characteristics make MXenes suitable for electrochemical, strain bio-electrochemical, photochemical, and pressure sensors.[85]

6. Application of $Ti_3C_2T_x$ MXene as a Gas Sensor

In order to make an accurate comparison of $Ti_3C_2T_x$ MXene's application in gas sensors, the nanomaterial can be divided into three distinct classes: pure $Ti_3C_2T_x$, functionalized $Ti_3C_2T_x$, and composite $Ti_3C_2T_x$.This review article investigates the functionalization of $Ti_3C_2T_x$ in terms of the alteration of functional groups and the intercalation of metal ions. Additionally,

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

combining $Ti_3C_2T_x$ with other materials is a successful tactic to enhance its gas-sensing performance. There are many various subgroups of materials that can be composited with $Ti_3C_2T_x$, including transition-metal dichalcogenides (TMDs), graphene derivatives, metal oxides, polymers, etc. $Ti_3C_2T_x$ can be used to make a variety of different composites. The application in gas sensors of all three different categories of $Ti_3C_2T_x$ will be discussed one by one.

6.1 Pure $Ti_3C_2T_x$ for Gas Sensor

After the discovery of MXene in 2011, $Ti_3C_2T_x$ was initially used as a gas sensor, as reported by Lee et al..[75] They reported the $Ti_3C_2T_x$-based volatile organic compounds (VOC) gas sensor at room temperature by integrating the $Ti_3C_2T_x$ nanosheets with solution-processed on flexible polyimide film. At ambient temperature, the $Ti_3C_2T_x$ sensors demonstrated p-type sensing behaviour while successfully measuring methanol, ethanol, acetone, and ammonia gas. The fabricated sensor exhibited the highest and lowest response when exposed to ammonia gas due to large adsorption energy, as shown in figure 4 b). Theoretically, the estimated value of the limit of detection (LOD) for acetone gas is 9.27 ppm. To detect VOC gases at parts per billion (10^{-9}) level, Kim et al. proposed a $Ti_3C_2T_x$ gas sensor with an extremely high signal-to-noise ratio (SNR).[74] They deposited a thin film of fully functionalized and high metallic conductivity of multi-layered $Ti_3C_2T_x$ on a pre-printed Au electrode SiO_2 substrate. The $Ti_3C_2T_x$ gas sensor's response and SNR were compared to those of typical 2D materials (including MoS_2, black phosphorus (BP), and reduced graphene oxide (r-GO)). It was revealed that the $Ti_3C_2T_x$ sensor's SNR was two orders of magnitude higher than that of other 2D materials because of their incredibly low noise level. $Ti_3C_2T_x$ MXene gas sensors showed a very low limit of detection (LOD) for VOC gases of 50–100 ppb at room temperature. A single-layer Ti_3C_2-based NH_3 sensor was proposed by Wu et al. at ambient temperature with high selectivity [figure 4 c) and d)].[86] Following intercalation, ultrasonic dispersion, and delamination, Ti_3C_2 was coated on the ceramic tube's surface to make the gas sensor ready to detect a variety of gases, including NO, CH_4, NH_3, H_2S, H_2O, acetone ethanol, and methanol with 500 ppm concentration at room temperature. According to their research, the single-layer Ti_3C_2sensor had the highest response, which was adequately explained by first-principles calculations. Several investigations on NH_3 showed that the manufactured Ti_3C_2-based sensors have good repeatability to ammonia at room temperature, with a detection limit of 10 ppm. Additionally, utilising practical first principles, Khakbaz et al. thoroughly investigated the possibility of pure$Ti_3C_2T_x$ for gas sensing.[87] On $Ti_3C_2T_x$, gas molecules such as NH_3, CO,

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

N_2O, NO, NO_2, CH_4, CO_2, and H_2S were adsorbed, and the effects of various ratios of numerous functional groups were examined. The results showed that $Ti_3C_2T_x$ is especially sensitive to NH_3 relative to the other gas molecules studied, with a charge transfer of 0.098 e and adsorption energy of 0.36 eV, respectively. Using first-principles analysis, Zeng et al. also reported that $Ti_3C_2T_x$ MXene is foreseen to be an encouraging sensor for SF_6 decomposition characteristic components.[88] The adsorption strength of analytes is unevenly exhibited by different terminated $Ti_3C_2T_x$ MXene, and hydroxyl groups greatly enhance the adsorption strength on the $Ti_3C_2T_x$ MXene surface. Based on their finding they also suggested that following the adsorption of the analytes, a clear charge transfer is what primarily causes variations in $Ti_3C_2T_x$ conductivity.

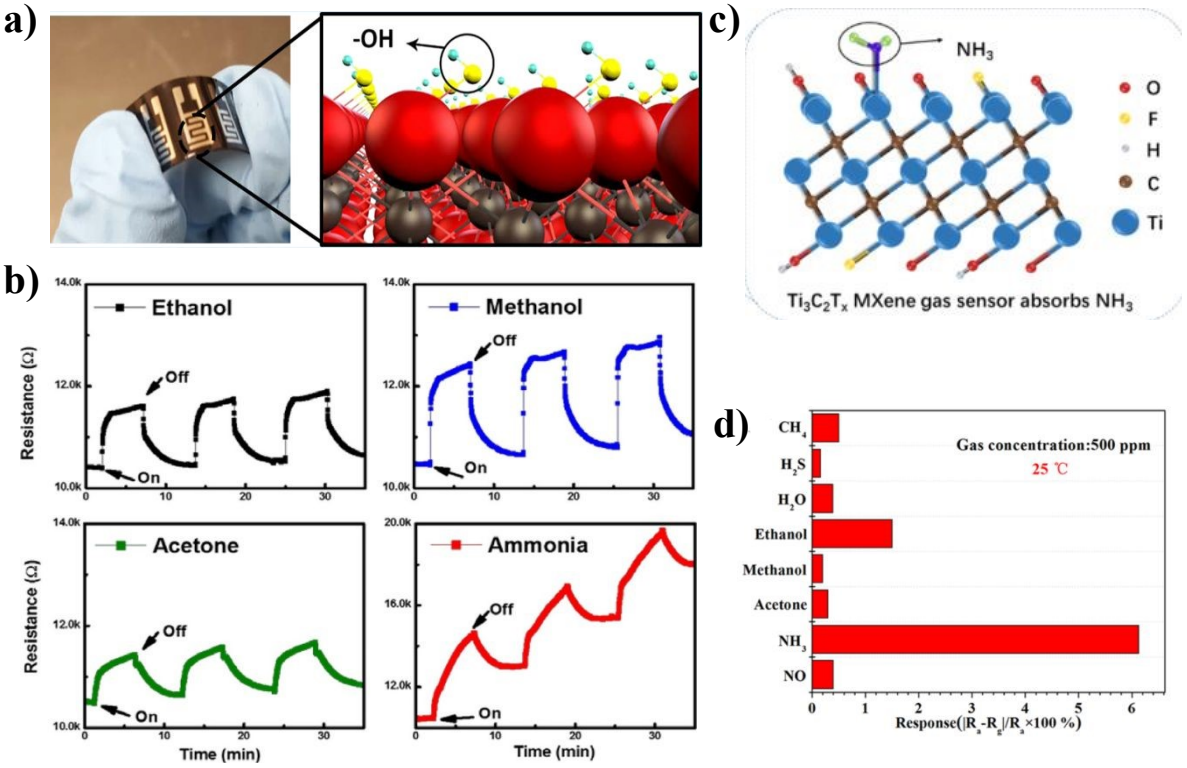


Figure 4: a) Schematic illustration of pure $Ti_3C_2T_x$ VOCs, **b)** Results of ethanol, methanol, acetone, and ammonia gas bubbling at 100 ppm using $Ti_3C_2T_x$ based sensor at 300 K, Reproduced with permission from reference [75] Copyright 2017. American Chemical Society, **c)** Pictorial representation of absorption of NH_3 on the surface of $Ti_3C_2T_x$ MXene and, **d)** Room temperature response of Ti_3C_2 MXene gas sensors toward 500 ppm CH_4, H_2S, H_2O, ethanol, methanol, acetone, NH_3, and NO, Reproduced with permission from reference [86] Copyright 2019. American Chemical Society

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

Numerous variables, including film thickness, MAX phase precursors, flake size, and degree of oxidation, have an impact on the gas sensing capabilities of pure $Ti_3C_2T_x$.[89, 90] For gas (acetone, ethanol, and NH_3) detection, Kim et al. also described an interfacial assembly of ultrathin $Ti_3C_2T_x$ films.[90] In particular, the addition of ethyl acetate to the $Ti_3C_2T_x$ solution promotes the production of a self-assembled layer, which had an excellent plane adhesion and stacking order. It was discovered that the self-assembly technique could improve the connectivity of small flakes, leading to minimal electrical noise in the gas sensor [figure 5 a) and b)]. The gas sensor made of ultrathin $Ti_3C_2T_x$ responded to NH_3 the best, as shown in figure 5 c). According to research by Shuck et al., the MAX phase carbon sources had an impact on the $Ti_3C_2T_x$'s structure and characteristics.[54] The three different types of $Ti_3C_2T_x$ that were produced through etching varied in terms of chemical composition, lateral flake size, electrical conductivity, and stability. They reported that $Ti_3C_2T_x$ made from graphite had the highest conductivity (4400 S/cm) and was the most stable (time constant, 10.1 days), whereas MXene made from TiC had a similar conductivity (3480 S/cm) but had the lowermost colloidal stability (4.8 days) as well as the lowest chemical stability (5.1 days). The most responsive to gases and with good selectivity to NH_3 was the TiC-produced $Ti_3C_2T_x$, as shown in figure 5 d). In spite of this, $Ti_3C_2T_x$ has not made significant progress in the field of gas sensors due to the fact that it is easily oxidized and has a tendency to be unstable. Chae et al. looked into the key variables (including water molecules, temperature, and solvents) controlling the oxidation of $Ti_3C_2T_x$ in order to get around this constraint. They proposed that low temperature and ethanol solvent can help in maintaining the intrinsic properties of prepared $Ti_3C_2T_x$.[91] For the purpose of detecting VOCs at room temperature, Liu et al. recently proposed termination-modified $Ti_3C_2T_x$ MXene ($Ti_3C_2T_x$–M) through alkali pre-treatment and trimethylacetic anhydride functionalization [photograph is shown in figure 5 e)].[92] In comparison to its unprepared counterpart, the $Ti_3C_2T_x$ sensor showed a fivefold increase in response to ethanol with a 29 % reduction in response to water vapor followed by the spray coating methods, as shown in figure 5 f).

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

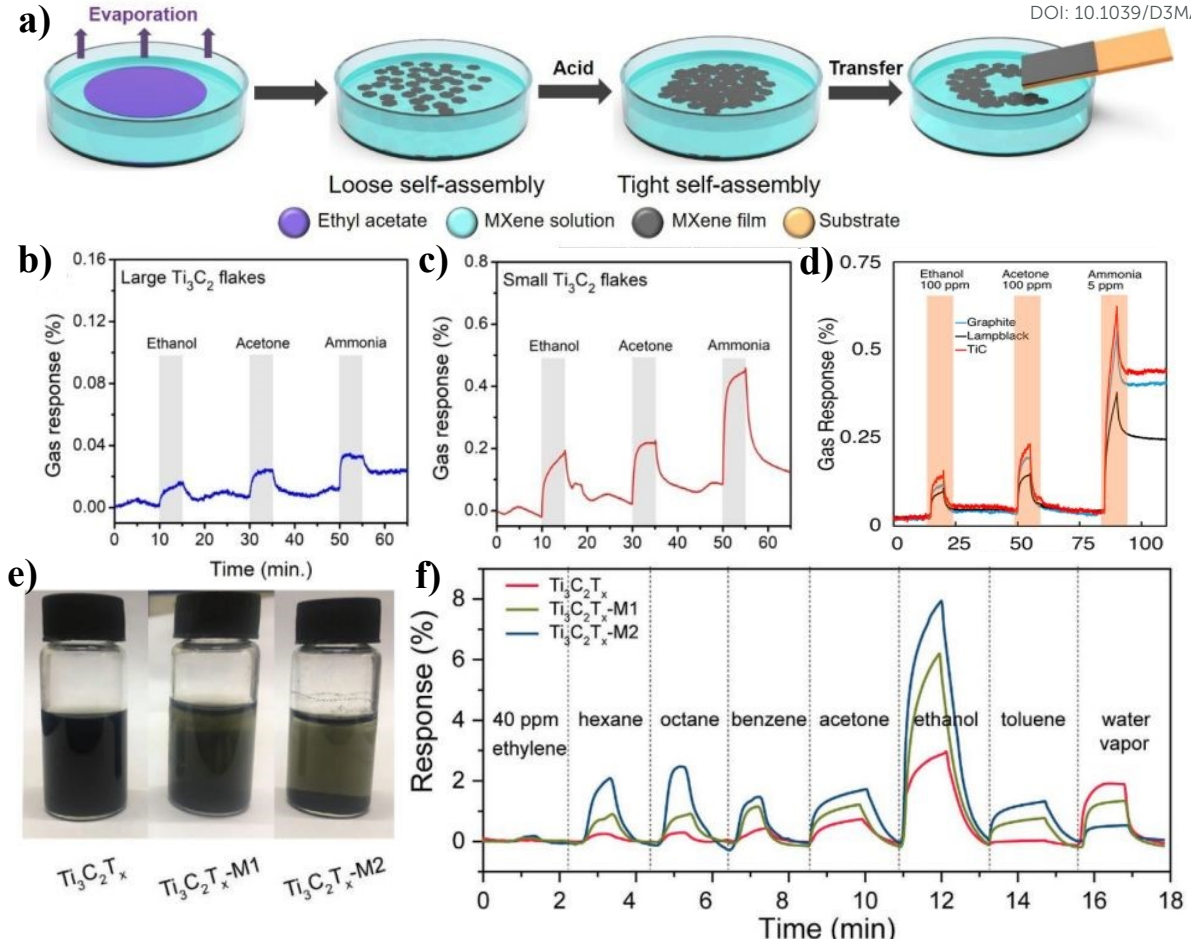


Figure 5: a) Schematic of the interfacial self-assembly technique for preparation of ultrathin MXene films. Gas sensing response of, **b)** Large Ti_3C_2 flakes, **c)** Small Ti_3C_2 flakes, Reproduced with permission from reference [90] Copyright 2019. American Chemical Society, **d)** $Ti_3C_2T_x$'s ability to detect gases after being made from various MAX precursors which showed different sensitivities with ethanol, acetone, and ammonia Reproduced with permission from reference [54] Copyright 2019. American Chemical Society , **e)** Photographs of pure $Ti_3C_2T_x$, $Ti_3C_2T_x$ -M1, and $Ti_3C_2T_x$ -M2 colloid solution and, **f)** Gas sensing performance of VOCs at 40 ppm concentration, Reproduced with permission from reference [92] Copyright 2021. Wiley

In summary, one notable advantage of pristine $Ti_3C_2T_x$ MXene-based gas sensors is their exceptional sensitivity and rapid response to a wide range of gases. These sensors exhibit high surface area, enabling effective gas adsorption, and their 2D layered structure provides numerous active sites for gas molecules to interact with, resulting in enhanced detection capabilities. Additionally, MXene-based sensors often possess excellent electrical conductivity, facilitating quick and precise signal readouts. These properties make them highly

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

versatile for gas sensing applications, ranging from environmental monitoring to industrial safety, where the ability to detect and respond to diverse gas species swiftly and accurately is of paramount importance. However, one significant drawback of pristine $Ti_3C_2T_x$ MXene-based gas sensors is their limited selectivity when detecting multiple gases simultaneously. The inherent reactivity of MXene materials can lead to cross-sensitivity issues, making it challenging to distinguish between different gas species accurately. The MXene's hydrophilic nature can lead to water absorption, which interferes with gas sensing performance and reduces sensitivity. To overcome this limitation, researchers have been exploring strategies such as surface functionalization and composite material design, which we will discuss one by one in next section. Surface modification with specific functional groups or the incorporation of other nanomaterials can enhance the selectivity of MXene-based sensors by tailoring their chemical affinity to target gases. These advancements in MXene-based gas sensors hold promise for enhancing their practical utility in real-world applications.

6.2 Functionalized $Ti_3C_2T_x$ for Gas Sensor

Even though $Ti_3C_2T_x$ has metallic properties when it is pure, its electronic properties can be changed by functionalization and become semiconducting. The functionalization can be done through metal ion intercalation[93], sulphur doping[94], oxidation[95], noble metal modification[96], and functional group modification.[97, 98] Researchers have started to pay more attention to functionalization because it can improve the adsorption effect of $Ti_3C_2T_x$, which is favorable for gas sensing applications. Metal ion intercalation was studied by Koh et al. as part of their investigation into how gas-induced interlayer swelling of $Ti_3C_2T_x$ influences the performance of gas sensing.[93] When NaOH was utilized as the ionic intercalating agent, the in-situ X-ray diffraction (XRD) experiments demonstrated a significant relationship between the gas response and the interlayer transport caused by the interlayer ions as shown in figure 6 a). The highest level of sensitivity was reported in $Ti_3C_2T_x$ after treatment with 0.3 mmol/L NaOH, and the highest response was achieved for ethanol, as shown in figure 6 b). Similar to this, Yang et al. enhanced the gas sensing performances of the organ-like $Ti_3C_2T_x$ by -OH functional alteration and alkaline metal ion (Na) intercalation.[79] The sensing components were uniformly coated on the Al_2O_3 ceramic substrate by drop-coating process to create the pristine $Ti_3C_2T_x$ and alkalized $Ti_3C_2T_x$ gas sensors to study the alkalized $Ti_3C_2T_x$'s gas sensing performance and the schematic is shown in figure 6 c). It was discovered that the alkalized $Ti_3C_2T_x$ exhibited the opposite tendency to the pristine $Ti_3C_2T_x$, whose resistance rose with the concentration of NH_3. The highest response with alkalized $Ti_3C_2T_x$ was achieved

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

in NH_3 gas (around 28.7 % more than pristine $Ti_3C_2T_x$) and the selectivity also enhanced with alkalized $Ti_3C_2T_x$ (ethanol, acetaldehyde, formaldehyde, methanol, acetone, methane, and NO_2) as shown in figure 6 d).

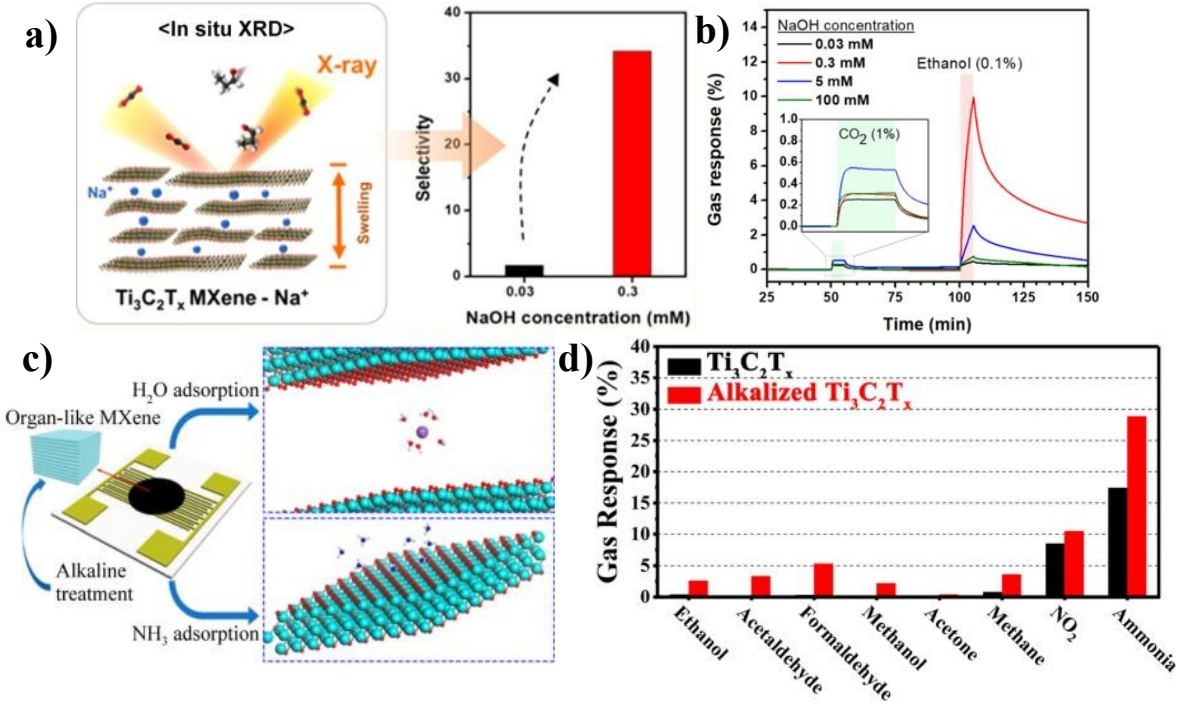


Figure 6: a) Gas sensor schematic based on $Ti_3C_2T_x$, where mM stands for mmol/L , **b)** Gas response of $Ti_3C_2T_x$ sensors with different concentrations of NaOH, Inset: Magnified gas response of CO_2, Reproduced with permission from reference [93] Copyright 2019. American Chemical Society, **c)** Schematic of an organ-like $Ti_3C_2T_x$ -based gas sensor with alkaline treatment, **d)** Selectivity test of alkalized $Ti_3C_2T_x$ and $Ti_3C_2T_x$ with a 100 ppm concentration to various test gases, Reproduced with permission from reference [79] Copyright 2019. American Chemical Society

The inherently metallic $Ti_3C_2T_x$ was functionalized with oxygen by Pazniak et al. to increase gas sensitivity. Figure 7 a) to g) [95] shows that Ti_3C_2T had been partially oxidized kept its conductivity at the optimum annealing temperature of 350 ℃, and the TiO_2 formed during oxidation improved its detection capacities to gases at concentrations as low as 10^{-6}. The partially oxidized multisensor array-based $Ti_3C_2T_x$ demonstrated strong response values and selectivity for acetone. Recently, Yao et al. also demonstrated a partially oxidized $Ti_3C_2T_x$ MXene-based sensor which showed a high response and excellent selectivity for NH_3.[99] The heat treatment of the detecting ceramic tube, which eliminates the adsorbed water and partially oxidizes the material at a temperature of 280° C, is an essential aspect of this work. The $Ti_3C_2T_x$

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

MXene-280 has a 147 % response to 500 ppm NH_3 in ambient conditions, whereas the equivalent response/recovery time are 67/157 s, respectively.[99]

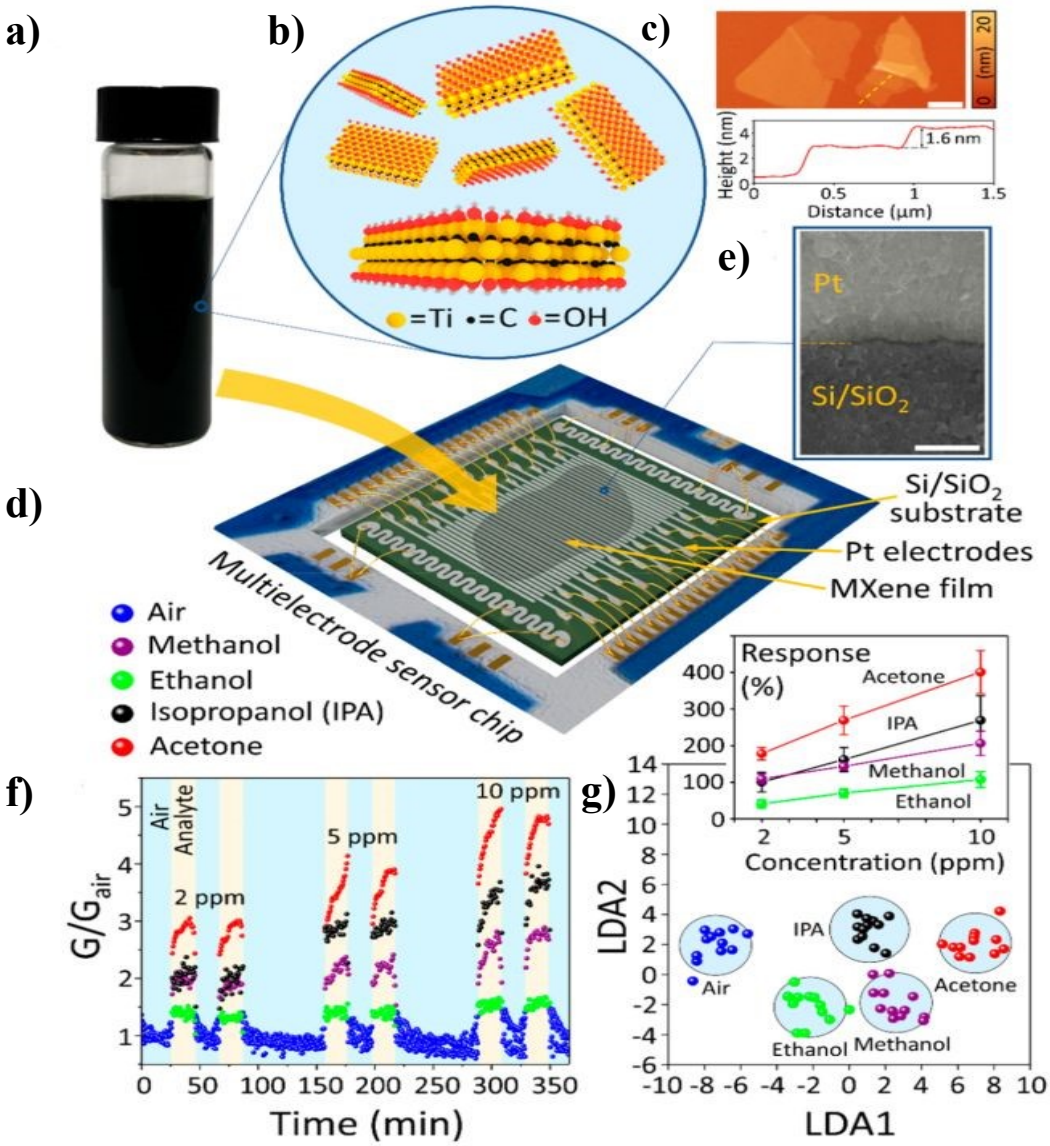


Figure 7: A partially oxidised $Ti_3C_2T_x$ MXene film's gas sensing capabilities on a multisensor chip: **a)** Image of a water-based, stable delaminated MXene solution, **b)** Monolayer $Ti_3C_2T_x$ MXene flakes' schematic structure, **c)** $Ti_3C_2T_x$MXene flakes on Si/SiO_2 as depicted in an AFM image with scale bar 1 μm, **d)** Schematic of a multielectrode chip with a coating of drop-cast MXene flakes, **e)** SEM image of a MXene film with contact area betweenSi/SiO_2 substrate and Pt electrode, **f)** Variation of MXene $G/(G_{air})$ with time, **g)** The multisensor array vector signals are anticipated by LDA into a two-component taste diagram for three organic vapours (10 ppm) and dry air; the circles associated to the gases are constructed with a 0.7 confidence around the relevant gravity centres (Main figure), Dependence of the chemiresistive response with

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

different concentration of gases (Inset), Reproduced with permission from reference [95] Copyright 2020. American Chemical Society

The electrochemical performance of the 2D $Ti_3C_2T_x$ sheet decorated with Pt nanoparticles (PtNPs) was studied by Lorencova et al..[100] The findings demonstrated that $Ti_3C_2T_x$/PtNP nanocomposite exhibited significantly superior and more stable redox activity in an anodic potential window compared to GCE treated with pure $Ti_3C_2T_x$ MXene. The H_2O_2 sensor fabricated with $Ti_3C_2T_x$ /PtNP on GCE provided LOD of 448 nM and a potential at which reduction begins of +250 mV (vs. Ag/AgCl), in contrast to 883 M and 160 mV observed for $Ti_3C_2T_x$ modified GCE. A flexible hydrogen sensor was also fabricated by Zhu et al., employing $Ti_3C_2T_x$ MXene that had been modified with the noble metal palladium (MXene@Pd CNC).[96] This sensor had a response time of (32 ± 7) s and a sensitivity range of S = (23 ± 4) % to 4 % H_2.

In summary, functionalized $Ti_3C_2T_x$ MXenes hold great promise for gas sensors due to their versatility and tunable surface chemistry, offering distinct advantages and potential drawbacks. One significant advantage is their enhanced selectivity and sensitivity towards specific gases. By introducing functional groups like -OH, -F, or -NH_2, the MXene surface can be tailored to interact selectively with target gas molecules, leading to improved detection capabilities for specific gases of interest. Moreover, functionalization can increase the active sites available for gas adsorption, further enhancing sensitivity. However, a notable drawback of excessive functionalization is the potential reduction in electrical conductivity. This can hinder electron flow within the sensor, impacting its overall performance. To overcome this limitation, researchers are actively working on optimizing the degree of functionalization to strike a balance between improved selectivity and maintaining good conductivity. Additionally, innovative techniques such as in situ functionalization and the development of hybrid structures that combine functionalized and pristine MXenes are being explored. These approaches aim to harness the benefits of functionalization while mitigating its negative impact on conductivity, thereby ensuring the gas sensor's overall effectiveness. Furthermore, the integration of functionalized MXenes into composite materials or heterostructure with other conductive materials like graphene or carbon nanotubes has shown promise in overcoming conductivity challenges while preserving the enhanced gas-sensing properties. In brief, functionalized $Ti_3C_2T_x$ MXenes offer tailored selectivity in gas sensors, but careful consideration of conductivity issues and the exploration of hybrid structures and composite materials are essential for optimizing their performance in practical applications.

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

6.3 $Ti_3C_2T_x$ composites-based Gas Sensor

A key tactic to enhance the gas sensing capabilities of $Ti_3C_2T_x$ is the formation of the composite. The diversity of materials, including metal oxides[101], TMDs[102], graphene[103], and polymers[104], have been effectively combined with $Ti_3C_2T_x$. The $Ti_3C_2T_x$ composite-based gas sensor attracts much attention because of the formation of a schottky barrier due to the combination of metal oxide and $Ti_3C_2T_x$, which improves the gas sensing performance of the device. To advance the NH_3 sensing properties of pure $Ti_3C_2T_x$, Tai et al. reported a gas sensor based on a bilayer film of TiO_2 and $Ti_3C_2T_x$.[81] In accordance with the findings, the $TiO_2/Ti_3C_2T_x$ sensor responds to 10 ppm NH_3 at room temperature with a response value that is 1.63 times greater and response/recovery times that are 0.65 and 0.52 times faster than those of a pure $Ti_3C_2T_x$ sensor as shown in figure 8 a) and b). The $TiO_2/Ti_3C_2T_x$ sensor, meanwhile, has an excellent linear response ($R^2 = 0.95596$) at NH_3 concentrations of 2 to 10 ppm and can detect NH_3 concentrations as low as 0.5 ppm. Sun et al. demonstrated how to synthesise $W_{18}O_{49}/Ti_3C_2T_x$ composites using an easy solvothermal technique to grow 1D $W_{18}O_{49}$ nanorods (NRs) on $Ti_3C_2T_x$ MXene sheets as shown in figure 8 c).[105] The reported $W_{18}O_{49}/Ti_3C_2T_x$ composites have excellent selectivity, enduring stability, a low LOD (170 ppb acetone), and quick response and recovery time (5.6 s and 6 s to 170 ppb acetone). They also reported a good response to low concentrations of acetone (11.6 to 20 ppm) and the operated temperature of the device was 300° C as shown in figure 8 d).

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

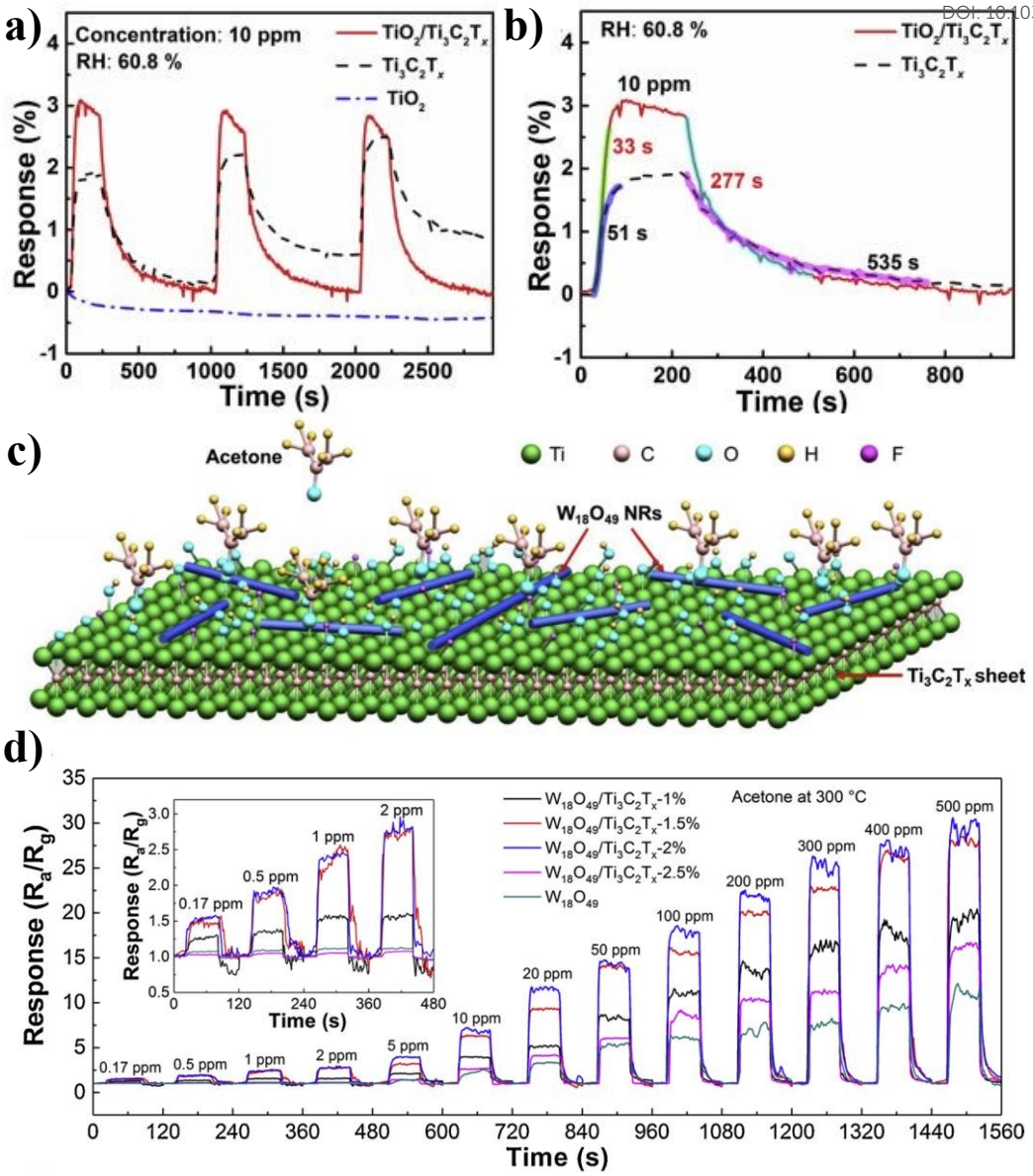


Figure 8: a) Response/recovery curve at a concentration of 10 ppm NH_3 for $TiO_2/Ti_3C_2T_x$, $Ti_3C_2T_x$, and TiO_2 gas sensors, **b)** Response and recovery times of the $Ti_3C_2T_x$ and $TiO_2/Ti_3C_2T_x$ gas sensors, Reproduced with permission from reference [81] Copyright 2019. Elsevier, **c)** Schematic of reaction between acetone and $Ti_3C_2T_x/W_{18}O_{49}$ composite and **d)** Response-recovery curves of different sensors at a concentration of 0.17–500 ppm acetone at 300 °C temperature, **Inset:** Response-recovery curves towards 0.17–2 ppm acetone (enlarged view) Reproduced with permission from reference [105] Copyright 2019. Elsevier

TMDs have received attention as prospective composite materials because of their numerous redox reaction active sites and high surface reactivity. To enhance the sensing capabilities of TMDs at ambient temperature, MXene can be combined with other TMDs, including In_2O_3, WS_2, SnO_2, graphene, CoS, WSe_2, and MOS_2, to generate heterojunctions.[39, 102, 106-110]

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

$Ti_3C_2T_x$ and WSe_2 were the model materials that Chen et al. decided to use in their implementation of the detection of various volatile organic compound (VOC) gases.[102] The nanohybrids gas sensor based on$Ti_3C_2T_x$/WSe_2wassynthesized by combining liquid-phase exfoliation $Ti_3C_2T_x$ with WSe_2 that had been functionalized with cetyltrimethyl ammonium bromide (CTAB), as shown in figure 9 a) and b). The gas sensor that was fabricated employing inkjet printing materials and a wireless monitoring system displayed a 12-fold improvement in ethanol sensitivity, very fast response and recovery time (9.7 s and 6.6 s), and a very low level of electrical noise. In addition, graphene and its derivatives have been the subject of a significant amount of research owing to the high surface area of these materials and the fact that their conductivity can be controlled. As a result, composites made of graphene and $Ti_3C_2T_x$ are deserving of future investigation in the realm of gas sensing. Yuan et al. have demonstrated a VOC sensor that is both high-performing and versatile, and it is built on a 3D MXene framework (3D-M).[111]The 3D-M was fabricated by combining a robust electrospinning method. The 3D-M sensor is able to function at ambient temperature thanks to its extremely interconnected porous structure. It is especially sensitive to extremely low concentrations of volatile organic compounds (ppb level) due to its capacity to quickly access and disperse gas molecules. In addition, the sensor has a very wide sensing range, a quick response and recovery time (less than two minutes), and excellent reversibility and versatility for a variety of volatile organic compounds. The researchers led by Lee et al. announced the fabrication of $Ti_3C_2T_x$ MXene/graphene hybrid fibres by a wet-spinning approach.[112] The remarkable mechanical and electrical qualities of these hybrid fibers make them ideally suited for use in flexible wearable gas sensors. The ability of the produced fibers to demonstrate significant sensitivity to NH_3 gas at ambient temperature is made possible by the synergistic impacts of the electrical characteristics and gas-adsorption capacities of MXene and graphene. In comparison to MXene and graphene on their own, the hybrid fibers demonstrated an NH_3 sensing response that was 6.77 % more sensitive ($\Delta R/R_0$ = 6.77 %). Even after being bent over a period of 2000 times, the hybrid fibres showed an exceptionally high degree of mechanical flexibility, with a resistance that changed by ± 0.2 % and a low level of noise resistance. This enabled the hybrid fibres to detect gas even as they were deformed, as shown in figure 9 c) and d). Recently, Yang et al. demonstrated the inherent process of the improvement of gas-sensing capabilities of oxidized $Ti_3C_2T_x$ by analyzing the performance of crumpled $Ti_3C_2T_x$ spheres.[113] This was done in order to show that the crumpled $Ti_3C_2T_x$ spheres have better gas-sensing properties. After explaining the structural and compositional modifications caused by the oxidation of $Ti_3C_2T_x$, it was found that the increase in gas-sensing capabilities was due to the presence of a

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

large number of Ti atom defects that had been introduced by the oxidation process. This was discovered after it was established that oxidation had brought about changes in structure and composition. Calculations using the first principles of density functional theory (DFT) were carried out, and the findings revealed that the adsorption capacity of Ti atom defect sites for gas molecules was significantly improved, particularly with respect to NO_2. Meanwhile, a reducing agent known as $Ti_3C_2T_x$ was utilized to assist in the process of GO reduction. The gas sensing performance was further improved by forming a $Ti_3C_2T_x/TiO_2$/rGO heterostructure, which resulted in high sensitivity, low level of detection limit, and a remarkable selectivity to NO_2, reaching up to 19.85 % for 5 ppm NO_2 at ambient temperature. All of these improvements were made possible by the formation of the heterostructure.

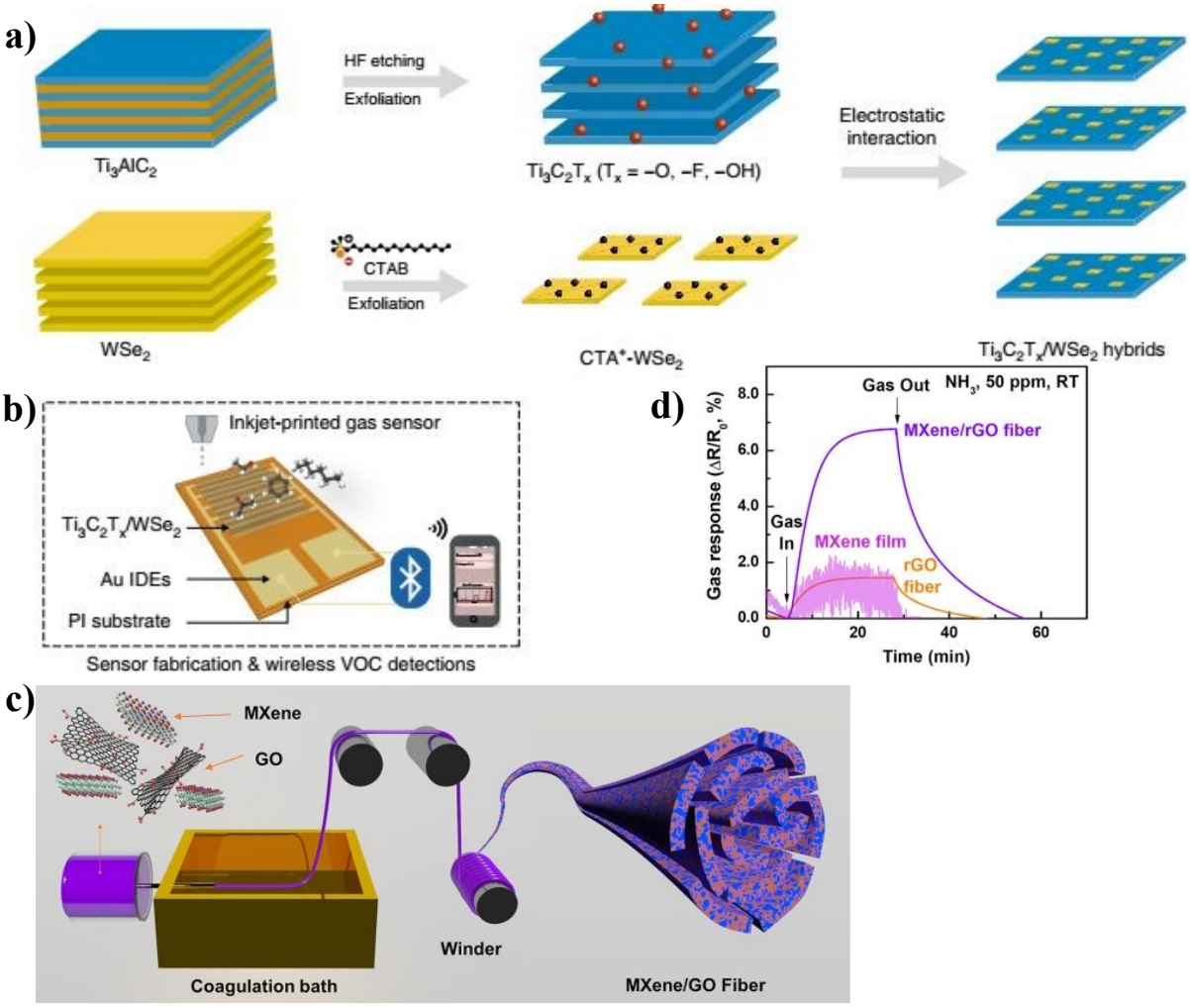


Figure 9: a) A schematic representation of the steps involved in the preparation of $Ti_3C_2T_x/WSe_2$ nanohybrids, **b)** Illustration of inkjet-printed gas sensors for the detection of VOC using a wireless monitoring system, Reproduced with permission from reference [102] Copyright 2020. Springer Nature (Open Access), **c)** Schematic of the spinning process for

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

MXene/GO hybrid fiber, and **d)** Comparison of the gas response of MXene film, rGO fiber, and MXene/rGO hybrid fiber (40 wt % MXene) Reproduced with permission from reference [112] Copyright 2020. American Chemical Society

Since polymers are sensitive, cheap, versatile, and may be used at room temperature, they have frequently been used to fabricate $Ti_3C_2T_x$ composites. Flexible gas sensors based on the $Ti_3C_2T_x$ / polyaniline (PANI) nanocomposite were studied by Zhao et al. using density functional theory in conjunction with electrosensitive experiments on large samples.[104] This composite sensing material showed a high sensitivity for ethanol (41.1 % at 200 ppm) and fast response/recovery time (0.4/0.5 second) at room temperature, as shown in figure 10 a) and b), which makes it the next-generation sensing material that can quickly detect real-time VOC gases. This is because the properties of composites and highly active $Ti_3C_2T_x$ MXene work synergistically with one another. To track NH_3 volatilization in farming operations, Li et al. report the development of a polyaniline (PANI)/$Ti_3C_2T_x$ hybrid sensitive film-based flexible chemiresistive gas sensor.[114] An in-situ method of self-assembly was utilized during the process of depositing the hybrid film onto the flexible polyimide substrate, as shown in figure 10 c). High sensitivity, excellent repeatability, a low detection limit, and strong air stability are all characteristics of this sensor, which is enabled by the gas-sensing enhancing effects of the PANI/$Ti_3C_2T_x$ schottky junction and the enhanced protonation degree of PANI in the hybrids. Particularly, the sensor demonstrates outstanding NH_3-sensing properties in situations with relative humidity (RH) of 20–80 % at a 10–40 °C temperature range. This makes it attractive for real-world agricultural applications because it can operate in a wide temperature range. Wang et al. investigated the possibility of using a conductive poly (3,4-ethylenedioxythiophene) doped with poly(styrene sulfonic acid) (PEDOT:PSS) composite $Ti_3C_2T_x$ in a gas sensing application, as shown in figure 10 d).[115] In comparison to the pristine materials, the $Ti_3C_2T_x$/PEDOT:PSS sensor, which was fabricated using the process of drop-coating, exhibited superior sensing behaviors when exposed to VOC vapors at ambient temperature. They also investigated the effect of composite ratio on gas sensing performance, and the findings revealed that the ideal mass ratio for PEDOT:PSS and $Ti_3C_2T_x$ was 4:1. This ratio produced the maximum response from both of these gas sensing components. Recently, Chen et al. reported a nanocomposite of $Ti_3C_2T_x$MXene and polypyrrole (PPy) by using *in-situ* polymerization reaction for gas sensing application.[116] The PPy/MXene-12, a nanocomposite-based gas sensor, displayed the best possible gas-detecting performance. It had a high response of 31.9 % to 100 ppm NH_3, a low detection limitation, and unfailing selectivity.

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

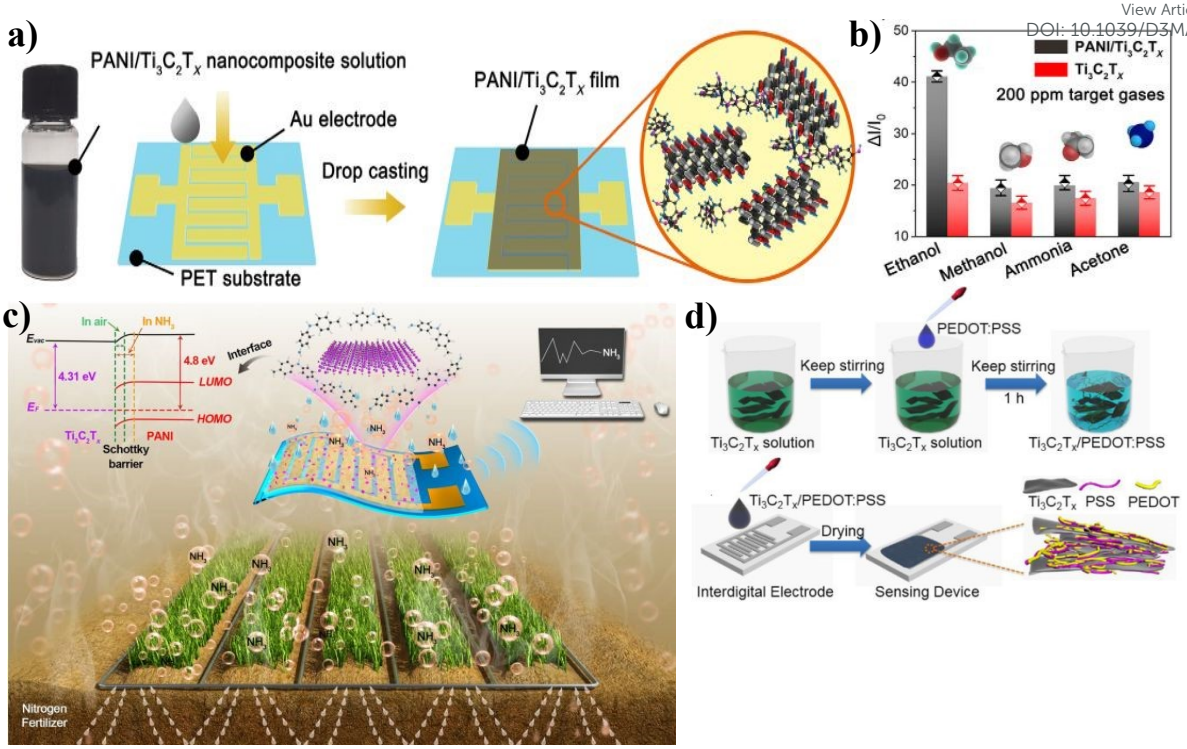


Figure 10: a) Interdigitated electrode schematic before and after coating with PANI/$Ti_3C_2T_x$ nanocomposites, **b)** Selectivity of $Ti_3C_2T_x$ and PANI/ $Ti_3C_2T_x$-based flexible sensors exposed to 200 ppm, Reproduced with permission from reference [104] Copyright 2019. John Wiley & Sons (Open Access), **c)** Agricultural NH_3 volatilization monitoring based on $Ti_3C_2T_x$/PANI NH_3 sensor, Reproduced with permission from reference [114] Copyright 2020. Elsevier, **d)** Schematic of fabrication of $Ti_3C_2T_x$/PEDOT:PSS materials and gas sensor devices, Reproduced with permission from reference [115] Copyright 2020. Elsevier

In summary, the $Ti_3C_2T_x$ composites-based gas sensors offer a range of advantages and drawbacks, necessitating innovative solutions for their optimal performance. One significant advantage is their improved conductivity and stability compared to pristine $Ti_3C_2T_x$, making them highly suitable for practical gas sensing applications. These composites often integrate $Ti_3C_2T_x$ with other materials like polymers, MOs, or carbon-based nanomaterials, harnessing the synergistic effects of each component. This enhances the sensor's electrical properties and provides a larger surface area for gas adsorption, leading to enhanced sensitivity and response times. However, a notable drawback is the potential loss of selectivity when using composites, as the additional materials may introduce cross-sensitivity to various gases. To address this, researchers are actively developing advanced composites with tailored properties that maintain or even improve selectivity while preserving high conductivity. Surface modification techniques and precise control over the composition and morphology of the composite

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

materials allow for a fine-tuning of gas-sensing characteristics. Another challenge lies in the potential degradation of the composites over time due to environmental factors or harsh operating conditions. To overcome this, protective coatings or encapsulation techniques are being explored to shield the sensitive composites from external influences, thus improving their long-term stability and reliability. Moreover, ongoing research into the optimization of composite fabrication processes and the development of novel composite combinations are continuously pushing the boundaries of gas sensor technology, aiming to strike a balance between enhanced performance and selectivity, as well as long-term durability for a wide range of gas detection applications.

Recently, lots of research going on $Ti_3C_2T_x$ MXene based self-powered gas sensor. One of the most remarkable aspects of MXene-based gas sensors is their ability to operate without the need for external power sources, such as batteries. This is achieved through the concept of self-powering, where the interaction between the gas molecules and the MXene-based sensing layer generates electricity. Two primary mechanisms have been explored for self-powering MXene-based gas sensors.

1. Piezoelectric Effect: When gas molecules adsorb or react with the MXene surface, they induce mechanical deformation or strain in the sensing layer. This strain can be harvested as piezoelectric energy, which is then converted into an electrical signal. This approach allows for the continuous and sustainable operation of gas sensors without the need for batteries.

2. Triboelectric Effect: In the triboelectric effect, friction or contact between the gas molecules and the MXene surface generates static electricity. This effect can also be harnessed to power gas sensors. As gas molecules interact with the MXene layer, they create charge imbalances that can be tapped into for sensing purposes.

Sardana et al. reported a self-powered humidity sensor using an electrospun anisotropic triboelectric nanogenerator (A-TENG) composed of aligned $Ti_3C_2T_x$ MXene nanofibers (NFs) and cellulose acetate (CA) NFs.[117] Significantly, anisotropic structure of TENG displays its potential in sensitive moisture detection upon fast adsorption/desorption of water molecules due to increased pore length of prepared NFs. They also demonstrated the applied application of detecting the skin moisture. The same group also reported, a new kind of MXene/TiO_2/C-NFs heterojunction-based sensory component for detection of NH_3 by using cellulose nanofibers (C-NFs) as a substrate. At room temperature, the proposed sensor shows excellent reproducibility, high selectivity, and sensitivity toward NH_3 (1–100 ppm) along with a fast

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

response/recovery time (76 s/62 s).[118] Wang et al. demonstrated self-powered and low-temperature resistant $Ti_3C_2T_x$ MXene-modified electronic-skin (e-skin) for multifunctional sensing.[119] The reported e-skin displays decent sensitivity to pressure (1.29 mV Pa^{-1}) and frequency (5.60 V Hz^{-1}) at a low temperature (−15 °C). In the meantime, utilizing this flexible e-skin, they successfully demonstrated temperature sensing over the range from −15 °C to 25 °C with a sensitivity of 76.6 nA cm^{-2} $°C^{-1}$. Gasso et al. reported $SnO_2/Ti_3C_2T_x$-MXene nanocomposite-based self-powered gas sensor for NO_2 sensing. The fabricated sensor has detection limit of 0.03 ppb NO_2 and power consumption as low as 1.2 μW. Moreover, they also reported SnO_2/MXene nanocomposite-based piezoelectric pressure sensor (PPs) with a high sensitivity of 2.088 kPa^{-1} under the pressure range of 1.63–6.23 kPa with a fast response time (265 ms) and recovery time (75.5 ms).[120] Wang et al. reported high-performance $Ti_3C_2T_x$ MXene/tungsten oxide nanofibers-based NO_2 sensor. The as fabricated self-powered $Ti_3C_2T_x$ MXene/WO_3 sensor driven by TENG has an outstanding response at room temperature (ΔUs/Usa = 510% at 50 ppm) for NO_2 gas, which is 15 times larger than that of the resistive MXene/WO_3 sensor.[121]

While MXene-based self-powered gas sensors have made significant strides, ongoing research is focused on further enhancing their sensitivity, selectivity, and stability. Additionally, efforts are underway to reduce production costs and facilitate large-scale manufacturing for wider adoption. In conclusion, MXene-based self-powered gas sensors are poised to play a pivotal role in shaping the future of gas detection technology. Their unique combination of material properties, self-powering mechanisms, and diverse applications make them a promising solution for addressing environmental, industrial, and healthcare challenges. As research in this field continues to evolve, we can anticipate even more sophisticated and accessible MXene-based gas sensors that contribute to a safer and healthier world.

7. MXene Gas Sensor: Theoretical Approach

The intricacy of the sensing mechanism employed by MXene sensors introduces considerable difficulties when attempting to fine-tune them through experimental methods. MXene sensors are equipped with a sophisticated and multifaceted system for detecting various environmental changes or target analytes, and this complexity can make it a formidable task to optimize their performance through practical experimentation. This complexity arises from the intricate interplay of materials and physical processes within the sensor, requiring a deep understanding of its underlying principles to effectively modify and enhance its sensing capabilities.

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Consequently, researchers face significant challenges in achieving the desired sensor performance due to the complicated nature of the MXene sensing mechanism, which necessitates a meticulous and thorough approach to experimentation and tuning. Hence, there exists a requirement for atomic-level modeling, specifically utilizing techniques such as density functional theory (DFT) and molecular dynamics (MD), to complement the empirical data and advance our understanding of the interactions between MXene materials and gas analytes.

The first computational study was conducted by Yu et al., who demonstrated the potential of MXenes as a gas sensor using monolayer Ti_2CO_2.[122] They investigated the interaction behavior of H_2, N_2, CO_2, CO, NH_3, O_2, NO_2, and CH_4 gas molecules on the semiconductor Ti_2CO_2. Because Ti_2CO_2 is the only one among its counterparts [Ti_2C, Ti_2CF_2, and $Ti_2C(OH)_2$] that exhibits semiconductor properties, the researchers focused their investigation on Ti_2CO_2 through first-principles simulations. Out of all the gases studied, NH_3 is the sole gas that chemisorbs onto the surface of the Ti_2CO_2 MXene. It accomplishes this by donating a relatively substantial charge of 0.174 e and establishing a N−Ti bond. In DOS analysis, there is a significant overlap of electronic charges, leading to orbital mixing and substantial charge transfer between NH_3 and Ti_2CO_2. This observation strongly indicates chemisorption, where chemical bonds are formed between NH_3 and Ti_2CO_2. Conversely, when examining other gas molecules, it was evident that they exhibited significantly lower gas adsorption energies and charge transfer compared to NH3. This points to the absence of overlapping electronic charges and signifies weak interactions between these molecules and the Ti_2CO_2 surface. The adsorption energy of NH_3 on Ti_2CO_2 was determined to be -0.37 electron volts (eV), which is quite less as compared to $Ti_2C(OH)_2$ [−0.48 eV], V_2CO_2 [−0.81 eV] and analogous to $Ti_3C_2O_2$ [−0.34]. The higher sensitivity of Ti_2CO_2 toward NH_3 was confirmed by calculating I-V curves before and after absorption of gas. Xiao et al. investigated the interactions of various gases, including NO, CO, H_2, CO_2, O_2, N_2, and CH_4, with Zr_2CO_2 MXene.[123] Their findings revealed that NH_3 exhibited strong chemisorption characterized by a noticeable charge transfer 0.188 e and a substantial adsorption energy -0.81 eV, whereas other gases were primarily physisorbed. The authors anticipated a similar response for M_2CO_2 (where M = Sc, Ti, Zr, and Hf) due to their analogous atomic and electronic structures. Additionally, it was observed that the transition from chemisorption to physisorption in NH_3 adsorption on Zr_2CO_2 occurred when two additional electrons were introduced into the MXene sheet, offering a practical means to release the gas molecule. In a subsequent study, Wang et al. conducted a comprehensive analysis of Hf_2CO_2's performance in detecting multiple gases, including H_2, O_2, NO_2, SO_2, CO,

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

CO_2, HCN, NH_3, and H_2S.[124] Consistent with the earlier report, Hf_2CO_2 exhibited remarkable sensitivity and selectivity toward NH_3, characterized by an exceptional charge transfer of 0.146 e and substantial adsorption energy of -0.834 eV. Additionally, it was observed that preadsorption of the MXene surface with H_2O, SO_2, and CO_2 molecules significantly enhanced NH_3 adsorption by fostering the formation of hydrogen bonds.

In addition to their experimental investigations, Kim et al. conducted a detailed examination of how the functional groups of $Ti_3C_2T_x$ MXene influence its sensing capabilities.[74] Employing the DFT approach, they analyzed the interactions between acetone and NH_3 with $Ti_3C_2T_x$ surfaces terminated by -OH, -O, and -F groups. Their findings indicated that $Ti_3C_2(OH)_2$ exhibited stronger binding energy with both acetone and NH_3 compared to $Ti_3C_2O_2$ and $Ti_3C_2F_2$ MXenes. Hajian et al. investigated the impact of the ratio of functional groups on NH_3 detection using $Ti_3C_2T_x$ MXene.[125] They examined two different ratios of -F functional groups in and observed that lower ratios of -F groups led to stronger NH_3 adsorption. This effect was attributed to the smaller charge transfer between NH_3 and fluorine in comparison to NH_3 and oxygen. Consequently, the synthesis method and the choice of MAX phase etchant, which dictate the functional group ratios, emerged as critical factors influencing the gas-sensing performance of MXene sensors. Zhang et al. examined the ability of $Ti_3C_2O_2$ nanosheets to capture formaldehyde effectively at room temperature.[126] The findings revealed a moderate adsorption energy of 0.45 eV and an adsorption capacity exceeding 6 mmol/g, indicating the promising potential of $Ti_3C_2O_2$ MXene for indoor formaldehyde removal. Ma et al. explored M_2CO_2 (where M = Sc, Ti, Zr, and Hf) MXenes to identify the most effective platforms for detecting toxic SO_2.[127] Their investigation revealed that monolayer Sc_2CO_2 exhibited a robust chemical bond with SO_2, characterized by an adsorption energy of -0.646 eV and a charge transfer of 0.453 e, establishing it as a favorable material for SO_2 sensing. The M_2CO_2 MXene was also explored for other gas detection such as NO and CO. Regarding NO, the most significant interactions were observed between the molecule and Sc_2CO_2, likely due to orbital hybridization occurring between the MXene and the gas molecule near the Fermi level. This was substantiated by a substantial charge transfer of 0.303 e and a reasonably moderate adsorption energy of -0.47 eV, affirming Sc_2CO_2 as a highly suitable material for NO sensing. Conversely, in the case of CO, it was primarily physisorbed onto Sc_2CO_2 MXene, with a minimal charge transfer of 0.017 e. However, the sensing capabilities of Sc_2CO_2 MXene for CO were enhanced through the introduction of Mn dopants into the structure, resulting in a charge transfer of 0.199 e and a strong adsorption energy of -0.85 eV.[128] Banu et al. utilized

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

M_2CO_2 MXene for the detection of NH_3 via oxygen functionalization.[129] Based on the detailed DFT study, they reported the adsorption behaviour of NH_3 on Cr_2C/Fe_2C went through the alteration from chemisorption (E_{ad} = −1.40/−1.33 eV) to physisorption (E_{ad} = −0.29/−0.22 eV). Recently Quan et al. reported fully flexible $Ti_3C_2T_x/WS_2$ based gas sensor for NO_2 detection at room temperature.[130] To delve further into the sensing mechanism, adsorption models for NO_2 on WS_2 (001), $Ti_3C_2T_x$ (001), and $Ti_3C_2T_x$ (001)/WS_2 (001) were computed using the DFT method. The calculated adsorption energies for NO_2 molecules on the surfaces of bare WS_2 (001), $Ti_3C_2T_x$ (001), and $Ti_3C_2T_x$ (001)/WS_2 (001) were found to be -0.20 eV, -3.37 eV, and -3.73 eV, respectively. Weng et al. reported highly sensitive and selective $Ti_3C_2O_2$ -based gas sensors for NH_3 sensing.[131] Through DFT study, they calculated adsorption energy of NH_3 on $Ti_3C_2O_2$ yields an appropriate value (−0.42 eV), and the charge transfer rate of 0.18 e. Based on their theoretical results, only NH_3 adsorbs onto the nanosheet through chemisorption. Reji et al. demonstrated the Sc_2CO_2 MXene nanosheets for detection of VOCs using first-principles DFT calculations.[132] They reported that 2-propanol (−0.63 eV), ethanol (−0.59 eV), and acetonitrile (−0.51 eV) show higher interaction with the Sc_2CO_2 MXene nanosheets. At the end, they also predict that Sc_2CO_2 MXene nanosheets can be utilized as a dual-mode sensor to detect potential VOC biomarkers in the exhaled breath.

8. Advantage and Drawback of $Ti_3C_2T_x$ MXene gas sensors and other Common Materials

$Ti_3C_2T_x$ MXene is a promising material for gas sensors, but like any material, it has its advantages and drawbacks when compared to common materials used in gas sensing applications. The comparison of $Ti_3C_2T_x$ MXene with some common materials is given in Table-1.

Table 1: Advantage and Drawback of $Ti_3C_2T_x$ MXene with some Common Material

S.N.	Materials	Advantages	Drawbacks
1.	**$Ti_3C_2T_x$ MXene based gas senor**	✓ High sensitivity ✓ Tunable properties ✓ Good conductivity ✓ Chemical stability	➢ Complex synthesis ➢ Limited gas selectivity ➢ Poor humidity sensitivity
2.	**Metal Oxide-based gas Sensors (e.g., SnO_2 and ZnO)**	✓ Low cost ✓ Commercial availability ✓ Good selectivity	➢ High operating temperature ➢ Slow response/recovery ➢ Limited sensitivity

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

3.	**Organic Polymer-based Gas Sensors**	✓ Low cost ✓ Low power consumptions ✓ Tailorability	➢ Limited sensitivity ➢ Limited durability ➢ Limited operating Range
Other 2D Materials (e.g., Graphene, TMDs like MoS_2)			
4.	**Graphene**	✓ Excellent electrical conductivity ✓ Low noise levels ✓ High surface area	➢ Limited gas selectivity. ➢ Difficulty in functionalizing for specific gas detection ➢ Susceptibility to environmental factors, such as humidity
5.	**TMDs**	✓ Good gas selectivity due to varying band structures ✓ Layer-dependent properties for tenability ✓ Compatibility with existing semiconductor technologies.	➢ Moderate surface area compared to MXenes. ➢ Synthesis Challenges ➢ Sensitivity to environmental conditions, including humidity.

In summary, $Ti_3C_2T_x$ MXene-based gas sensors offer unique advantages such as high surface area, tunability, and fast response times. However, they also face challenges related to synthesis complexity, moisture sensitivity, and limited gas range. The choice of gas sensor material depends on the specific application and requirements for sensitivity, selectivity, and operational conditions.

Our detailed study suggested that $Ti_3C_2T_x$ MXene-based gas sensors have a high response value for NH_3 gas. The response of NH_3 gas based on $Ti_3C_2T_x$ MXene with different literature has been shown in figure 11.

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

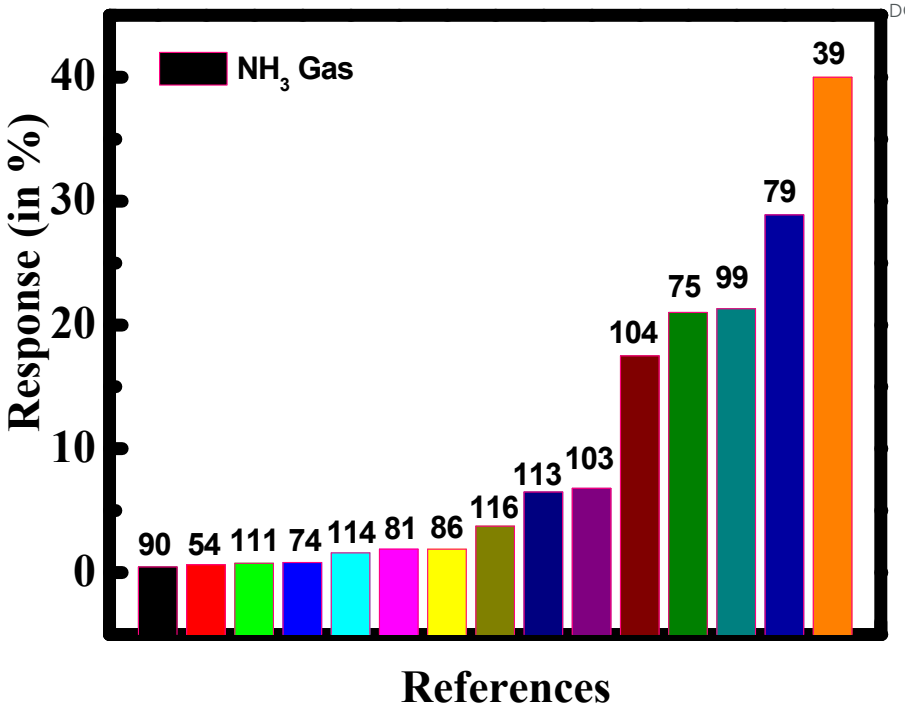


Figure 11. Gas response of NH_3 gas with $Ti_3C_2T_x$ MXene-based gas sensors

We have summarised the different types of $Ti_3C_2T_x$ MXene (pure, functionalized and compoite) based gas sensors in terms of sensitivity, limit of detection and working range in Table-2.

Table 2: Summary of $Ti_3C_2T_x$ MXene-based gas sensor

MXene	MXene Type	Target Gas	Sensitivity in %	Limit of Detection (LOD)	Working Range	References
	Pure	Ammonia	0.21	NR	25 to 200 ppm	[75]
	Pure	Acetone	0.97	0.011 ppb	50 ppb to 100 ppm	[74]
	Pure	Nitrogen Dioxide	4.5	NR	5 ppm	[91]
	Pure	Ammonia	0.62	NR	5 ppm	[54]

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

$Ti_3C_2T_x$	Pure	Ammonia	6.13	NR	10 to 700 ppm	[86]
	Pure	Ammonia	1.22	NR	1 to 100 ppm	[133]
	Pure	Methane	68	NR	200 to 10000 ppm	[134]
	Pure	Toluene	19.32	NR	1 to 50 ppm	[94]
	Functionalized	Ethanol	9.99	NR	0.3 mM/L NaOH	[93]
	Functionalized	Hydrogen	23 ± 4	NR	0.5 to 40 % H_2	[96]
	Functionalized	Toluene	79.5	NR	1 to 50 ppm	[94]
	Functionalized	Ammonia	28.87	NR	10 to 500 ppm	[79]
	Composite	Ethanol	14	NR	5 to 120 ppm	[135]
	Composite	Toluene	79.5	NR	1 to 50 ppm	[94]
	Composite	Hydrogen	70	NR	100 to 500 ppm	[136]
	Composite	Ammonia	6.2	NR	1 to 80 ppm	[137]
	Composite	Ammonia	3.1	NR	10 ppm	[81]
	Composite	Ammonia	0.4	NR	1 to 100 ppm	[138]
	Composite	Ammonia	24.8	NR	1 to 100 ppm	[139]
	Composite	Ammonia	22.3	NR	1 to 5 ppm	[80]
	Composite	Ammonia	40	4.29 ppb	0.5 to 100 ppm	[39]
	Composite	Ammonia	100.7	NR	20 ppm	[140]

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

Composite	Toluene	11.4	NR	10 to 50 ppm	[141]
Composite	Nitrogen Dioxide	41.93	NR	5 to 100 ppm	[142]
Composite	Nitrogen Dioxide	37	125 ppb	0.125 to 5 ppm	[101]
Composite	Nitrogen Dioxide	165	50 ppb	0.05 to 20 ppm	[143]
Composite	Acetone	0.25	0.05 ppm	0.05 to 30 ppm	[144]
Composite	Acetone	52	NR	10 to 500 ppm	[145]
Composite	Ethanol	27.4	NR	50 to 200 ppm	[104]
Composite	Carbon dioxide	1	NR	8 to 3000 ppm	[146]
Composite	Hexanal	3.4	217 ppb	10 to 40 ppm	[147]
Composite	Methanol	29.6	NR	5 to 100 ppm	[106]

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

Conclusions and future prospective

MXenes are a newly discovered family of 2D materials that have been the subject of intensive research due to the exceptional properties that make them exceedingly valuable for use in various applications. MXene-based sensors have been among the most extensively researched applications because they are able to exploit the full potential of the one-of-a-kind 2D structure of MXenes. This is because MXenes have an inherently large specific surface area and surface-mediated properties that can be controlled. Thus, investigations on MXene-based gas sensors have been enhancing the selectivity and sensitivity using approaches such as composite formation, surface functionalization, and macro-structuring. The fact that the alternation of the MXene surface does not forfeit its fundamental outstanding electrical conductivity and improves its other physical and chemical properties makes it one of the most promising materials for sensors. For instance, it is conceivable to boost the adsorption energy of analytes

View Article Online
DOI: 10.1039/D3MA00631J

on the MXene surface while maintaining the material's conductivity. The aforementioned methods have proven to be successful in the majority of the research aimed at enhancing the performance of sensors.

The development of MXene-based sensors is still in its early stages; therefore, there are still issues that need to be resolved before they can be used commercially. There needs to be more information about these requirements because the performance standards and other requirements for putting these sensors on the market have not yet been developed in a systematic way. In this review, we have discussed the recent developments in the field of gas sensors based on 2D $Ti_3C_2T_x$ from three different perspectives: etching and delamination of $Ti_3C_2T_x$; pure, functional, and composite $Ti_3C_2T_x$ applied in gas sensors; and their sensing mechanism. The $Ti_3C_2T_x$ MXene has garnered a lot of attention and has been demonstrated to offer a lot of potential in the field of gas sensor. This is because of their one-of-a-kind structure and features. However, the application of $Ti_3C_2T_x$ in the real world is constrained by a number of different reasons. The first difficulty is associated with the process of preparing the MXene precursor, also known as the MAX phase. In most cases, the formation of a MAX phase occurs as a result of the high-temperature processing of combustible materials in powder forms, such as titanium, aluminum, and others. To obtain a MAX powder of the desired fineness, multiple milling processes are necessary. If HF is to be used as an etchant, the handling of HF and treatment of waste must be done correctly. The second obstacle is the mass production of MXene-based sensors, which is difficult because MXene does not have sufficient mechanical stability to be processed by industrial machines. In order to overcome this obstacle, researchers have been concentrating on the fabrication of nanocomposites by combining MXenes with a variety of other materials, particularly polymers. The third obstacle is determining the ideal proportion of MXene to incorporate into a polymer matrix in order to produce a high-performance sensor. The fourth obstacle pertains to applications that take place in the real world and require sensors to be durable and functional even after being damaged. This can be accomplished by modifying the surface of the components that make up a sensor, which in most cases results in improved mechanical properties of those components. The fifth obstacle is the requirement for clean and flawless MXene surfaces, which increases the sensitivity of MXene-based sensors. It is necessary to develop etchants that are superior to both HF and HCl/LiF. In general, the focus of research efforts on this subject at the moment is on exploring the maximal capability and potential of $Ti_3C_2T_x$ MXenes in terms of transferring their advantageous properties to sensors in order to obtain high performance of sensitivity and LOD.

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

These aspects are absolutely necessary for the successful commercialization of MXene-based sensors. However, in order for MXenes to be successfully employed in the industry, several practical characteristics, including productivity, stability, and multifunctionality, should be examined and improved.

As a result, the area of future research should be broadened, and the primary objective should be to enhance sensor performance in terms of the variables described above so that it is compatible with industry requirements. In order to speed up the development of next-generation sensor technologies based on MXenes, it will be necessary to address the requirements listed above for commercialization. This will eventually pave the way for the creation of low-cost, high-performance, and multimodal sensors that can be used in soft-electronics applications. We hope this work will provide a roadmap for the commercialization of $Ti_3C_2T_x$ MXene-based gas sensors and encourage a wide range of in-depth and applicable studies in this area.

Acknowledgment

This work was supported by "**Science and Engineering Research Board**,". Nitesh K. Chourasia acknowledges "Science and Engineering Research Board'' for National Post-Doctoral Fellowship (**PDF/2021/001490**).

Conflict of interest-

All authors declare that they have no conflicts of interest. We certify that the submission is original work and is not under review at any other publication.

References

[1] Zhang H, Wang L, Zou Y, Li Y, Xuan J, Wang X, et al. Enhanced ammonia sensing response based on Pt-decorated Ti3C2T x/TiO2 composite at room temperature. Nanotechnology 2023;34:205501.

[2] Lai QT, Zhao XH, Sun QJ, Tang Z, Tang XG, Roy VA. Emerging MXene-Based Flexible Tactile Sensors for Health Monitoring and Haptic Perception. Small 2023:2300283.

[3] Vishnuraj R, Aleem M, Nair KG, Pullithadathil B. 1D aligned, n–p and n–n type ZnO heterojunction nanofibers for NO 2 sensors: exploration of conduction mechanism using in situ impedance spectroscopy. Materials Advances 2023;4:3010-25.

[4] Kong X, Tarakanova EN, Du X, Tomilova LG, Chen Y. Discrimination and detection of NO 2, NH 3 and H 2 S using sensor array based on three ambipolar sandwich

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

tetradiazepinoporphyrazinato/phthalocyaninato europium double-decker complexes. Materials Advances 2023;4:1515-22.

[5] Song H, Liu Y, Fang Y, Zhang D. Carbon-based electrochemical sensors for in vivo and in vitro neurotransmitter detection. Critical Reviews in Analytical Chemistry 2023;53:955-74.

[6] Vivaldi FM, Dallinger A, Bonini A, Poma N, Sembranti L, Biagini D, et al. Three-dimensional (3D) laser-induced graphene: structure, properties, and application to chemical sensing. ACS Applied Materials & Interfaces 2021;13:30245-60.

[7] Qin P, Okur S, Li C, Chandresh A, Mutruc D, Hecht S, et al. A photoprogrammable electronic nose with switchable selectivity for VOCs using MOF films. Chemical Science 2021;12:15700-9.

[8] Yeung H-M, Yoshikawa G, Minami K, Shiba K. Strain-based chemical sensing using metal–organic framework nanoparticles. Journal of Materials Chemistry A 2020;8:18007-14.

[9] Chen J, Lv H, Bai X, Liu Z, He L, Wang J, et al. Synthesis of hierarchically porous Co3O4/Biomass carbon composites derived from MOFs and their highly NO2 gas sensing performance. Microporous and Mesoporous Materials 2021;321:111108.

[10] Li X, Zhang Y, Cheng Y, Chen X, Tan W. MOF-derived porous hierarchical ZnCo2O4 microflowers for enhanced performance gas sensor. Ceramics International 2021;47:9214-24.

[11] Chourasia NK, Singh AK, Rai S, Sharma A, Chakrabarti P, Srivastava A, et al. A Lithography-Free Fabrication of Low-Operating Voltage-Driven, Very Large Channel Length Graphene Field-Effect Transistor With NH 3 Sensing Application. IEEE Transactions on Electron Devices 2020;67:4385-91.

[12] Chourasia NK, Singh VK, Sharma A, Srivastava A, Pal BN. Lithography-free fabrication of low operating voltage and large channel length graphene transistor with current saturation by utilizing Li+ of ion-conducting-oxide gate dielectric. AIP Advances 2020;10:085313.

[13] Lei G, Lou C, Liu X, Xie J, Li Z, Zheng W, et al. Thin films of tungsten oxide materials for advanced gas sensors. Sensors and Actuators B: Chemical 2021;341:129996.

[14] Sharma A, Chourasia NK, Sugathan A, Kumar Y, Jit S, Liu S-W, et al. Solution processed Li 5 AlO 4 dielectric for low voltage transistor fabrication and its application in metal oxide/quantum dot heterojunction phototransistors. Journal of Materials Chemistry C 2018;6:790-8.

[15] Chourasia NK, Sharma A, Acharya V, Pal N, Biring S, Pal BN. Solution processed low band gap ion-conducting gate dielectric for low voltage metal oxide transistor. Journal of Alloys and Compounds 2019;777:1124-32.

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

[16] Pal N, Sharma A, Acharya V, Chourasia NK, Biring S, Pal BN. Gate interface engineering for subvolt metal oxide transistor fabrication by using ion-conducting dielectric with Mn2O3 gate interface. ACS Applied Electronic Materials 2019;2:25-34.

[17] Sharma A, Chourasia NK, Pal N, Biring S, Pal BN. Role of electron donation of TiO2 gate interface for developing solution-processed high-performance one-volt metal-oxide thin-film transistor using ion-conducting gate dielectric. The Journal of Physical Chemistry C 2019;123:20278-86.

[18] Acharya V, Sharma A, Chourasia NK, Pal BN. Solution-processed Pb0. 8Ba0. 2ZrO3 as a gate dielectric for low-voltage metal-oxide thin-film transistor. Emergent Materials 2020;3:57-62.

[19] Chourasia NK, Sharma A, Pal N, Biring S, Pal BN. Dielectric/Semiconductor Interfacial p-Doping: A New Technique to Fabricate Solution-Processed High-Performance 1 V Ambipolar Oxide Transistors. physica status solidi (RRL)–Rapid Research Letters 2020;14:2000268.

[20] Sharma A, Chourasia NK, Acharya V, Pal N, Biring S, Liu S-W, et al. Ultra-low voltage metal oxide thin film transistor by low-temperature annealed solution processed LiAlO2 gate dielectric. Electronic Materials Letters 2020;16:22-34.

[21] Singh AK, Chourasia NK, Pal BN, Pandey A, Chakrabarti P. Low Operating Voltage Solution Processed (Li_2ZnO_2) Dielectric and (SnO_2) Channel-Based Medium Wave UV-B Phototransistor for Application in Phototherapy. IEEE Transactions on Electron Devices 2020;67:2028-34.

[22] Singh AK, Chourasia NK, Pal BN, Pandey A, Chakrabarti P. A Proposed All ZnO Based Thin Film Transistor For UV-B Detection. IEEE Photonics Technology Letters 2020;32:1548-51.

[23] Chourasia NK, Pal BN. Solution-processed photodetectors. Chemical Solution Synthesis for Materials Design and Thin Film Device Applications: Elsevier; 2021. p. 649-64.

[24] Pandey U, Chourasia NK, Pal N, Biring S, Pal BN. Functional Dielectric Properties of Solution-Processed Lithium Indium Tin Oxide ($LiInSnO_4$) and Its Application as a Gate Insulator of a Low Voltage Thin Film Transistor. IEEE Transactions on Electron Devices 2022;69:1077-82.

[25] Saini R, Mahajan A, Bedi R, Aswal D, Debnath A. Phthalocyanine based nanowires and nanoflowers as highly sensitive room temperature Cl 2 sensors. RSC Advances 2014;4:15945-51.

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

[26] Chaloeipote G, Prathumwan R, Subannajui K, Wisitsoraat A, Wongchoosuk C. 3D printed CuO semiconducting gas sensor for ammonia detection at room temperature. Materials Science in Semiconductor Processing 2021;123:105546.

[27] Wang Y, Zhang S, Huang C, Qu F, Yao D, Guo H, et al. Mesoporous WO 3 modified by Au nanoparticles for enhanced trimethylamine gas sensing properties. Dalton Transactions 2021;50:970-8.

[28] Sun H, Yao Y, Yang R, Yan Z, Cao C, Deng Y, et al. A ZnO Gas Sensor with an Abnormal Response to Hydrogen. Energies 2023;16:5847.

[29] Kumar S, Sharma AK, Sohal MK, Sharma DP, Debnath A, Aswal D, et al. Room temperature highly sensitive chlorine sensor based on reduced graphene oxide anchored with substituted copper phthalocyanine. Sensors and Actuators B: Chemical 2021;327:128925.

[30] Naguib M, Kurtoglu M, Presser V, Lu J, Niu J, Heon M, et al. Two-dimensional nanocrystals produced by exfoliation of Ti3AlC2. Advanced materials 2011;23:4248-53.

[31] Dhamodharan D, Dhinakaran V, Byun H-S. MXenes: An emerging 2D material. Carbon 2022;192:366-83.

[32] Ho DH, Choi YY, Jo SB, Myoung JM, Cho JH. Sensing with MXenes: progress and prospects. Advanced Materials 2021;33:2005846.

[33] Ling Z, Ren CE, Zhao M-Q, Yang J, Giammarco JM, Qiu J, et al. Flexible and conductive MXene films and nanocomposites with high capacitance. Proceedings of the National Academy of Sciences 2014;111:16676-81.

[34] Chen H, Wen Y, Qi Y, Zhao Q, Qu L, Li C. Pristine titanium carbide MXene films with environmentally stable conductivity and superior mechanical strength. Advanced Functional Materials 2020;30:1906996.

[35] Rafieerad A, Yan W, Sequiera GL, Sareen N, Abu-El-Rub E, Moudgil M, et al. Application of Ti3C2 MXene quantum dots for immunomodulation and regenerative medicine. Advanced healthcare materials 2019;8:1900569.

[36] Shahzad F, Iqbal A, Kim H, Koo CM. 2D transition metal carbides (MXenes): applications as an electrically conducting material. Advanced Materials 2020;32:2002159.

[37] Yang Q, Wang Y, Li X, Li H, Wang Z, Tang Z, et al. Recent Progress of MX ene-Based Nanomaterials in Flexible Energy Storage and Electronic Devices. Energy & Environmental Materials 2018;1:183-95.

[38] Zeng R, Wang W, Chen M, Wan Q, Wang C, Knopp D, et al. CRISPR-Cas12a-driven MXene-PEDOT: PSS piezoresistive wireless biosensor. Nano Energy 2021;82:105711.

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

[39] He T, Liu W, Lv T, Ma M, Liu Z, Vasiliev A, et al. MXene/SnO2 heterojunction based chemical gas sensors. Sensors and Actuators B: Chemical 2021;329:129275.

[40] Yang L, Wang H, Yuan W, Li Y, Gao P, Tiwari N, et al. Wearable pressure sensors based on MXene/tissue papers for wireless human health monitoring. ACS Applied Materials & Interfaces 2021;13:60531-43.

[41] Chen Y, Ge Y, Huang W, Li Z, Wu L, Zhang H, et al. Refractive index sensors based on Ti3C2Tx MXene fibers. ACS Applied Nano Materials 2020;3:303-11.

[42] Guo F, Feng C, Zhang Z, Zhang L, Xu C, Zhang C, et al. A room-temperature NO2 sensor based on Ti3C2TX MXene modified with sphere-like CuO. Sensors and Actuators B: Chemical 2023;375:132885.

[43] Li X-P, Li Y, Li X, Song D, Min P, Hu C, et al. Highly sensitive, reliable and flexible piezoresistive pressure sensors featuring polyurethane sponge coated with MXene sheets. Journal of colloid and interface science 2019;542:54-62.

[44] Hong W, Wyatt BC, Nemani SK, Anasori B. Double transition-metal MXenes: Atomistic design of two-dimensional carbides and nitrides. Mrs Bulletin 2020;45:850-61.

[45] Kumar AN, Pal K. Amine-functionalized stable Nb 2 CT x MXene toward room temperature ultrasensitive NO 2 gas sensor. Materials Advances 2022;3:5151-62.

[46] Deysher G, Shuck CE, Hantanasirisakul K, Frey NC, Foucher AC, Maleski K, et al. Synthesis of Mo4VAlC4 MAX phase and two-dimensional Mo4VC4 MXene with five atomic layers of transition metals. ACS nano 2019;14:204-17.

[47] Alhabeb M, Maleski K, Anasori B, Lelyukh P, Clark L, Sin S, et al. Guidelines for synthesis and processing of two-dimensional titanium carbide (Ti3C2T x MXene). Chemistry of Materials 2017;29:7633-44.

[48] Meng W, Liu X, Song H, Xie Y, Shi X, Dargusch M, et al. Advances and challenges in 2D MXenes: From structures to energy storage and conversions. Nano Today 2021;40:101273.

[49] Gupta M, Verma A, Chaudhary P, Yadav B. MXene and their integrated composite-based acetone sensors for monitoring of diabetes. Materials Advances 2023.

[50] Pei Y, Zhang X, Hui Z, Zhou J, Huang X, Sun G, et al. Ti3C2TX MXene for sensing applications: recent progress, design principles, and future perspectives. ACS nano 2021;15:3996-4017.

[51] Sun Z, Music D, Ahuja R, Li S, Schneider JM. Bonding and classification of nanolayered ternary carbides. Physical Review B 2004;70:092102.

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

[52] Mashtalir O, Naguib M, Mochalin VN, Dall'Agnese Y, Heon M, Barsoum MW, et al. Intercalation and delamination of layered carbides and carbonitrides. Nature communications 2013;4:1-7.

[53] Li M, Lu J, Luo K, Li Y, Chang K, Chen K, et al. Element replacement approach by reaction with Lewis acidic molten salts to synthesize nanolaminated MAX phases and MXenes. Journal of the American Chemical Society 2019;141:4730-7.

[54] Shuck CE, Han M, Maleski K, Hantanasirisakul K, Kim SJ, Choi J, et al. Effect of Ti3AlC2 MAX phase on structure and properties of resultant Ti3C2T x MXene. ACS Applied Nano Materials 2019;2:3368-76.

[55] Halim J, Lukatskaya MR, Cook KM, Lu J, Smith CR, Näslund L-Å, et al. Transparent conductive two-dimensional titanium carbide epitaxial thin films. Chemistry of Materials 2014;26:2374-81.

[56] Ghidiu M, Lukatskaya MR, Zhao M-Q, Gogotsi Y, Barsoum MW. Conductive two-dimensional titanium carbide 'clay'with high volumetric capacitance. Nature 2014;516:78-81.

[57] Lipatov A, Alhabeb M, Lukatskaya MR, Boson A, Gogotsi Y, Sinitskii A. Effect of synthesis on quality, electronic properties and environmental stability of individual monolayer Ti3C2 MXene flakes. Advanced Electronic Materials 2016;2:1600255.

[58] Nguyen V-H, Nguyen B-S, Hu C, Nguyen CC, Nguyen DLT, Nguyen Dinh MT, et al. Novel architecture titanium carbide (Ti3C2Tx) MXene cocatalysts toward photocatalytic hydrogen production: a mini-review. Nanomaterials 2020;10:602.

[59] Halim J, Palisaitis J, Lu J, Thörnberg J, Moon E, Precner M, et al. Synthesis of two-dimensional Nb1. 33C (MXene) with randomly distributed vacancies by etching of the quaternary solid solution (Nb2/3Sc1/3) 2AlC MAX phase. ACS Applied Nano Materials 2018;1:2455-60.

[60] Wang L, Liu D, Lian W, Hu Q, Liu X, Zhou A. The preparation of V2CTx by facile hydrothermal-assisted etching processing and its performance in lithium-ion battery. Journal of Materials Research and Technology 2020;9:984-93.

[61] Maleski K, Mochalin VN, Gogotsi Y. Dispersions of two-dimensional titanium carbide MXene in organic solvents. Chemistry of Materials 2017;29:1632-40.

[62] Mashtalir O, Lukatskaya MR, Kolesnikov AI, Raymundo-Pinero E, Naguib M, Barsoum M, et al. The effect of hydrazine intercalation on the structure and capacitance of 2D titanium carbide (MXene). Nanoscale 2016;8:9128-33.

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

[63] Yazdanparast S, Soltanmohammad S, Fash-White A, Tucker GJ, Brennecka GL. Synthesis and surface chemistry of 2D TiVC solid-solution MXenes. ACS Applied Materials & Interfaces 2020;12:20129-37.

[64] Chia HL, Mayorga-Martinez CC, Antonatos N, Sofer Zk, Gonzalez-Julian JJ, Webster RD, et al. MXene titanium carbide-based biosensor: strong dependence of exfoliation method on performance. Analytical chemistry 2020;92:2452-9.

[65] Xu C, Wang L, Liu Z, Chen L, Guo J, Kang N, et al. Large-area high-quality 2D ultrathin Mo2C superconducting crystals. Nature materials 2015;14:1135-41.

[66] Verger L, Xu C, Natu V, Cheng H-M, Ren W, Barsoum MW. Overview of the synthesis of MXenes and other ultrathin 2D transition metal carbides and nitrides. Current Opinion in Solid State and Materials Science 2019;23:149-63.

[67] Wang D, Zhou C, Filatov AS, Cho W, Lagunas F, Wang M, et al. Direct synthesis and chemical vapor deposition of 2D carbide and nitride MXenes. Science 2023;379:1242-7.

[68] Qin R, Li X, Hu M, Shan G, Seeram R, Yin M. Preparation of high-performance MXene/PVA-based flexible pressure sensors with adjustable sensitivity and sensing range. Sensors and Actuators A: Physical 2022;338:113458.

[69] Zeng X, Zhao C, Jiang X, Yu R, Che R. Functional Tailoring of Multi-Dimensional Pure MXene Nanostructures for Significantly Accelerated Electromagnetic Wave Absorption. Small 2023:2303393.

[70] Li R, Ma X, Li J, Cao J, Gao H, Li T, et al. Flexible and high-performance electrochromic devices enabled by self-assembled 2D TiO2/MXene heterostructures. Nature communications 2021;12:1-11.

[71] Naguib M, Mashtalir O, Lukatskaya MR, Dyatkin B, Zhang C, Presser V, et al. One-step synthesis of nanocrystalline transition metal oxides on thin sheets of disordered graphitic carbon by oxidation of MXenes. Chemical communications 2014;50:7420-3.

[72] Zhao X, Vashisth A, Prehn E, Sun W, Shah SA, Habib T, et al. Antioxidants unlock shelf-stable Ti3C2Tx (MXene) nanosheet dispersions. Matter 2019;1:513-26.

[73] Khazaei M, Arai M, Sasaki T, Chung CY, Venkataramanan NS, Estili M, et al. Novel electronic and magnetic properties of two-dimensional transition metal carbides and nitrides. Advanced Functional Materials 2013;23:2185-92.

[74] Kim SJ, Koh H-J, Ren CE, Kwon O, Maleski K, Cho S-Y, et al. Metallic Ti3C2T x MXene gas sensors with ultrahigh signal-to-noise ratio. ACS nano 2018;12:986-93.

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

[75] Lee E, VahidMohammadi A, Prorok BC, Yoon YS, Beidaghi M, Kim D-J. Room temperature gas sensing of two-dimensional titanium carbide (MXene). ACS applied materials & interfaces 2017;9:37184-90.

[76] Chen WY, Jiang X, Lai S-N, Peroulis D, Stanciu L. Nanohybrids of a MXene and transition metal dichalcogenide for selective detection of volatile organic compounds. Nature communications 2020;11:1302.

[77] Anasori B, Shi C, Moon EJ, Xie Y, Voigt CA, Kent PR, et al. Control of electronic properties of 2D carbides (MXenes) by manipulating their transition metal layers. Nanoscale Horizons 2016;1:227-34.

[78] Zhang Y, Jiang Y, Duan Z, Huang Q, Wu Y, Liu B, et al. Highly sensitive and selective NO2 sensor of alkalized V2CTx MXene driven by interlayer swelling. Sensors and Actuators B: Chemical 2021;344:130150.

[79] Yang Z, Liu A, Wang C, Liu F, He J, Li S, et al. Improvement of gas and humidity sensing properties of organ-like MXene by alkaline treatment. ACS sensors 2019;4:1261-9.

[80] Guo X, Ding Y, Kuang D, Wu Z, Sun X, Du B, et al. Enhanced ammonia sensing performance based on MXene-Ti3C2Tx multilayer nanoflakes functionalized by tungsten trioxide nanoparticles. Journal of Colloid and Interface Science 2021;595:6-14.

[81] Tai H, Duan Z, He Z, Li X, Xu J, Liu B, et al. Enhanced ammonia response of Ti3C2Tx nanosheets supported by TiO2 nanoparticles at room temperature. Sensors and Actuators B: Chemical 2019;298:126874.

[82] Zhang D, Yu S, Wang X, Huang J, Pan W, Zhang J, et al. UV illumination-enhanced ultrasensitive ammonia gas sensor based on (001) TiO2/MXene heterostructure for food spoilage detection. Journal of Hazardous Materials 2022;423:127160.

[83] Kalambate PK, Gadhari NS, Li X, Rao Z, Navale ST, Shen Y, et al. Recent advances in MXene–based electrochemical sensors and biosensors. TrAC Trends in Analytical Chemistry 2019;120:115643.

[84] Ran J, Gao G, Li F-T, Ma T-Y, Du A, Qiao S-Z. Ti3C2 MXene co-catalyst on metal sulfide photo-absorbers for enhanced visible-light photocatalytic hydrogen production. Nature communications 2017;8:1-10.

[85] Liao H, Guo X, Wan P, Yu G. Conductive MXene nanocomposite organohydrogel for flexible, healable, low-temperature tolerant strain sensors. Advanced Functional Materials 2019;29:1904507.

[86] Wu M, He M, Hu Q, Wu Q, Sun G, Xie L, et al. Ti3C2 MXene-based sensors with high selectivity for NH3 detection at room temperature. ACS sensors 2019;4:2763-70.

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

[87] Khakbaz P, Moshayedi M, Hajian S, Soleimani M, Narakathu BB, Bazuin BJ, et al. Titanium carbide MXene as NH3 sensor: realistic first-principles study. The Journal of Physical Chemistry C 2019;123:29794-803.

[88] Zeng F, Feng X, Chen X, Yao Q, Miao Y, Dai L, et al. First-principles analysis of Ti3C2Tx MXene as a promising candidate for SF6 decomposition characteristic components sensor. Applied Surface Science 2022;578:152020.

[89] Huang S, Mochalin VN. Understanding chemistry of two-dimensional transition metal carbides and carbonitrides (MXenes) with gas analysis. ACS nano 2020;14:10251-7.

[90] Kim SJ, Choi J, Maleski K, Hantanasirisakul K, Jung H-T, Gogotsi Y, et al. Interfacial assembly of ultrathin, functional MXene films. ACS applied materials & interfaces 2019;11:32320-7.

[91] Chae Y, Kim SJ, Cho S-Y, Choi J, Maleski K, Lee B-J, et al. An investigation into the factors governing the oxidation of two-dimensional Ti 3 C 2 MXene. Nanoscale 2019;11:8387-93.

[92] Li X, An Z, Lu Y, Shan J, Xing H, Liu G, et al. Room Temperature VOCs Sensing with Termination-Modified Ti3C2Tx MXene for Wearable Exhaled Breath Monitoring. Advanced Materials Technologies 2022;7:2100872.

[93] Koh H-J, Kim SJ, Maleski K, Cho S-Y, Kim Y-J, Ahn CW, et al. Enhanced selectivity of MXene gas sensors through metal ion intercalation: in situ X-ray diffraction study. ACS sensors 2019;4:1365-72.

[94] Shuvo SN, Ulloa Gomez AM, Mishra A, Chen WY, Dongare AM, Stanciu LA. Sulfur-doped titanium carbide MXenes for room-temperature gas sensing. ACS sensors 2020;5:2915-24.

[95] Pazniak H, Plugin IA, Loes MJ, Inerbaev TM, Burmistrov IN, Gorshenkov M, et al. Partially Oxidized Ti3C2T x MXenes for Fast and Selective Detection of Organic Vapors at Part-per-Million Concentrations. ACS Applied Nano Materials 2020;3:3195-204.

[96] Zhu Z, Liu C, Jiang F, Liu J, Ma X, Liu P, et al. Flexible and lightweight Ti3C2Tx MXene@ Pd colloidal nanoclusters paper film as novel H2 sensor. Journal of hazardous materials 2020;399:123054.

[97] Muckley ES, Naguib M, Wang H-W, Vlcek L, Osti NC, Sacci RL, et al. Multimodality of structural, electrical, and gravimetric responses of intercalated MXenes to water. ACS nano 2017;11:11118-26.

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

[98] Shpigel N, Levi MD, Sigalov S, Mathis TS, Gogotsi Y, Aurbach D. Direct assessment of nanoconfined water in 2D Ti3C2 electrode interspaces by a surface acoustic technique. Journal of the American Chemical Society 2018;140:8910-7.

[99] Yao L, Tian X, Cui X, Zhao R, Xiao X, Wang Y. Partially oxidized Ti3C2Tx MXene-sensitive material-based ammonia gas sensor with high-sensing performances for room temperature application. Journal of Materials Science: Materials in Electronics 2021;32:27837-48.

[100] Lorencova L, Bertok T, Filip J, Jerigova M, Velic D, Kasak P, et al. Highly stable Ti3C2Tx (MXene)/Pt nanoparticles-modified glassy carbon electrode for H2O2 and small molecules sensing applications. Sensors and Actuators B: Chemical 2018;263:360-8.

[101] Choi J, Kim YJ, Cho SY, Park K, Kang H, Kim SJ, et al. In situ formation of multiple schottky barriers in a Ti3C2 MXene film and its application in highly sensitive gas sensors. Advanced Functional Materials 2020;30:2003998.

[102] Chen WY, Jiang X, Lai S-N, Peroulis D, Stanciu L. Nanohybrids of a MXene and transition metal dichalcogenide for selective detection of volatile organic compounds. Nature communications 2020;11:1-10.

[103] Lee SH, Eom W, Shin H, Ambade RB, Bang JH, Kim HW, et al. Room-Temperature, Highly Durable $Ti_3C_2T_x$ MXene/Graphene Hybrid Fibers for NH_3 Gas Sensing. 2020.

[104] Zhao L, Wang K, Wei W, Wang L, Han W. High-performance flexible sensing devices based on polyaniline/MXene nanocomposites. InfoMat 2019;1:407-16.

[105] Sun S, Wang M, Chang X, Jiang Y, Zhang D, Wang D, et al. W18O49/Ti3C2Tx Mxene nanocomposites for highly sensitive acetone gas sensor with low detection limit. Sensors and Actuators B: Chemical 2020;304:127274.

[106] Liu M, Wang Z, Song P, Yang Z, Wang Q. In2O3 nanocubes/Ti3C2Tx MXene composites for enhanced methanol gas sensing properties at room temperature. Ceramics International 2021;47:23028-37.

[107] Xia Y, He S, Wang J, Zhou L, Wang J, Komarneni S. MXene/WS 2 hybrids for visible-light-activated NO 2 sensing at room temperature. Chemical Communications 2021;57:9136-9.

[108] Pandey AK. Graphene–Ti3C2Tx MXene hybrid nanostructure: a promising material for sensitivity enhancement in plasmonic sensor. Applied Physics A 2021;127:1-6.

[109] Hu Z-D, Sun X-Y, Li H-F, Kang Y-R, Song X-Q, Wang P, et al. Cobalt monosulfide nanofibers: ethanol sensing and magnetic properties. Rare Metals 2021;40:1554-60.

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

[110] Yan H, Chu L, Li Z, Sun C, Shi Y, Ma J. 2H-MoS2/Ti3C2Tx MXene composites for enhanced NO2 gas sensing properties at room temperature. Sensors and Actuators Reports 2022;4:100103.

[111] Yuan W, Yang K, Peng H, Li F, Yin F. A flexible VOCs sensor based on a 3D Mxene framework with a high sensing performance. Journal of Materials Chemistry A 2018;6:18116-24.

[112] Lee SH, Eom W, Shin H, Ambade RB, Bang JH, Kim HW, et al. Room-temperature, highly durable Ti3C2T x MXene/graphene hybrid fibers for NH3 gas sensing. ACS applied materials & interfaces 2020;12:10434-42.

[113] Yang Z, Zou H, Zhang Y, Liu F, Wang J, Lv S, et al. The Introduction of Defects in Ti3C2Tx and Ti3C2Tx-Assisted Reduction of Graphene Oxide for Highly Selective Detection of ppb-Level NO2. Advanced Functional Materials 2022;32:2108959.

[114] Li X, Xu J, Jiang Y, He Z, Liu B, Xie H, et al. Toward agricultural ammonia volatilization monitoring: A flexible polyaniline/Ti3C2Tx hybrid sensitive films based gas sensor. Sensors and Actuators B: Chemical 2020;316:128144.

[115] Wang X, Sun K, Li K, Li X, Gogotsi Y. Ti3C2Tx/PEDOT: PSS hybrid materials for room-temperature methanol sensor. Chinese Chemical Letters 2020;31:1018-21.

[116] Chen P, Zhao Z, Shao Z, Tian Y, Li B, Huang B, et al. Highly selective NH3 gas sensor based on polypyrrole/Ti3C2Tx nanocomposites operating at room temperature. Journal of Materials Science: Materials in Electronics 2022;33:6168-77.

[117] Sardana S, Singh Z, Sharma AK, Kaur N, Pati PK, Mahajan A. Self-powered biocompatible humidity sensor based on an electrospun anisotropic triboelectric nanogenerator for non-invasive diagnostic applications. Sensors and Actuators B: Chemical 2022;371:132507.

[118] Sardana S, Kaur H, Arora B, Aswal DK, Mahajan A. Self-powered monitoring of ammonia using an MXene/TiO2/cellulose nanofiber heterojunction-based sensor driven by an electrospun triboelectric nanogenerator. ACS sensors 2022;7:312-21.

[119] Wang M, Liu W, Shi X, Cong Y, Lin S, Sun T, et al. Self-powered and low-temperature resistant MXene-modified electronic-skin for multifunctional sensing. Chemical Communications 2021;57:8790-3.

[120] Gasso S, Mahajan A. Self-Powered Wearable Gas Sensors Based on l-Ascorbate-Treated MXene Nanosheets and SnO2 Nanofibers. ACS Applied Nano Materials 2023;6:6678-92.

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

[121] Wang D, Zhang D, Guo J, Hu Y, Yang Y, Sun T, et al. Multifunctional poly (vinyl alcohol)/Ag nanofibers-based triboelectric nanogenerator for self-powered MXene/tungsten oxide nanohybrid NO2 gas sensor. Nano Energy 2021;89:106410.

[122] Yu X-f, Li Y-c, Cheng J-b, Liu Z-b, Li Q-z, Li W-z, et al. Monolayer Ti2CO2: a promising candidate for NH3 sensor or capturer with high sensitivity and selectivity. ACS applied materials & interfaces 2015;7:13707-13.

[123] Xiao B, Li Y-c, Yu X-f, Cheng J-b. MXenes: Reusable materials for NH3 sensor or capturer by controlling the charge injection. Sensors and Actuators B: Chemical 2016;235:103-9.

[124] Wang Y, Ma S, Wang L, Jiao Z. A novel highly selective and sensitive NH3 gas sensor based on monolayer Hf2CO2. Applied Surface Science 2019;492:116-24.

[125] Hajian S, Khakbaz P, Moshayedi M, Maddipatla D, Narakathu BB, Turkani VS, et al. Impact of different ratios of fluorine, oxygen, and hydroxyl surface terminations on Ti3C2T x MXene as ammonia sensor: a first-principles study. 2018 IEEE SENSORS: IEEE; 2018. p. 1-4.

[126] Zhang Y, Zhou Z, Lan J, Zhang P. Prediction of Ti3C2O2 MXene as an effective capturer of formaldehyde. Applied Surface Science 2019;469:770-4.

[127] Ma S, Yuan D, Jiao Z, Wang T, Dai X. Monolayer Sc2CO2: a promising candidate as a SO2 gas sensor or capturer. The Journal of Physical Chemistry C 2017;121:24077-84.

[128] Yang D, Fan X, Zhao D, An Y, Hu Y, Luo Z. Sc2CO2 and Mn-doped Sc2CO2 as gas sensor materials to NO and CO: A first-principles study. Physica E: Low-dimensional Systems and Nanostructures 2019;111:84-90.

[129] Banu AA, Sinthika S, Premkumar S, Vigneshwaran J, Karazhanov SZ, Jose SP. DFT study of NH3 adsorption on 2D monolayer MXenes (M2C, M= Cr, Fe) via oxygen functionalization: Suitable materials for gas sensors. FlatChem 2022;31:100329.

[130] Quan W, Shi J, Luo H, Fan C, Lv W, Chen X, et al. Fully Flexible MXene-based Gas Sensor on Paper for Highly Sensitive Room-Temperature Nitrogen Dioxide Detection. ACS sensors 2023;8:103-13.

[131] Weng K, Peng J, Shi Z, Arramel A, Li N. Highly NH3 Sensitive and Selective Ti3C2O2-Based Gas Sensors: A Density Functional Theory-NEGF Study. ACS omega 2023;8:4261-9.

[132] Reji RP, Balaji SKC, Sivalingam Y, Kawazoe Y, Velappa Jayaraman S. First-Principles Density Functional Theory Calculations on the Potential of Sc2CO2 MXene Nanosheets as a Dual-Mode Sensor for Detection of Volatile Organic Compounds in Exhaled Human Breath. ACS Applied Nano Materials 2023;6:5345-56.

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

[133] Jian Y, Qu D, Guo L, Zhu Y, Su C, Feng H, et al. The prior rules of designing Ti3C2Tx MXene-based gas sensors. Frontiers of Chemical Science and Engineering 2021;15:505-17.

[134] Wang J, Xu R, Xia Y, Komarneni S. Ti2CTx MXene: A novel p-type sensing material for visible light-enhanced room temperature methane detection. Ceramics International 2021;47:34437-42.

[135] Chen WY, Lai S-N, Yen C-C, Jiang X, Peroulis D, Stanciu LA. Surface functionalization of Ti3C2T x MXene with highly reliable superhydrophobic protection for volatile organic compounds sensing. ACS nano 2020;14:11490-501.

[136] Chachuli SAM, Hamidon MN, Ertugrul M, Mamat MS, Coban O, Tuzluca FN, et al. Effects of MWCNTs/graphene nanoflakes/MXene addition to TiO2 thick film on hydrogen gas sensing. Journal of Alloys and Compounds 2021;882:160671.

[137] Kuang D, Guo X, Zhu Z, Ding Y, Sun X, Wu Z, et al. Enhanced room temperature ammonia response of 2D-Ti3C2Tx MXene decorated with Ni (OH) 2 nanoparticles. Ceramics International 2021;47:19471-80.

[138] Sun Q, Wang J, Wang X, Dai J, Wang X, Fan H, et al. Treatment-dependent surface chemistry and gas sensing behavior of the thinnest member of titanium carbide MXenes. Nanoscale 2020;12:16987-94.

[139] Wang D, Zhang D, Yang Y, Mi Q, Zhang J, Yu L. Multifunctional latex/polytetrafluoroethylene-based triboelectric nanogenerator for self-powered organ-like MXene/metal–organic framework-derived CuO nanohybrid ammonia sensor. ACS nano 2021;15:2911-9.

[140] Liu Z, He T, Sun H, Huang B, Li X. Layered MXene heterostructured with In2O3 nanoparticles for ammonia sensors at room temperature. Sensors and Actuators B: Chemical 2022;365:131918.

[141] Hermawan A, Zhang B, Taufik A, Asakura Y, Hasegawa T, Zhu J, et al. CuO Nanoparticles/Ti3C2T x MXene Hybrid Nanocomposites for Detection of Toluene Gas. ACS Applied Nano Materials 2020;3:4755-66.

[142] Yang Z, Jiang L, Wang J, Liu F, He J, Liu A, et al. Flexible resistive NO2 gas sensor of three-dimensional crumpled MXene Ti3C2Tx/ZnO spheres for room temperature application. Sensors and Actuators B: Chemical 2021;326:128828.

[143] Song Y, Xu Y, Guo Q, Hua Z, Yin F, Yuan W. MXene-derived TiO2 nanoparticles intercalating between RGO nanosheets: an assembly for highly sensitive gas detection. ACS Applied Materials & Interfaces 2021;13:39772-80.

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

View Article Online
DOI: 10.1039/D3MA00631J

[144] Tang Y, Xu Y, Yang J, Song Y, Yin F, Yuan W. Stretchable and wearable conductometric VOC sensors based on microstructured MXene/polyurethane core-sheath fibers. Sensors and Actuators B: Chemical 2021;346:130500.

[145] Liu M, Wang Z, Song P, Yang Z, Wang Q. Flexible MXene/rGO/CuO hybrid aerogels for high performance acetone sensing at room temperature. Sensors and Actuators B: Chemical 2021;340:129946.

[146] Zhou Y, Wang Y, Wang Y, Li X. Humidity-Enabled Ionic Conductive Trace Carbon Dioxide Sensing of Nitrogen-Doped Ti3C2T x MXene/Polyethyleneimine Composite Films Decorated with Reduced Graphene Oxide Nanosheets. Analytical Chemistry 2020;92:16033-42.

[147] Kuang D, Wang L, Guo X, She Y, Du B, Liang C, et al. Facile hydrothermal synthesis of Ti3C2Tx-TiO2 nanocomposites for gaseous volatile organic compounds detection at room temperature. Journal of Hazardous Materials 2021;416:126171.

Open Access Article. Published on 18 October 2023. Downloaded on 11/3/2023 5:14:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Materials Advances Accepted Manuscript

Use of Poly Vinyl Chloride (PVC) Tape to Increase the Strength of Bamboo

Rishabh Rastogi, Yogesh, Nikhil Kumar Pankaj, B. R. G. Robert

Abstract: ***Poly vinyl chloride is suggested to use as confining material around bamboo to increase its strength. Investigation is done on PVC confined bamboo by performing tensile and compressive strength test. Then we compared the mechanical performance of unconfined bamboo and PVC confined bamboo. In this research bamboo is also tested for water absorption with and without PVC tape reinforcement. Later, on these grounds we also derived adhesive value of PVC tape, which can be used to find out length of tape required for any percent of increment in strength of bamboo. Preliminary results of tensile tests, compressive tests and water absorption test on bamboo (scientific name Bambusa vulgaris native name chadao bans or jad vala bans) are presented in this research. As bamboo confined in PVC is 27.25% more strong in tension and 77.19% more strong in compression than unconfined bamboo which depicts that bamboo confined with PVC tape can be a future alternative for steel in reinforced concrete.***

Keywords**: Poly Vinyl Chloride (PVC) tape, Bamboo splint, Bamboo culm, Splitting buckling, End bearing buckling, Tape adhesion.***

I. INTRODUCTION

There are mainly three basis on which building materials are selected and these are financial, functional and technical specification since long time. But these days one important aspect is sustainability which emerged as even more critical basis to select a building material to reduce environmental pollution and efficient utilization of resources. The building industry, creating a substantial part of total environmental deterioration. One must invent new environment friendly ways in construction and thus contributing in sustainable use of resources. This can be done by many ways like using renewable material in construction, using green and ecofriendly resources in construction, reusing same resource as many time as possible and least utilization of resources. In this paper we focused on using renewable material for construction purpose. Bamboo a fast growing, natural and renewable resource which can be used as a reinforcement may be partially or fully in concrete structure.

Revised Manuscript Received on July 22, 2020.

Rishabh Rastogi, Department of Civil Engineering, Delhi Technology University, New Delhi, India.

Yogesh, Department of Civil Engineering, Delhi Technology University, New Delhi, India.

Nikhil Kumar Pankaj, Department of Civil Engineering, Delhi Technology University, New Delhi, India.

B.R.G. Robert, Assistant Professor, Department of Civil Engineering, Delhi Technological University, New Delhi, India.

But bamboo strength is not as much high as that of steel so in this paper we tried to increase the strength of bamboo by confining it with PVC tape. India stands at the second spot in the production of bamboo resources in the world, next to China. Bamboo can be sustainable alternative for reinforcement by adequate research. Being hollow in nature bamboo can resist bending forces upto larger extent. Keeping in mind it is not easy to do connections of bamboo but this problem can be overcome by research once its strength is increased considerably. So, our prime objective is to increase strength of bamboo anyhow. Around 23% of air pollution, 40% of water pollution, 50% of landfill wastes and 50% of climatic changes is contributed by construction industry. More than 1 ton of carbon dioxide gas is emitted in the atmosphere for production of 1 ton of cement. Similarly, more than 2 ton of carbon dioxide gas is emitted in atmosphere just to produce 1 ton of steel. As per researches, bamboo fails in tension at strength of around 100 to 130 MPa and in compression at strength of around 40 to 50 MPa. Due to compressive buckling failure of wall of bamboo it has a low compressive strength. In dry environment wall of bamboo get cracked easily and these cracks also contribute in low compressive strength. Hence considering these drawbacks of bamboo we used PVC tape as confining material to increase its load carrying capacity.

II. MATERIAL USED

Bamboo (scientific name *Bambusa vulgaris* native name chadao bans or jad vala bans) was brought from R K plywood Bawana road Delhi, Poly vinyl chloride (PVC) tape was acquired from electrical shop.

III. USE OF PVC TAPE FOR CONFINING BAMBOO

In steel or concrete engineering carbon fiber reinforced polymer (CFRP) is already in extensive use. But it is costly so, we have used a cheaper durable alternative that is PVC as confining material and tested the strength of bamboo.

In this report we have suggested the use of PVC tape as confining material by circularly confining bamboo to increase its mechanical strength.

IV. TENSILE STRENGTH TEST OF BAMBOO SPLINT

A. Testing bamboo specimen without PVC confinement

Tensile strength of bamboo splint without PVC confinement was studied firstly.

Bamboo (scientific name *Bambusa vulgaris* native name chadao bans or jad vala bans) was used as specimen for conducting all the test. From the wall of bamboo culm four equal size bamboo splint specimen was cut along the length of bamboo. We sanded all bamboo splint sample of equal dimensions into dog-bone shape and then tested two of them without confining in PVC tape. We have shown one sample in figure 1.

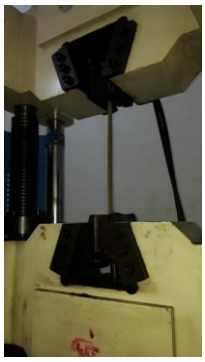
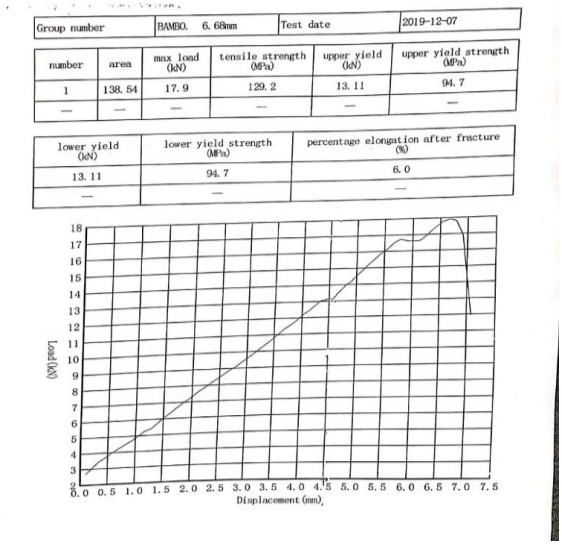


Fig. 1 Bamboo splint without PVC tape in UTM

Each bamboo splint specimen was 430mm long. Although the specimen cross-section is in the shape of quadrant of arch but we have considered it as rectangular in shape for ease of calculations. To calculate cross-section area, each specimen area is calculated at three points that is at top node, bottom node and middle position then their average is taken. Cross-sectional area coming out to be using this method is 138.5mm^2 for one of the two sample. The sample fractured in the middle portion of its length. The average strength of two test sample was found to be 124.4 MPa.

Fig. 2 Fractured sample after loading

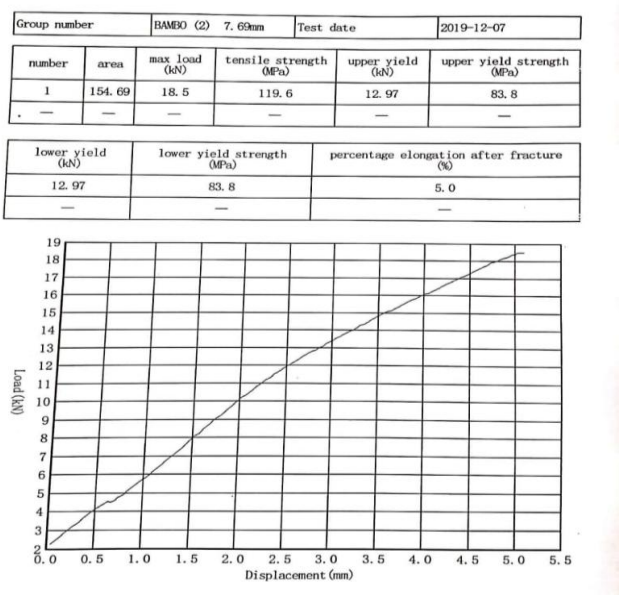
Maximum stress obtained in bamboo at failure is measured by dividing the maximum tensile load taken by specimen at failure with the average area of cross-section. Strength of two samples was found to be 129.2MPa and 119.6MPa, with a mean of 124.4MPa. The Percentage elongation of specimen after fracture was 5.5%. In figure 3 load vs displacement curves are shown for tensile specimens. In table I results of tensile test without PVC tape confinement are summarized.



Group number	BAMBO. 6.68mm		Test date		2019-12-07
number	area	max load (kN)	tensile strength (MPa)	upper yield (kN)	upper yield strength (MPa)
1	138.54	17.9	129.2	13.11	94.7
—	—	—	—	—	—

lower yield (kN)	lower yield strength (MPa)	percentage elongation after fracture (%)
13.11	94.7	6.0
—	—	—

Fig. 3 Load vs Displacement curve of first sample



Group number	BAMBO (2) 7.69mm		Test date		2019-12-07
number	area	max load (kN)	tensile strength (MPa)	upper yield (kN)	upper yield strength (MPa)
1	154.69	18.5	119.6	12.97	83.8
—	—	—	—	—	—

lower yield (kN)	lower yield strength (MPa)	percentage elongation after fracture (%)
12.97	83.8	5.0
—	—	—

Fig. 4 Load vs Displacement curve of second sample

Table – I : Results of the tensile tests without PVC confinement

	Sample 1	Sample 2	Average
Maximum load (kN)	17.9	18.5	18.2
Tensile strength (MPa)	129.2	119.6	124.4
Upper yield (kN)	13.11	12.97	13.04
Upper yield strength (MPa)	94.7	83.8	89.25
Lower yield (kN)	13.11	12.97	13.04

Lower yield strength (MPa)	94.7	83.8	89.25
Percentage elongation after fracture (%)	6.0	5.0	5.5

B. Testing bamboo specimen with PVC tape confinement

In this case, the fiber strength test was performed on the tensile sample wrapped with the PVC tape. The variety of bamboo was same as earlier and it was *Bambusa vulgaris* native name chadao bans or jad vala bans. A sample of that kind has been shown in Figure 5 and 6.

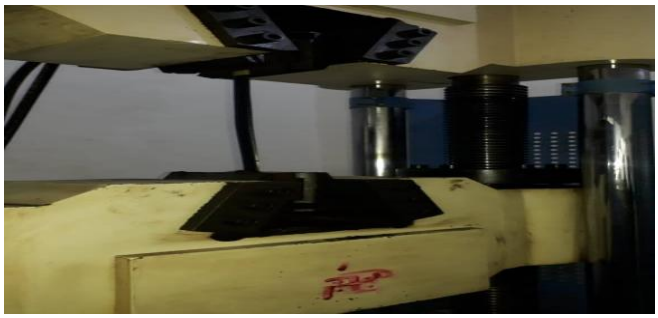


Fig. 5 Testing of bamboo with tape confinement in UTM

Fig. 6 Dog bone shape sample of bamboo splint

Each of the sample used for the test was found to be 430 mm long. The sample is finished into 'Dog bone' shape having a average cross sectional area of 154.72mm^2. The width of PVC tape is 15mm. The number of PVC tape wrapping is 24 without overlapping of PVC tape over a length of 360mm out of 430mm leaving a space of 35mm on both sides. The sample fractured in the middle portion of its length. The strength of tensile test sample was found to be 158.3Mpa.

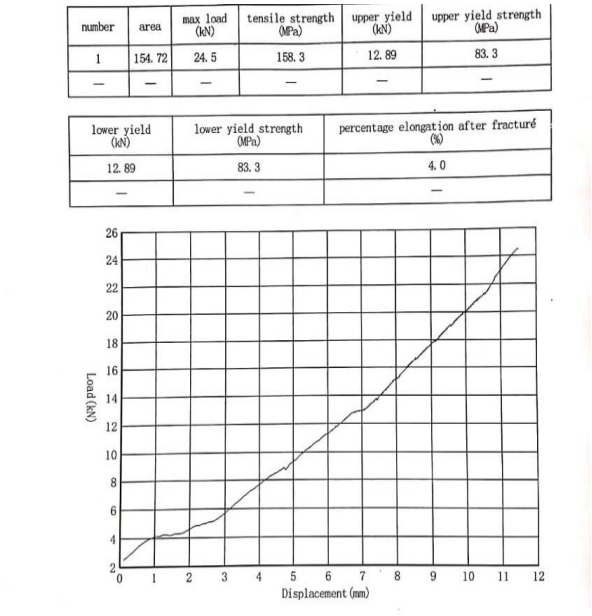


Fig. 7 Load vs Displacement curve for confined bamboo sample

The strength of the tensile samples without PVC tape confinement gave an average value of 124.4Mpa. On the other hand, the strength of the samples with PVC tape gave an average value of 158.3Mpa. Thus, the mean strength of tensile bamboo sample confined with PVC tape can be enhanced to 27.25% more than the mean strength of bamboo sample without PVC tape.

V. DERIVATION OF TAPE ADHESION ON BAMBOO

Maximum tensile load without tape confinement = 18.2kN
Maximum tensile load with tape confinement = 24.5kN
Increase in load bearing because of tape confinement = 24.5 - 18.2 = 6.3kN

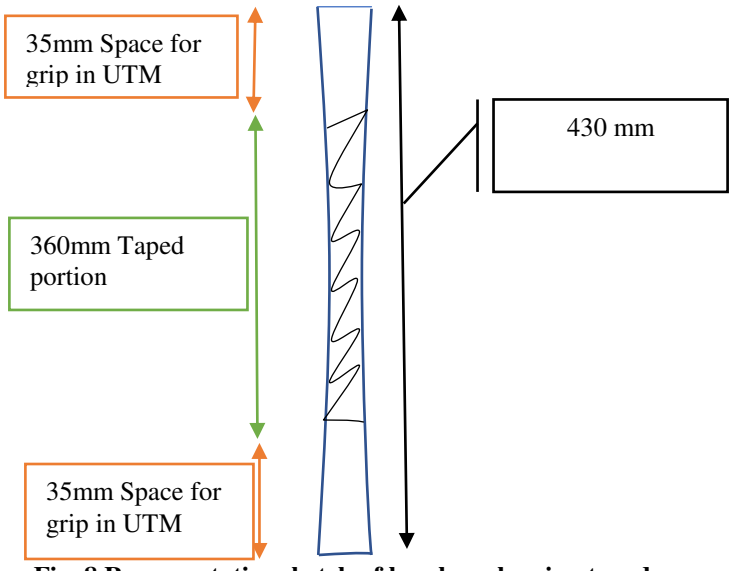


Fig. 8 Representative sketch of bamboo showing taped portion

Tape width = 15mm

No. of tape wrappings around bamboo splint = $\frac{360}{15}$ = 24

Perimeter of bamboo splint = $2(20.12 + 7.69)$ = 55.62mm

Length of total tape used = No. of wrappings × Perimeter of bamboo splint

= 24 × 55.62 mm

= 1335mm

Using increment in load and total tape length used we can derive the value of Adhesion force per unit length of tape wrapped around the bamboo.

So, **Adhesion of tape on bamboo** = $\frac{6.3 \times 1000}{1335}$ N/mm = **4.72 N/mm**

Now using this Adhesion value, we can find out the length of tape required corresponding to the load increment.

VI. COMPRESSIVE STRENGTH TEST OF BAMBOO CULM

A. Testing bamboo specimen without PVC confinement

Each of the samples used in the test has a minimum length which was approximately three times the external diameter of culms of bamboo, but it was not exceeding 200mm.The internal diameter of the sample was 12.58mm and the external diameter was 30.7mm so the net area was 619.8mm^2.

Fig. 9 Splitting failure of bamboo culm

After putting sample in CTM we applied compressive load and then noted the compressive strength by dividing it with net cross section area of bamboo.

Fig.10 Compressive load is applied on sample

Due to applied load the bamboo fractured at the middle. The failure mode of specimens without the PVC tape is splitting fracture. The splitting fracture occurred due to the compressive buckling failure of the walls of bamboo. The maximum compressive load taken by sample was 26kN.

B. Testing bamboo specimen with PVC tape confinement

This test was performed as per the Guiqiu Huang, Zhen Huang where they used carbon fiber as confining material and we have used PVC tape instead of carbon fiber. The number of wrapping of PVC tape around the sample was 10. The length of sample was 150mm. The external diameter was 27.56mm and internal diameter was 14.54mm. The net area was 430.51mm^2.

Fig. 11 PVC confined bamboo culm

Fig.13 is showing the compressive failure modes along with PVC tape. The mode of failure of the samples with the PVC tape was the end bearing failure. The maximum compressive load taken by the specimen with PVC confinement was 32kN.

Fig.12 Applying load on PVC confined sample using UTM

Fig. 13 End bearing failure of sample

Table – II : Results of the Compressive tests without and with PVC confinement

Properties	Sample without PVC tape	Sample with PVC tape
Load (N)	26000	32000
Area (sq.mm)	619.8	430.51
Compressive strength (MPa)	41.95	74.33
Failure mode	Splitting buckling	End bearing buckling

VII. WATER ABSORPTION TEST

A. Bamboo without PVC tape confinement

Splint of bamboo with dimensions 20.74mm×6.68mm and height 430 mm was left dipped in water for 24 hours. The weight of the splint before and after dipping was recorded.

Weight before dipping, $W_1 = 20g$

Weight after dipping, $W_2 = 50g$

Weight of water = $W_2 - W_1 = 30g$

Water content, $M_1 = \{(W_2 - W_1)/ W_1\}\times100$

$= 150\%$

B. Bamboo with PVC tape confinement

Splint of bamboo with same dimensions 20.74mm x6.68mm and height 430 mm was also left dipped in water for 24 hours. The weight of the splint before and after dipping was recorded.

Weight before dipping, $W_1= 20g$

Weight after dipping, $W_2 = 30g$

Weight of water = $W_2 - W_1 = 10g$

Water content, $M_2 = \{(W_2 - W_1)/ W_1\}\times100$

$= 50\%$

Fig. 14 Water bath of samples for 24 hours

VIII. RESULTS AND DISCUSSION

A. Results of tension test

The strength of the tensile samples without PVC tape confinement gave an average value of 124.4Mpa. On the other hand, the strength of the samples with PVC tape gave an average value of 158.3Mpa. Thus, the mean strength of tensile bamboo sample confined with PVC tape can be enhanced to 27.25% increased than the mean strength of bamboo sample without PVC tape.

B. Result of compression test

A Bamboo culm of height 150 mm without PVC tape was tested for compression strength test and the maximum compressive load was coming as 26kN.

The compressive strength was recorded as 41.95 MPa. When a Bamboo culm of same height 150 mm with PVC tape was tested for compression strength test and the maximum compressive load was coming as 32kN. The compressive strength was recorded as 74.33 MPa. Thus, the mean compressive strength of bamboo confined with PVC tape can be enhanced to 77.19 % greater than the value obtained for bamboo not confined with PVC tape.

C. Results of water absorption test

As w2 < w1, so splint with PVC tape wrapped over it has less water absorbed in comparison to splint without PVC tape.
So, we can use PVC tape as it does not absorb much water from the concrete thus enhancing its durability.

IX. CONCLUSION

Due to wrapping of PVC tape there was substantial increase in tensile strength and compressive strength of splint of bamboo and bamboo culm respectively. Hence it can be a future alternative for reinforcement in concrete structures.
The test for water absorption was also performed on two bamboo splints of same kind that were used in tensile test. The results were positive and specified the norms as per the standard research papers.

REFERENCES

1. Bhalla, S., Gupta, S., Puttaguna, S. and Suresh, R. (2009), "Bamboo as Green Alternative To Concrete and Steel for Modern Structures", Journal of Environmental Research and Development.
2. Fei Ye1 and Wenxi Fu, Ph.D., "Physical and Mechanical Characterization of Fresh Bamboo for Infrastructure Projects" American Society of Civil Engineers.
3. A. W. Mekonnen and J. N. Mandal, "Model Studies on Bamboo-Geogrid Reinforced Fly Ash Walls under Uniformly Distributed Load" 2017 American Society of Civil Engineers.
4. Wenqing Wu, M.ASCE, "Experimental Analysis of Bending Resistance of Bamboo Composite I-Shaped Beam" 2013 American Society of Civil Engineers.
5. Vishal Puri and Pradipta Chakrabortty, "Policy Issues in Affordable Housing Made with Bamboo Reinforced Structural Component" 2013 American Society of Civil Engineers.
6. Nathan Schneider, Weichiang Pang, Mengzhe Gu, "Application of Bamboo for Flexural and Shear Reinforcement in Concrete Beams" Structures Congress 2014 © ASCE 2014.
7. Guiqiu Huang, Zhen Huang, Xueyuan Deng, Jing Jiang, "Experimental Study on Carbon Fiber Polymer Reinforced Bamboos" Civil Engineering and Urban Planning 2012 (CEUP 2012) ©ASCE 2012.

AUTHORS PROFILE

Rishabh Rastogi, Bachelor of technology in Civil engineering from Delhi Technology University, New Delhi.
Membership - Present member of Indian Geotechnical Society Delhi Chapter.
Achievement - Secured 2nd position in INSPIRE AWARD, DST, Govt. of India at Delhi state level at thyagraj stadium, august-2014. Project – A chimney which works on the principle of electrostatic charge to reduce air pollution. Participated in INSPIRE AWARD, DST, Govt. of India at national level at Pragati maidan (ITPO),Oct-2014. Awarded with a novel of Sudha Murthy and completion certificate of USIP internship by Prof. Yogesh Singh (vice chancellor, DTU) at USIP valediction ceremony (30 May,2019), Ex-intern at L&T Panipat Elevated Corridor Limited (L&T PECL), Panipat Haryana (June, 2018- July, 2018).

Yogesh, Bachelor of technology in Civil engineering from Delhi Technology University, New Delhi.
Membership - Member of University Student Internship Program (USIP) during the year 2018. It is an Internship Program to engage DTU students in various activities of the university
Achievement-Awarded with a novel of Chanakya and completion certificate of USIP internship by Prof. Yogesh Singh (Vice Chancellor DTU) at USIP valediction ceremony held on 30 May 2019. Did summer training at BAANI City Center Sector – 63, Gurugram, Haryana from June 2019-July 2019.It was a commercial project offering High street retail food court and Service apartments. There I gained knowledge about the construction activities held at the site and structural members like footing, columns, etc. Achieved Consolation prize in Bharat Ko Jaano Quiz Competition in 2013.

Nikhil Kumar Pankaj, Bachelor of technology in civil engineering from Delhi Technological University, New Delhi.
Ex intern at Larsen and turbo barapullah flyover (June -July) 2019.
Achievement- Member of Innova (tech and management team of DTU).
Membership - Present member of Indian Geotechnical Society Delhi Chapter.

B.R.G. Robert, Assistant professor in Delhi Technological University, New Delhi.
Areas of Interest - Geology, Surveying, Building materials & Construction. Testing of Materials, Design Mix of Concrete & Job mix for road works, Third party Quality Assurance. Conducting Survey camp & Industrial visits for students. Teaching Building material and Construction and Surveying to IV Sem Students of Civil Engineering.

2023 3rd Asian Conference on Innovation in Technology (ASIANCON)
Pune, India. Aug 25-27, 2023

Using discussion thread context in Sentiment analysis for improving Cyberbullying detection

Manoj Kumar
Dept. of Computer Science and Engineering
Delhi Technological University
Delhi, India
mkumarg@dce.ac.in

Himanshu
Dept. of Computer Science and Engineering
Delhi Technological University
Delhi, India
zhimanshuk135@gmail.com

Vishal Choudhary
Dept. of Computer Science and Engineering
Delhi Technological University
Delhi, India
tusharchoudharyjaat100@gmail.com

Yash Nishal
Dept. of Computer Science and Engineering
Delhi Technological University
Delhi, India
yashnishal@gmail.com

***Abstract*— Cyberbullying is the act of harassing, demeaning, or threatening someone via technology such as social networking, texting, or instant messaging. It is defined as "willful and repeated harm inflicted through the medium of electronic text" (Patchin and Hinduja 2006:152) [1]. Cyberbullying is an underlying problem in social networking services, threatening users' mental and physical health. The aim of this study is to improve the detection of cyberbullying in discussion threads by utilizing the repeated negative context formed in the presence of cyberbullying.**

***Keywords*— Cyberbullying, discussion thread, adjusted sentiments, LSTM, RNN**

I. Introduction

In the recent decade, due to the non-stop widespread of the internet and the popularity of several social media networks, cyberbullying has become progressively more prevalent. This has adverse psychological and mental health impact on people of all ages and genders. The recent pandemic of COVID-19 created even more opportunities for cyberbullying incidents, forcing everyone to stay indoors and increase the consumption of the internet. India has the greatest rate of cyberbullying, at 38%, followed by Brazil at 29% and the US at 26%, according to Techjury research [2]. Denmark, Finland, the Netherlands, and Norway are among the European nations where minors are most likely to be exposed to dangerous content and engage in online sexual exploitation.

According to Pew Research Center's figures [3] on online cyberbullying, 49% of teenagers between the ages of 15 and 17 have engaged in at least one of the six online behaviours, compared to 42% of those between the ages of 13 and 14. In a poll of more than 1000 parents of kids older than 5 years old [3], Comparitech discovered that 60% of parents of kids between the ages of 14 and 18 said their kids experienced bullying in the year 2019. Online bullying and harassment are serious problems for kids their age, according to 59% of American teenagers who have experienced it. [Pew Research Centre] (2018, September 27) [3]. Instagram had the highest rate of youth cyberbullying (42%) compared to all other platforms, with Facebook not far behind at 37%. At 31%, Snapchat came in third place according to the annual bullying survey, 2017 [5].

Cyberbullying incidents typically involve two main parties: the cyberbully and the victim. The reasons for cyberbullying may vary depending on the specific type of cyberbullying that occurs. However, it's worth noting that in addition to the cyberbully and victim, there may be other roles that come into play and exacerbate the situation. There are three main features to identify a potential cyberbullying case i.e., Negative or hostile intent, targeted subject and repetition of aggressive behaviour towards the subject.

Another fact that needs to be considered while detecting cyberbullying is that every negative behaviour cannot be interpreted as a potential bullying case as there are many situations in which disagreement between thoughts of different individuals or debates, where people have the right to express their feelings by negating the ideas and values defined by other people. All these points refer to the necessity of having the whole context of discussion for identifying a situation as potential cyberbullying or not. Thus, we propose a model that will consider the discussion context of the thread to identify it as a potential case of cyberbullying.

II. Literature Review

A. Sentiment analysis

(Smith et al. 2012) [13] proposed a modified lexicon-based approach for sentiment analysis by including valence shifters to tackle the shortcomings of traditional Bag-of-Word (BoW) methods. (Saif et al. 2012) [14] suggested the significance of including semantic features in sentiment analysis. They provided comparisons that proved including semantic features in sentiment analysis outperformed the traditional approaches.

(Barbieri et al. 2018) [15] explained how sentiment analysis can be effectively performed on emojis and emoticons and the improvement upon including emojis sentiments with textual sentiment of the data. Using this study, (Barbieri et al. 2022) [10] proposed a state-of-art sentiment analysis model having multilingual, emojis and slang support.

B. Detecting cyberbullying

(Yin et al. 2009) [7] provided an improved version of the BoW model for cyberbullying detection by providing localised weights to words using Term Frequency/Inverse Document Frequency (TF-iDF) which gave an accuracy of 39.4% and recall of 61.9% in detecting online harassment. (Reynolds et al. 2011) [12] found out in their study that

2023 3rd Asian Conference on Innovation in Technology (ASIANCON) | 979-8-3503-0228-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ASIANCON58793.2023.10270005

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:44:30 UTC from IEEE Xplore. Restrictions apply.

cyberbullying sentences often contain some commonly used curse words and bad slang, and utilised their severity in a BoW model to achieve 81.7% accuracy. (Mody et al. 2018) [8] proposed a hybrid model by integrating knowledge-based and machine learning approaches to reinforce sentiment analysis in the detection of cyberbullying.

C. Extensions to cyberbullying detection

(Pawar et al. 2019) [16] developed a multilingual cyberbullying detection system with support for English, Hindi and Marathi languages. (Bozyi˘git et al. 2021) [11] utilized the social media features of users to improve the classification of cyberbullying tweets by machine learning approaches. (Dadvar et al. 2012) [17] highlighted that cyberbullying done by different genders has different key points of action.

(Balakrishnan et al. 2019) [18] studied the significance of psychological features of Twitter users on the classification of their tweets. (Li et al. 2016) [6] studied the relationship between parent and child comments in discussion threads and proposed the adjustment of sentiments of parent comments by including context from child comments.

III. Proposed Architecture

A. Data Set

We required a dataset that contained discussion threads and their tags as cyberbullying or not. For our experiment, we decided to go with YouTube's comment threads as discussion threads. As there is no open data set available for YouTube comment threads, we created our data set by scraping comments using YouTube API provided by Google. A total of 4 channels, each from different video categories were chosen and comment threads of the top ten videos by popularity from each channel were scraped. A total of 1040 discussion threads consisting of a total of 25971 comments were collected.

Due to the contextual sensitivity of such large and varied discussion threads, we manually labelled the discussion threads for the presence of cyberbullying. We included a new attribute 'label' for the parent comments of each thread and provided a value of either 'cyberbullying' to indicate the presence of cyberbullying in the thread or 'normal' to indicate its absence. The overall structure of our model is illustrated in Fig. 1.

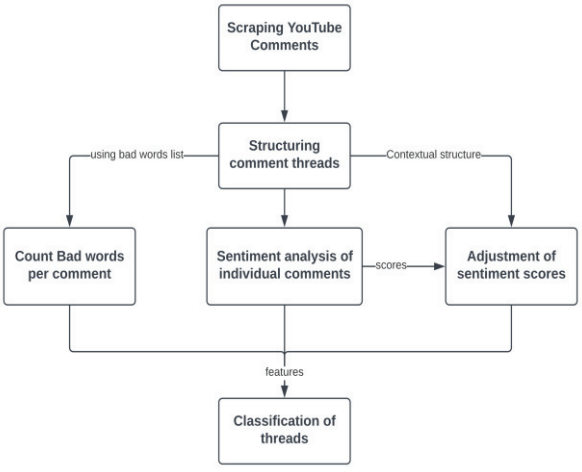


Fig. 1. Architecture diagram

B. Tree-based thread structure

We propose a tree-based discussion thread structure with adjustments upon individual sentiment scores of the comments, which will help improve the overall efficiency of cyberbullying detection in the threads. Our model considers the discussion thread as an N-ary tree in which each node represents the comments from a single thread and the edges are drawn from a parent comment to its subsequent replies, as shown in Fig. 2. This helps us consider the repetition of negative intent and targeted subjects while adjusting their sentiment scores. The tree is created by utilising the '@author-name' property added to the start of a reply comment where 'author-name' points to the author of the parent comment of the respective reply.

Sentiment scores of individual comments are calculated for each thread, using the 'twitter-XLM-roBERTa-base for Sentiment Analysis' model [10]. It returns positive, negative and neutral sentiment scores for each comment passed through it. We have chosen this model due to its ability to consider Emojis and common online vocabulary while providing high accuracy. We consider only the negative sentiment scores for each comment as they are the only significant attribute while considering cyberbullying. We also considered a count of bad words for each comment calculated by utilising the 'Bad words list' from Kaggle and updated it by adding some more words to the list.

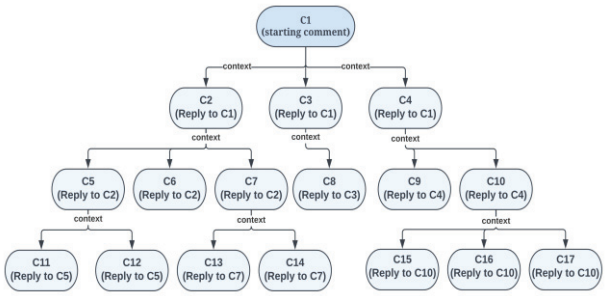


Fig. 2. Comment thread's tree-structure example

C. Adjustment of sentiment score

The original sentiment scores of each comment are adjusted to include context from the discussion. This allows iteratively passing the context of the discussion to each reply, updating the actual sentiment of the discussion at each comment. The sentiment score of the starting comment of the thread is not altered as there is no context present till then. For each reply, the original sentiment score is adjusted from the adjusted sentiment score of its parent comment, given by the formula:

$$AS = CS + \beta * ELU(ASP - CS, \alpha) \qquad (1)$$

Here, AS stands for the adjusted score of comment, CS stands for the current score of the comment, ASP stands for the adjusted score of parent comment and ELU represents the Exponential Linear Unit function. α and β are two parameters that define the properties of our adjustment. The α parameter determines the curvature of the ELU function when the current sentiment score of the comment is more than adjusted score of the parent while β restricts the magnitude of adjustment of the score from previous scores.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:44:30 UTC from IEEE Xplore. Restrictions apply.

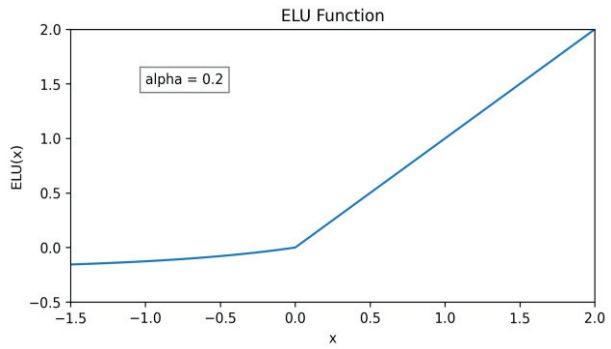


Fig. 3. ELU function with α = 0.2

The ELU function is designed such that it is linear for positive inputs but gives a sigmoid-like curvature for negative input. This helps us retain the repetition of negative sentiments in the context and restrict the effect of positive sentiments of parents to its replies. We consider the value of β = 0.7 and α = 0.2 to be the best fit for our model by experimentation. Fig. 3 provides the graph of ELU function with α = 0.2. Such a small value of α will ensure that the repeated negative sentiment's score will affect the adjusted score more intensely than repeated positive scores, which will help us extract the repeated negative sentiments from the discussion.

D. Training and testing

Each comment of a thread has an original sentiment score, adjusted score and count of bad words as its features. As the number of comments per thread is variable, we have taken Long Short-Term Model (LSTM) and Recurrent Neural Network (RNN) models for sequential input of the threads. We experimented with different attributes of the model, such as the number of neural layers and the number of neurons per layer. These models were trained to classify each thread as cyberbullying thread or not. 80% of the dataset was used for training purposes while the remaining 20% was used for testing purposes. Each model was trained and tested for 100 epochs and the model with best F-score was chosen for the results. The results of our experiments are provided in the next section.

IV. Results

As already mentioned (in Section Training and testing), we conducted experiments using LSTM and RNN models with various different configurations for each model. The configurations included the number of neuron layers (single or double layer) and the number of neurons per layer (64, 128 or 256 neurons per layer). We considered the overall F-score of classification as the result parameter for each model. The results achieved for different configurations are shown in table 1 for LSTM model and table 2 for RNN model.

TABLE I. Results for configurations of LSTM

Configuration wise F-score for LSTM		
Layers	***Neurons per layer***	***F-score (%)***
1 layer	64 neurons	68.292
	128 neurons	70.440
	256 neurons	67.143
2 layers	64 neurons	68.831
	128 neurons	67.175
	256 neurons	67.153

TABLE II. Results for configurations of RNN

Configuration wise F-score for RNN		
Layers	***Neurons per layer***	***F-score (%)***
1 layer	64 neurons	70.807
	128 neurons	71.084
	256 neurons	67.879
2 layers	64 neurons	72.000
	128 neurons	69.333
	256 neurons	67.516

The best result was shown by RNN model having a double layer consisting of 64 neurons per layer, showing an F-score of 72%. We have provided the confusion matrix for this configuration in table 3. It gave a total recall of 80.59% with a precision of 65.06%.

TABLE III. Confusion Matrix of Best Model

		Prediction	
		Positive	***Negative***
Actual	***Positive***	54	29
	Negative	13	216

V. Conclusion And Future Work

This study aims to improve the efficiency of cyberbullying detection by considering the context of the discussion. We proposed the method of adjusting sentiments sequentially according to the discussion context flow for this purpose. With this approach, we were able to get significant results using Recurrent Neural Network for sequential input of the threads. We believe that our work contributes to the pursuit of exploiting discussion thread context for cyberbullying detection.

There are many potential extensions of our work. Some of the significant ones according to us are as follows:

- We can train our model on very large datasets to validate our approach.
- We can combine more features such as emotional analysis, sarcasm detection, etc. with our approach to improve accuracy.
- This approach can be combined with various different knowledge-based classification approaches to reinforce the detection system.
- Multi-lingual support can be added to our current approach by using multi-lingual sentiment analysis model.

References

[1] Patchin, J. W., & Hinduja, S. (2006). Bullies Move Beyond the Schoolyard: A Preliminary Look at Cyberbullying. Youth Violence and Juvenile Justice, 4(2), 148–169.

[2] Pew Research Center, December 2022, "Teens and Cyberbullying 2022" https://www.pewresearch.org/internet/2022/12/15/teens-and-cyberbullying-2022/

[3] Petrov, C., 50 alarming cyberbullying statistics to know in 2023, Techjury, p. 1 (2023) https://techjury.net/blog/cyberbullying-statistics/

[4] Department of Mathematics, Indian Institute Of Technology Madras , Chennai, "Percentage of deaths caused by cyberbullying" (2023)

[5] Ditch the Label - UK Study, The annual bullying survey 2017 | ditch the label, p. 26

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:44:30 UTC from IEEE Xplore. Restrictions apply.

[6] Ziyi Li, Junpei Kawamoto, Yaokai Feng, and Kouichi Sakurai. 2016. Cyberbullying detection using parent-child relationship between comments. In Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services (iiWAS '16). Association for Computing Machinery, New York, NY, USA, 325–334

[7] Yin, Dawei & Xue, Zhenzhen & Hong, Liangjie & Davison, Brian & Edwards, April & Edwards, Lynne. (2009). Detection of harassment on Web 2.0.

[8] A. Mody, S. Shah, R. Pimple and N. Shekokar, "Identification of Potential Cyber Bullying Tweets using Hybrid Approach in Sentiment Analysis," 2018 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), Msyuru, India, 2018, pp. 878-881, doi: 10.1109/ICEECCOT43722.2018.9001476.

[9] Shi Yuan, Junjie Wu, Lihong Wang and Qing Wang, "A hybrid method for multi-class sentiment analysis of micro-blogs," 2016 13th International Conference on Service Systems and Service Management (ICSSSM), Kunming, China, 2016, pp. 1-6, doi: 10.1109/ICSSSM.2016.7538628.

[10] Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 258–266, Marseille, France. European Language Resources Association.

[11] Bozyigit, Alican & Utku, Semih & Nasibov, Efendi. (2021). Cyberbullying Detection: Utilizing Social Media Features. Expert Systems with Applications. 179. 115001. 10.1016/j.eswa.2021.115001.

[12] K. Reynolds, A. Kontostathis and L. Edwards, "Using Machine Learning to Detect Cyberbullying," 2011 10th International Conference on Machine Learning and Applications and Workshops, Honolulu, HI, USA, 2011, pp. 241-244, doi: 10.1109/ICMLA.2011.152.

[13] Smith, Phillip & Lee, Mark. (2012). A CCG-based Approach to Fine-Grained Sentiment Analysis. AAAI Spring Symposium - Technical Report. 3-16.

[14] Saif, Hassan & Alani, Harith. (2012). Semantic Sentiment Analysis of Twitter. The Semantic Web--ISWC 2012. 7649. 508-524. 10.1007/978-3-642-35176-1_32.

[15] Francesco Barbieri, Jose Camacho-Collados, Francesco Ronzano, Luis Espinosa-Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. 2018. SemEval 2018 Task 2: Multilingual Emoji Prediction. In Proceedings of the 12th International Workshop on Semantic Evaluation, pages 24–33, New Orleans, Louisiana. Association for Computational Linguistics.

[16] R. Pawar and R. R. Raje, "Multilingual Cyberbullying Detection System," 2019 IEEE International Conference on Electro Information Technology (EIT), Brookings, SD, USA, 2019, pp. 040-044, doi: 10.1109/EIT.2019.8833846.

[17] Dadvar, Maral & de Jong, Franciska & Ordelman, Roeland & Trieschnigg, Dolf. (2012). Improved Cyberbullying Detection Using Gender Information. Cognitive Processing - COGN PROCESS.

[18] Balakrishnan, Vimala & Khan, Shahzaib & Arabnia, Hamid. (2020). Improving Cyberbullying Detection using Twitter Users' Psychological Features and Machine Learning. Computers & Security. 90. 101710. 10.1016/j.cose.2019.101710.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 05:44:30 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2023.3317426

Visual-Motion-Interaction Guided Pedestrian Intention Prediction Framework

[1]Neha Sharma, *Member, IEEE,* [2]Chhavi Dhiman, *Member, IEEE and* [3]S. Indu, *Senior Member, IEEE*

***Abstract*—The capability to comprehend the intention of pedestrians on the road is one of the most crucial skills that the current autonomous vehicles (AVs) are striving for, to become fully autonomous. In recent years, multimodal methods have gained traction employing trajectory, appearance, context, etc. for predicting pedestrian crossing intention. However, most existing research works still lag rich feature representational ability in a multimodal scenario, restricting their performance. Moreover, less emphasis is put on pedestrian interactions with the surroundings for predicting short-term pedestrian intention in a challenging ego-centric vision. To address these challenges, an efficient Visual Motion Interaction (VMI) guided intention prediction framework has been proposed. This framework comprises of Visual Encoder (VE), Motion Encoder (ME) and Interaction Encoder (IE) to capture rich multimodal features of the pedestrian and its interactions with the surroundings, followed by temporal attention and adaptive fusion module to integrate these multimodal features efficiently. The proposed framework outperforms several SOTAs on benchmark datasets: PIE/JAAD with Accuracy, AUC, F1-score, Precision and Recall as 0.92/0.89, 0.91/0.90, 0.87/0.81, 0.86/0.79, 0.88/0.83 respectively. Furthermore, extensive experiments are carried out to investigate different fusion architectures and design parameters of all encoders. The proposed VMI framework predicts pedestrian crossing intention 2.5 sec ahead of the crossing event. Code is available at: https://github.com/neha013/VMI.git**

***Index Terms*-- Pedestrians, Intention Prediction, Autonomous Vehicles**

I. Introduction

As to the Global Status Report on Road Safety 2018, vehicle crashes cause numerous annual deaths, particularly impacting vulnerable road users[1]. Pedestrians, lacking protective gear, face high vulnerability and substantial injury risk in collisions. Consequently, the growing advancement of Autonomous Vehicle (AV) technology is being explored to enhance road safety and convenience for all users. AV technology holds the potential to reduce accidents attributed to human errors like fatigue, misperception, and inattention. Leading automotive manufacturers and tech giants like BMW, Tesla, and Google are actively advancing AV technology in this pursuit.

Predicting pedestrians' road-crossing decisions is pivotal for achieving a reliable driverless experience through AVs. Initial studies emphasized pedestrian dynamics to anticipate crossing intent[2]. Yet, analysing merely the trajectory proves inadequate for understanding underlying intentions[3]. Beyond trajectory, various factors impact pedestrian road-crossing decisions. These factors fall into three primary modalities: *pedestrian-specific* (encompassing pose, appearance, etc.), *context-specific* (involving scene infrastructure and social interaction with co-pedestrians), and *hybrid* modality encompassing comprehensive human cognitive aspects while observing a pedestrian on the road[3].

Inspired by how human drivers interact with pedestrians and make decisions intuitively, recent endeavours[4], [5] aim to decipher pedestrians' crossing intentions by analysing a varied combination of pedestrian-specific and context-specific features. Nonetheless, dealing with such diverse modalities necessitates an efficient multimodal fusion framework that can capture adequate discriminatory features for classification. Moreover, interpreting pedestrian interactions with the surrounding environment is highly challenging in a dynamic ego-centric setting. Quite a few approaches[6], [7] could exploit the social interaction features for short-term intention prediction, so far. Therefore, to address these issues, the proposed multi-modality fusion framework for pedestrian intention prediction focuses on a holistic understanding of the pedestrian and the surrounding environment. This work utilises appearance, context, pose, motion dynamics of pedestrians, social interactions and ego-vehicle speed to extract discriminant and relevant pedestrian feature representations to improve upon SOTA approaches. The main contributions of this work are as follows:

i) A novel multimodal fusion framework to anticipate the short-term intention of pedestrians by combining visual, motion and interaction encoders i.e., VE, ME and IE, is defined to capture a rich multimodal representation of pedestrian appearance, dynamics and interactions in space.

ii) A novel Multi-Head Attention-based Adjacency Matrix ($MHA - AdjMat$) is presented, as given in Algorithm 1, to learn efficient interaction features using Graph Convolution Network (GCN).

iii) An extensive ablation study is carried out to investigate the effects of time-to-event (TTE), Observation sequence length, multimodal fusion architectures and varying design parameters on the performance of the proposed approach.

iv) The experimental observations validate that the proposed framework outperforms other SOTA models on benchmark datasets: JAAD and PIE with pedestrian crossing intention prediction 2.5 sec ahead of the crossing event.

The organization of the paper is as follows: Section II explores the current state-of-the-art and related works: Section III describes the proposed framework; Section IV highlights the significant experimental analysis and lastly, Section V outlines the relevant contributions of the proposed work and its future scope.

First N. Sharma, Research Scholar, Department of Electronics and Communication and Engineering, Delhi Technological University (DTU), Delhi, India-110042 (e-mail: nehashrm013@gmail.com). Second C. Dhiman, Assistant Professor, Department of Electronics and Communication Engineering, Delhi, India-110042 (e-mail: chhavi.dhiman@dtu.ac.in). Third S. Indu, Professor, Department of Electronics and Communication Engineering, Delhi, India-110042 (e-mail: s.indu@dtu.ac.in)

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 08:30:04 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2023.3317426

II. Related Works

The increase in the efficiency and generalisation capability of deep learning techniques has led to a huge surge of data-driven models for pedestrian intention prediction in recent years. The models differ in the type of features [2], [8] they utilise, the fusion approach employed [4], [5] and the way social interactions are modelled [6], [9]. In this section, a brief review of such related works has been given as follows:

A. Multimodal features

Within existing literature, various features have been employed to alleviate the cognitive load of employed intelligent frameworks. The predominant feature for predicting pedestrian intent has been trajectory or historical motion data, evident in numerous studies [10]. Nonetheless, relying solely on trajectory proves inadequate when no historical data exists or when the trajectory is abrupt [3]. Combining pose key-point information with trajectory has shown promise in advancing intention prediction [8]. Visual appearance features also offer significant cues regarding pedestrian intent and future actions [4], [5], [11]. Recent pioneering research emphasizes the importance of visual context features in comprehending the pedestrian's traffic environment [5], [6]. Furthermore, in dynamic scenes, incorporating ego-vehicle motion information aids in assessing target pedestrians' relative motion with respect to onboard cameras [4], [7].

B. Multimodal architectures

Quite a lot of architectures [4], [5], [11] are proposed hitherto, that endeavour to fuse multi-source inputs optimally for efficient and accurate pedestrian crossing prediction. SF-GRU [4] fused local context, appearance, bounding box, pose and ego-vehicle speed hierarchically using GRU as the encoder. On similar lines, Yang et al.[5], proposed fusion of two different channels for visual: local context and global context and non-visual: features-pose, location and ego-speed respectively. Nonetheless, these approaches were restricted in a sense since they do not take into consideration the impact of human social conduct and interactions with the surrounding environment that plays an inevitable role in assessing a pedestrian's short-term intention. Moreover, these works lack rich feature representations of distinct pedestrian modalities and efficient integration of these modalities for enhanced classification performance.

C. Modelling interactions

Modelling subtle nuances of interactions among pedestrians in a dynamic traffic scene, influencing their crossing intention is pivotal in mimicking human-like subconscious decision-making in AVs and ADAS systems. The inherent randomness and dynamic nature of these interactions both in space and time, pose challenges for learning models. Existing methods [2], [12] often limit utilising pedestrian interaction features in long-term intention prediction. Only a few works[6], [9] could employ interaction information for short-term crossing intention provided the high order of randomness in ego-centric videos. Several pioneering works have employed recurrent networks(RNNs)[2], graph convolutional networks(GCNs)[7], [13], etc. and achieved some success. A Category Interaction Module in Biped[6] is developed, to model pedestrians' interactions with various road traffic elements. Pedestrian Graph+.[7] infers spatiotemporal relationships autonomously through network learning. In another seminal work, Yao et al.[9] designed a human visual learning-inspired Attention Relation Network for deeper traffic scene comprehension. However, such evolved multimodal architectures that seamlessly integrate diverse modalities, enhancing both the learning efficacy of the model and the intention prediction performance, remain limited.

Different from the above approaches, this work proposes a novel holistic framework to integrate social interaction features, trajectory, appearance, context and ego speed adaptively. Furthermore, feature extraction module CBAM[14] in the Visual Encoder, HRNet[15] based pose extraction module in the Motion Encoder and GCN employing novel Multi-Head

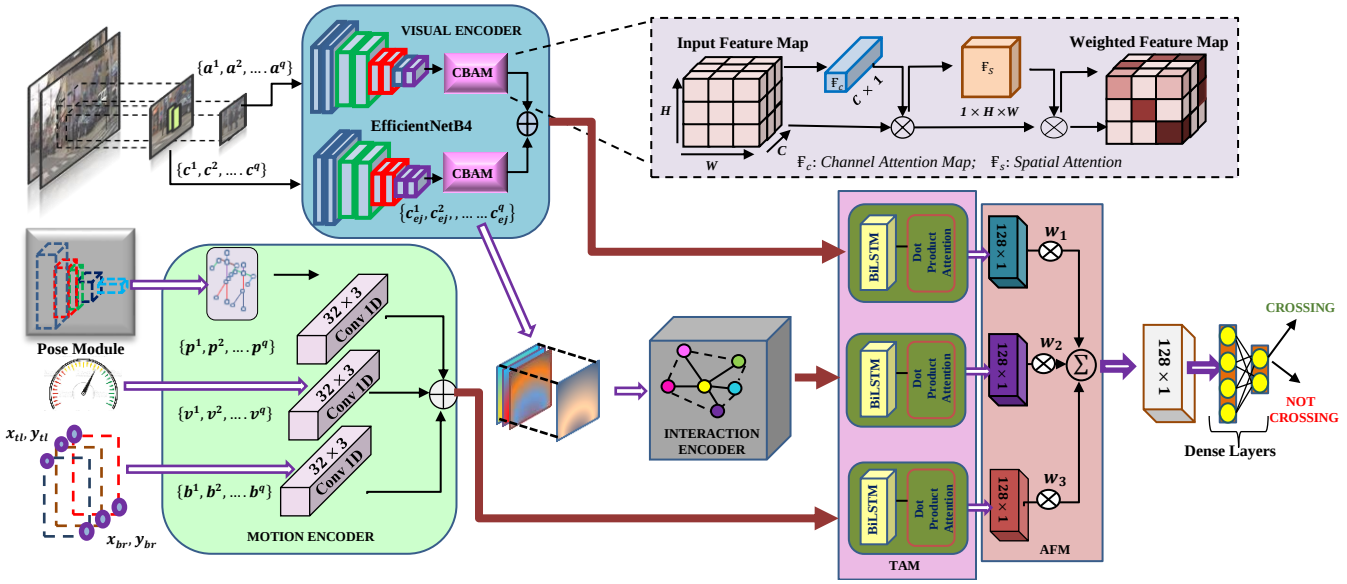


Fig. 1. Illustration of the Visual-Motion-Interaction (V-M-I) guided framework for pedestrian intention prediction integrating multimodal features from a traffic scene, including visual dynamics $A_j = \{a_j^1, a_j^2, a_j^3, \dots \dots a_j^q\}, C_j = \{c_j^1, c_j^2, c_j^3, \dots \dots c_j^q\}$ and non-visual motion dynamics $P_j = \{p_j^1, p_j^2, p_j^3, \dots \dots p_j^q\}, B_j = \{b_j^1, b_j^2, b_j^3, \dots \dots b_j^q\}$, $S_j = \{s_j^1, s_j^2, s_j^3, \dots \dots s_j^q\}$. This framework adeptly combines five essential components: Visual Encoder (VE), Motion Encoder (ME), Interaction Encoder (IE), Temporal Attention Module (TAM), and Adaptive Fusion Module (AFM).

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 08:30:04 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2023.3317426

Attention-based Adjacency Matrix ($MHA-AdjMat$) to extract social interaction features in the Interactio Encoder have enhanced features for superior performance.

III. Proposed Framework

The traffic scene environment is dynamically varying and with onboard cameras, not just the traffic scene but the relative size and distance of the objects/pedestrian to the ego-vehicle in motion is also changing continuously. Unlike human drivers who can decipher non-verbal cues of the surrounding traffic environment and take decisions accordingly, AVs lack this inherent capability. This complex scenario motivates us to utilise multimodal information involving visual, motion and contextual features from the dynamic traffic scene to make AVs mimic human's cognitive ability to anticipate pedestrians' intentions on the road. In this work, the pedestrian crossing intention prediction can be formulated as a binary classification task wherein the motive is to find the probability of a pedestrian 'j' intention to cross or not, $ב_j \in (0,1)$, provided the past observations of visual, dynamic and interaction information of the pedestrian and the ego-vehicle speed for $'q'$ time steps. The proposed implementation of the Visual-Motion-Interaction (V-M-I) guided framework for intention prediction of pedestrians is described in Fig. 1. This architecture employs multimodal features extracted from a traffic scene, involving target pedestrian visual appearance and non-visual dynamic features. Furthermore, interaction features are extracted from the surrounding context, since pedestrian interactions with co-pedestrians on the road also play a pivotal role in influencing crossing behaviour. This architecture is comprised of the following essential components:

i) **Visual Encoder (VE)**: This block encodes pedestrians' appearance and the contextual cues of surroundings to learn visual scene semantics for the target pedestrian.

ii) **Motion Encoder (ME):** This block entails pedestrian pose, trajectory and speed of the ego-vehicle to model the motion dynamics of the pedestrian in a scene.

iii) **Interaction Encoder (IE)**: It leverages the potential of graph convolution networks to infer the relationship of a pedestrian with others in space for a given time frame.

iv) **Temporal Attention Module (TAM):** This penultimate module extracts the temporal relationship of the feature maps of the visual, motion and interaction encoder. This is followed by attention to assigning weights to the most significant parts of the temporal sequence.

v) **Adaptive Fusion Module (AFM):** It adaptively fuses the three stream TAM features of visual, motion and interaction encoders optimising the framework for multimodal inputs.

A. Visual Encoder (VE)

The VE encodes the visual features of the pedestrian and its surroundings as described below:

Appearance – The sequence of RGB images of a traffic scene captures the variations of pedestrians' appearance temporally [5], [11]. The visual appearance features $A_j = \{a_j^1, a_j^2, a_j^3, \dots \dots a_j^q\}$ of the pedestrian $'j' \in (1, m)$ for past observed $'q'$ time steps are captured using image portions of the size of the bounding box consisting of the pedestrian in a scene.

Context – The local contextual information of the target pedestrian depicts its relationship with the dynamic traffic scene elements in its surrounding [3], [13]. The surrounding local environment features $C_j = \{c_j^1, c_j^2, c_j^3, \dots \dots c_j^q\}$ are extracted using a larger image portion size to include the immediate contextual details. This is achieved by extending the dimension by at least twice the size of the pedestrian $'j'$ bounding box and masking it to include only the surrounding details.

In the realm of generating image representations, transfer learning has emerged as a formidable approach, leveraging pre-trained models to extract meaningful features. Inspired by the insights derived from such transfer learning paradigms[16]–[18], the proposed framework endeavours to balance the trade-off between the model's performance and complexity in terms of size and number of parameters of the model by utilising the EfficientNetB4[19] model. This model has 82.9% Top-1 Accuracy on ImageNet[20], while being 17.47x smaller and having 18x fewer parameters than best existing ConvNet[21] so far. Both the appearance and the local context features are processed in parallel through EfficientNetB4 pre-trained on ImageNet. This is followed by Convolutional Block Attention Module (CBAM)[14] to emphasize relevant features across both channel and spatial dimensions as depicted in Fig. 1. Let the appearance and context features after processing via EfficientNetB4 $A_e = \{a_{ej}^1, a_{ej}^2, a_{ej}^3, \dots \dots a_{ej}^q\}$ and $C_e = \{c_{ej}^1, c_{ej}^2, c_{ej}^3, \dots \dots c_{ej}^q\}$ of dimension $C \times H \times W$ where C, H and W denotes the number of channels, height and width dimension of the feature space. These processed features are then passed through an average pooling mechanism to reduce feature dimensions. The parallel branches are finally concatenated to give a modified visual feature representation as shown in equations (1)- (3) as follows:

$$A_e{}' = \mathbb{F}_c(A_e) \otimes A_e;\ A_e{}'' = \mathbb{F}_s(A_e{}') \otimes A_e{}' \quad (1)$$

$$C_e{}' = \mathbb{F}_c(C_e) \otimes C_e :\ C_e{}'' = \mathbb{F}_s(C_e{}') \otimes C_e{}' \quad (2)$$

$$A_e"\oplus C_e" = \{A_e"C_e"_j^1, A_e"C_e"_j^2, \dots \dots A_e"C_e"_j^q\}\ \forall\, j(1,n) \quad (3)$$

where $\mathbb{F}_c \in \mathbb{R}^{C\times1\times1}$ and $\mathbb{F}_s \in \mathbb{R}^{1\times H\times W}$ corresponds to 1-D channel attention and 2-D spatial attention map respectively. The procedure of mapping attention both spatially and channel-wise via CBAM is represented in Fig.1. whereby the darker the colour, the higher the weight assigned to the feature. The CBAM output then undergoes global average pooling to reduce the dimensions of the feature vector for further computations.

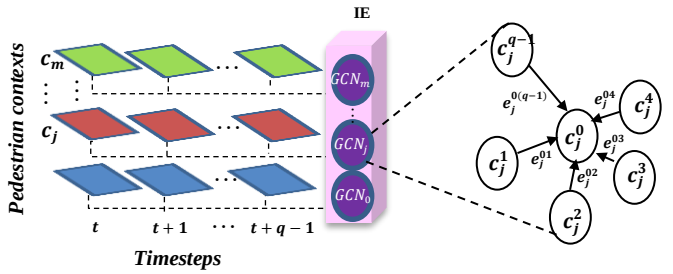
B. Motion Encoder (ME)

The representation of these features is described as follows:

Pose: In this work, HRNet[15] is employed to generate a set of 17 pose key points per pedestrian sample using a bounding box sequence. This network has achieved ~76% mAP on the MS COCO[22] dataset while maintaining higher resolution representation throughout the estimation process. This network surpasses the performance of several SOTA pose estimation modules[23] utilised in quite a few existing works[4], [5]. The key points obtained are in the form of x and y coordinates resulting in a vector of 34 values per pedestrian sample. These values undergo normalization and then concatenation into a feature vector for further processing for $'q'$ past observations, and are represented as $P_j = \{p_j^1, p_j^2, p_j^3, \dots \dots p_j^q\}$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 08:30:04 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2023.3317426



q: Number of timesteps
m: Number of pedestrian samples
c_j^t: j^{th} pedestrian context feature map at t^{th} timestep
e_j^{tu}: Degree of association between t and u timesteps

Fig. 2. Visualisation of Interaction Encoder that models temporal relationship of the target pedestrian's context at different time steps. The architecture uses a graph G: $\{N_j, E_j\}$, with nodes N_j as $'t^{th}'$ time-step encoded feature maps of the target pedestrian $'j'$ and edges E_j representing interconnections among these maps in a traffic scene.

Trajectory: The location of a pedestrian $'j'$ in a 2D coordinate space is provided with top-left $\{x_{tl}, y_{tl}\}$ and bottom-right $\{x_{br}, y_{br}\}$ coordinate points represented as $B_j = \{b_j^1, b_j^2, b_j^3, \dots \dots b_j^q\}$

Speed: This consists of speed value measurements of the ego-vehicle in km/h given as - $S_j = \{s_j^1, s_j^2, s_j^3, \dots \dots s_j^q\}$

Motivated by the scientific findings[24] about the low computational complexity and faster training time of Conv 1-D over Conv 2-D for one-dimensional data, the ME processes pose, trajectory and speed through a Conv 1-D parallelly with 32 filters of kernel size, 3 and stride, 1. These transformed features are then combined and passed to the TAM.

C. Interaction Encoder (IE)

Pedestrians' interactions with traffic scene elements dynamically impact the road crossing intention [3], [7]. Inspired by some of the seminal works [7], [13], the proposed approach leverages graph convolutional networks (GCNs)[25] to model the temporal relationship of pedestrian interactions across consecutive frames. In the proposed architecture, the nodes of the graph G: $\{N_j, E_j\}$ represents $'t^{th}'$time-step encoded feature map of the target pedestrian $'j'$ context and edges E_j represents the associations existing among these feature maps corresponding to different time steps in a traffic scene as shown in Fig. 2. These inter- associations within a graph are reflected via an adjacency matrix $R_{q \times q}$ computed as shown in Algorithm 1, where 'q' is the total time steps considered per sample. For each j^{th} pedestrian sample, received context feature map is applied with linear transformation followed by a reshape operation to extract Query, Key and Value matrices, as described in Step 4, Algorithm 1. Step 5 computes multi-head attention $H_{q \times b \times d}$ by calculating individual attention heads h^i and then concatenating all $'b'$ attention heads. Following this, $H_{q \times b \times d}$ is then mapped with number of timesteps again as $H_{q \times r}$. Hence, the Multi-Head Attention-based Adjacency Matrix[26], $R_{q \times q}$, derived at Step 6, ensures extraction of the implicit contextual details of the dynamic interactions of the target pedestrian at consecutive timesteps. Further, GCN is applied to the graph representation derived above with nodes $'N_j'$ and multi-head attention-based adjacency matrix $'R_j'$

Algorithm 1:
Multi-Head Attention based Adjacency Matrix Computation
$(MHA - AdjMat)$

1: Input:
a) Context Visual encoder Feature Map $X \in \{A_{ej}''\}$ corresponding to pedestrian j:
b) Size of attention vector:$'r'$,
c) Number of feature maps 'c'
d) Number of attention heads :$'b'$

2: Output: Adjacency Matrix R of dimension $q \times q$

3: Initialisation:
Initialize an empty matrix R with zeros.

4: **Feature Map Processing:**
for each feature map X corresponding to pedestrian j:
- Extract $Query\ (Q), Key(K)\ and\ Value(V)$ matrices by applying ***Dense*** operation to feature map X as:
$$Q_{q\times r}, K_{q\times r}, V_{q\times r} \leftarrow Dense(X)$$
Where *Dense* performs the linear transformation with respective weight matrices W_q, W_k and W_v
- ***Reshape*** Q, K and V such that the last dimension becomes a multiple of number of heads 'b' where $b \times d = r$
$Q_{q\times b\times d}$, $K_{q\times b\times d}, V_{q\times b\times d} = Reshape(Q_{q\times r}, K_{q\times r}, V_{q\times r})$

5: **Mulit-Head Attention Computation:**
- ***for*** i in $range\ (1, b)$:
Compute scaled dot product attention for i_{th} head:
$$h^i \leftarrow softmax\left(\frac{Q^i . K^i}{\sqrt{r}}\right) V^i$$
- ***Concatenate*** all heads to get the final head vector as
$H \leftarrow concat[h^1 : h^2 : h^3 : \dots \dots \dots \dots \dots : h^b]$
- ***Reshape*** H: $H_{q\times b\times d}$ to get $H_{q\times r}$

6: ***Linear transformation***:
Perform a linear transformation to obtain the Adjacency Matrix $R_{q\times q}$ by applying the *Dense* operation to H:$R_{q\times q} \leftarrow Dense(H)$ with the weight matrix W_r.

return $R_{q\times q}$

depicting the relationship between nodes. The adjacency matrix requires a normalisation step to curb issues of vanishing or exploding gradients as network training may be sensitive to the range of scale of values. The normalization step includes generating the Laplacian matrix as represented below:

$$\check{R} = D^{-\frac{1}{2}}(R + I)D^{-\frac{1}{2}} \quad (4)$$

wherein self-loops are also considered to incorporate the target node's features in the propagation, by adding Identity matrix I to the adjacency matrix R and D is the diagonal degree matrix containing row-wise summation of $(D + I)$ matrix. Here, the spectral propagation rule for the convolution of the graph features is employed [27]. Where, GCN layer is modelled by applying a linear transformation operation as a scalar product of the adjacency matrix and the hidden feature, followed by a GeLU[28] activation function for the next layer $'l'$ as shown in equation (5)

$$G(l+1) = GeLU\left(\left(D^{-\frac{1}{2}}(R + I)D^{-\frac{1}{2}}\right)X_l W_l\right) \quad (5)$$

where $X(l)$ is the previously hidden layer output of the GCN convolution layer. For $l = 0$, $X(0) = C_e{}''$.

D. Temporal Attention Module (TAM)

In this module, the BiLSTM layer provides enhanced temporal representations of the input sequence by leveraging learning through bi-directional layers. The following attention module weighs the most relevant parts of the feature map $'\Upsilon'$along the temporal dimension. The attention weight vector $\hat{\alpha}_k$ for k^{th} branch of the architecture is given in equations (6)-(7) as follows:

$$s_k = softmax(\tanh(W\Upsilon_k + bias)) \quad (6)$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 08:30:04 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2023.3317426

TABLE I
Training specifications of the proposed framework

Training Parameters	JAAD	PIE
Optimizer	ADAM	ADAM
Learning Rate	2×10^{-5}	5×10^{-5}
# Epochs	60	70
L2 Regularization	0.0001	0.0001
Loss Function	Binary Cross Entropy	Binary Cross Entropy
Batch Size	8	16

$$\hat{\alpha}_k = \sum_d \alpha_k \Upsilon_k \quad (7)$$

Where d is the dimension along which attention vector computation is carried out. In this case, it is the output vector length of the BiLSTM layer. The number of hidden units of the BiLSTM layer employed is 64.

E. Adaptive Fusion Module (AFM)

The generated features from the VE, IE and ME, followed by the TAM, have varying impacts on the pedestrian intention prediction, therefore, an adaptive fusion is introduced to accordingly weigh all the encoded features. These hidden representations of the proposed fusion have been accumulated in a vector representation, $'E'$ as shown in the equation below:

$$E = \sum_{k=1}^{3} w_k \hat{\alpha}_k \quad (8)$$

where $\hat{\alpha}_k$ represents the encoded hidden states of the prior TAM and w_k are the trainable weights with HeNormal initialisation[29]. This is followed by dense layers of 64 and 8 units and a final activation function to give *'crossing'* or *'not crossing'* predictions.

IV. EXPERIMENTAL RESULTS

The proposed framework is evaluated using two publicly available benchmark datasets, namely JAAD[30] and PIE[31]. The JAAD dataset is having 346 video clips with a total duration of 240 hours recorded at 30 frames per sec(fps). Each clip ranges from $5-15$ *sec* with a resolution of 1920×1080 and 1280×720. The bounding boxes and tracking ids are provided for each pedestrian. The driver's action is implicitly encoded as vehicle speed for training the model. The PIE dataset consists of 1842 pedestrian tracks having longer sequences and increased pedestrian samples with annotations in comparison to JAAD. The dataset configuration follows the training/validation/test split as recommended in [32] for JAAD [31] for PIE. The performance of the proposed framework has been compared against the following SOTA methods:

- ***Pie_traj*** [31]**:** employs an RNN-based encoder-decoder to extract vital information regarding pedestrian appearance and surrounding context and pedestrian dynamics.
- ***Stacked Fusion GRU*** (SF-GRU) [4]**:** hierarchically stacks GRU encoder that fuses features of high relevance like appearance at the beginning and others like ego-speed at the last.
- ***Feature Fusion and Spatio-temporal Attention*** (FFSTA)[5]**:** employs a hybrid architecture to fuse features from visual and non-visual branches. The local and global context are combined in the visual branch while other features like pose, bounding box and vehicle speed are hierarchically fused in the non-visual branch
- ***BiPed*** [6]**:** proposes a bifold encoding approach for the individual and shared representation encompassing trajectory, grid locations, ego speed and global context.
- ***Multi-model Atrous***(MMA)[11]**:** processes diverse input modalities through the visual encoding and dynamics encoding branch followed by corresponding attention modules and subsequent fusion for joint representation.
- ***Intent and Action***(IA)[9]**:** presents a human visual learning-inspired Attention Relation Network ensuring a deeper understanding of the scene semantics along with other pedestrian-specific features.
- ***Pedestrian Graph+***(PG+)[7]**:** employs a fully convolutional graph-based neural network that takes context, human pose key points and ego-speed as input.

TABLE II
Comparison of existing SOTAs with the proposed method on the PIE and JAAD dataset

Methods	Year	PIE/JAAD				
		Acc	AUC	F1	Prec	Rec
PIE_tra[31]	2019	0.79	-	0.87	-	-
SF-GRU[4]	2020	0.87/0.84	0.85/0.80	0.78/0.62	0.74/0.54	0.64/0.73
FFSTA[5]	2022	0.85/0.83	0.83/0.82	0.71/0.63	0.69/0.51	0.72/0.81
BiPed[26]	2020	0.91	0.90	0.85	0.82	0.88
MMA[11]	2020	0.89/0.89	0.88/0.88	0.81/0.81	0.77/0.77	0.85/**0.85**
IA[9]	2021	0.84/0.87	0.90/0.70	**0.88/0.92**	**0.96** /0.66	0.81
PG+[7]	2022	0.89/0.86	0.90/0.88	0.81/0.65	0.83/0.58	0.79/0.75
Ours	-	**0.92/0.89**	**0.91/0.90**	0.87/0.81	0.86/**0.79**	**0.88**/0.83

The experimental settings for implementation are outlined in Table 1. The computation of the pose key points, appearance and contextual features is done before training. Data augmentation involves horizonalt image flipping for balancing crossing/non-crossing samples to mitigate prediction bias.

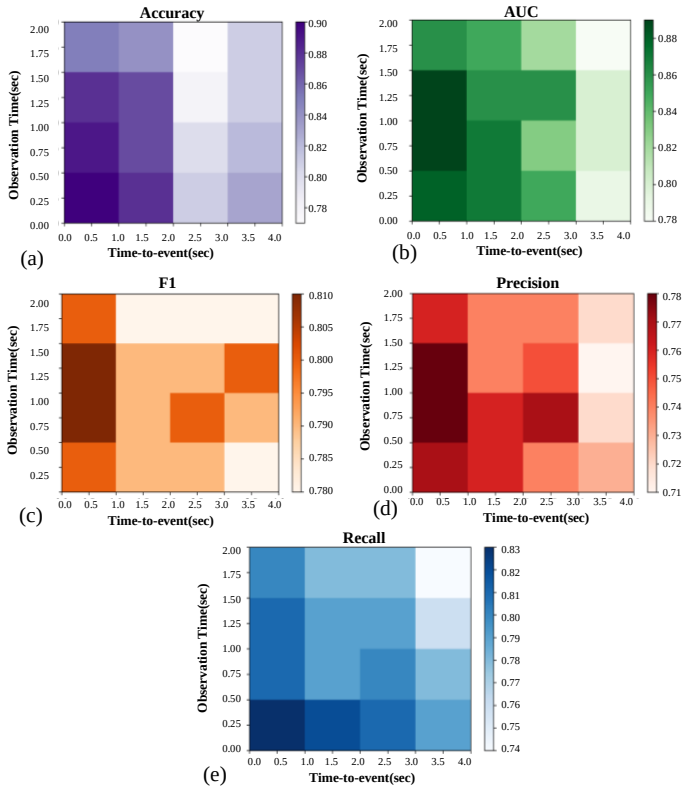


Fig. 3. Impact of Time to Event (TTE) and Observation Sequence Lengths (OSL) on Crossing Intention Prediction, evaluated in terms of (a) Accuracy, (b) AUC, (c) F1 Score, (d) Precision, and (e) Recall metrics. Given the dynamic nature of traffic scenes, TTE intervals ranging from 0 to 4 seconds, sampled at 0.2-second intervals, are combined with a 2-second OSL, also sampled at 0.2-second intervals, to investigate prediction performance.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 08:30:04 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2023.3317426

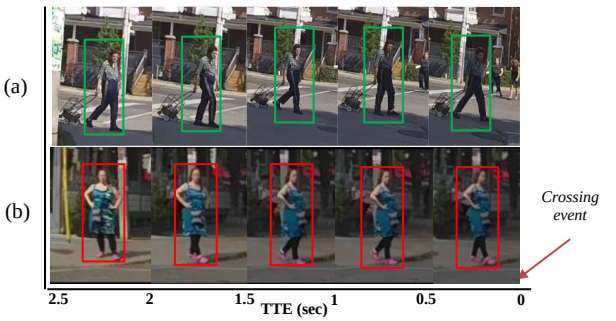


Fig. 4. Qualitative samples of pedestrian short-term intention prediction using our proposed approach, illustrate its effectiveness in anticipating intent 2.5 seconds before the crossing event. (a) An instance of correctly predicted intention (b) An instance of failure: pedestrian is crossing the road; however, the proposed approach predicted the pedestrian's intention inaccurately as non-crossing due to a sudden change in direction. The 'Green' bounding box indicates crossing intention, while the 'Red' bounding box signifies non-crossing intention with respect to the Time-to-Event (TTE).

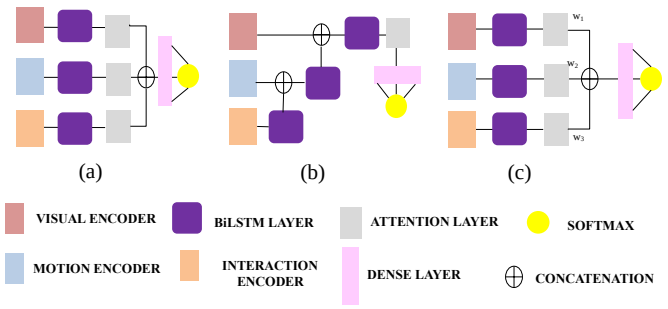


Fig. 5. Diagrammatic Overview of various Pedestrian Intention Prediction Fusion Architectures. (a) Parallel Fusion (PF), (b) Hierarchical Fusion (HF), and (c) Adaptive Fusion (AF) approaches, showcasing distinct strategies for enhancing pedestrian intention prediction.

The performance of the proposed approach in comparison to the state of the arts is represented in Table II. It is evident from the table that the proposed VMI-guided framework performs better in terms of all evaluation metrics against SOTA models, except F1 score and Precision on the PIE dataset and F1 score and Recall on the JAAD dataset. The intent and action [9] perform a little better (∼1%) in the F1 score and also have a significant leap in precision approximately by 10 %. This drop-in precision in our proposed work is compensated with an overall improvement in accuracy (∼ 8.7%), AUC score (∼1%) and recall (∼8.7%) on the PIE dataset. Similarly, the considerable jump in performance metrics like accuracy (∼8.64%), AUC(∼2.3%) and precision(∼28.6%) overcomes the decline in F1 score (∼12%) in the JAAD dataset against [9] by our proposed approach. The optimal observation length of 0.5 *sec* with 2.5 *sec* time-to-event has been found empirically which will be discussed later in the ablation study.

A. Ablation Study:

Impact of choosing different Time to event (TTE) and Observation Sequence Lengths (OSL): The high variability in traffic scene dynamics with every passing second has a considerable impact on the crossing intention prediction of the pedestrian. OSL also impacts the prediction process. The longer the OSL, the greater the information acquired. However, it may also add some insignificant details that may result in erroneous prediction results. In this section, different TTE points are considered on the timeline of the crossing/not crossing event ranging from 0-4 sec, sampled at every 0.2 sec and 2 secs long

TABLE III
Ablation study on different fusion architectures

Fusion	Encoder	PIE/JAAD Acc	AUC	F1	Prec	Rec
HF	GRU	0.83/0.84	0.78/0.81	0.73/0.71	0.76/0.71	0.71/0.71
	LSTM	0.85/0.84	0.80/0.83	0.74/0.74	0.78/0.72	0.70/0.76
	BiLSTM	0.85/0.85	0.80/0.85	0.77/0.75	0.78/0.74	0.75/0.77
	BiLSTM+A	0.86/0.85	0.85/0.86	0.77/0.76	0.79/0.74	0.76/0.78
PF	GRU	0.85/0.83	0.83/0.82	0.78/0.72	0.77/0.69	0.78/0.75
	LSTM	0.86/0.83	0.83/0.82	0.78/0.74	0.77/0.72	0.79/0.77
	BiLSTM	0.86/0.85	0.81/0.84	0.78/0.76	0.80/0.74	0.77/0.79
	BiLSTM+A	0.88/0.86	0.84/0.87	0.80/0.77	0.81/0.75	0.79/0.80
AF	GRU	0.87/0.85	0.85/0.81	0.80/0.77	0.79/0.75	0.81/0.79
	LSTM	0.87/0.85	0.84/0.84	0.82/0.78	0.84/0.76	0.80/0.81
	BiLSTM	0.90/0.87	0.90/0.86	0.84/0.79	0.85/0.77	0.83/0.82
	BiLSTM+A	**0.92/0.90**	**0.91/0.89**	**0.87/0.81**	**0.85/0.78**	**0.88/0.83**

A : Attention

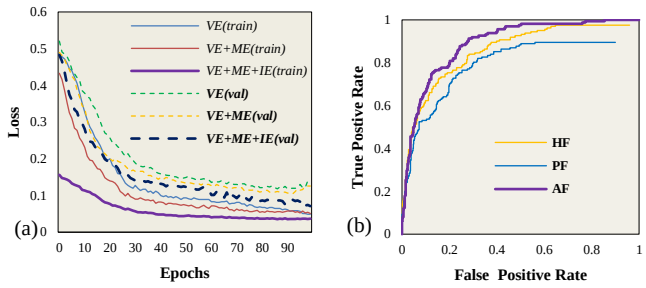


Fig. 6. (a) Analysis of Training and Validation Loss across different encoder combinations, namely VE, VE+ME, and VE+ME+IE. Validation losses are annotated as (---), while (—) signifies training losses. The best performing combination (*VE+ME+IE*) is highlighted in **bold,** showcasing the model's effective convergence during training process. (b) ROC Curves illustrate the Performance of Parallel Fusion (PF), Hierarchical Fusion (HF), and Proposed Adaptive Fusion (AF) architectures. The most optimal ROC curve of the

OSL is sampled at every 0.2 sec as shown in Fig. 3. It is observed that there is a gradual decline in the overall performance of the proposed approach as the TTE increases and vice versa as the TTE approaches close to 0 since the intention of the pedestrians becomes evident by that time. There is also a slight gain in accuracy, AUC and precision up to 1.5 *sec* of OSL but at the expense of a decrease in recall. Moreover, a more balanced metric F1 score shows no huge variations (∼3%) with the increase in OSL. Notably, the proposed approach works satisfactorily with all parameters ∼(> 80%) even 3 *secs* to TTE as opposed to SF-GRU[4] where F1 score, precision and recall drop to $\sim 60-65\%$ at 3 sec to TTE. Furthermore, the overall performance of the proposed approach does not drop below $12-13\%$ with respect to the highest attained metric value even at the end TTE and OSL values. Nonetheless, in SF-GRU[4], the TTE and OSL variations cause a drop as high as 33.3%. This proves the robustness of our approach against varying OSL and TTE. Fig. 4.(a) illustrates a qualitative sample prediction on the PIE dataset 2.5 sec ahead of the crossing event, while Fig. 4 (b) depicts a failure case by a sudden pedestrian direction change.

Relevance of different Encoders: This section studies the training performance of VE, ME and IE both individually and jointly. The analysis of loss curves in Fig. 6(a) indicates effective training for all combinations, with a steeper validation loss curve observed when utilising all three encoders VE+ME+IE, showcasing the model's enhanced generalisation capability achieved through integration of multiple modalities.

Fusion Strategies: Different fusion architectures for multimodal features have been employed in this ablation study inspired by the works [4], [5] as depicted in Fig. 5. Table III shows the performance of a few of the fusion architectures on

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 08:30:04 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2023.3317426

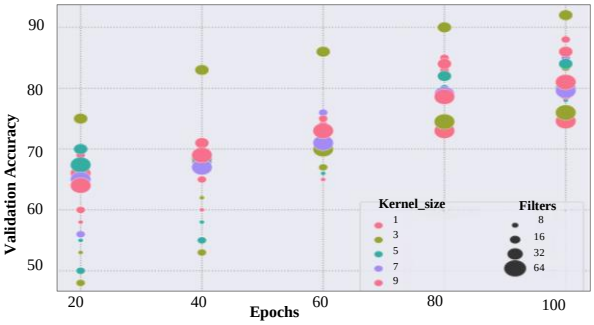


Fig. 7. Hyperparameter Analysis of the Convolutional Layer in the Motion Encoder. The impact of varying kernel sizes and filter counts on the Motion Encoder's performance in extracting significant motion features from pedestrian pose, trajectory, and ego speed. The colour and size of the bubble denotes size of kernel and number of filters respectively.

Ablation study on Interaction Encoder Components

Graph Node Features	Adjacency Matrix	PIE/JAAD Acc	AUC	F1	Prec	Rec
Pose	Uni	0.81/0.81	0.84/0.82	0.78/0.75	0.77/0.72	0.78/0.77
	DI	0.82/0.82	0.83/0.80	0.79/0.74	0.78/0.71	0.79/0.76
	SHA	0.84/0.84	0.83/0.82	0.80/0.76	0.80/0.73	0.79/0.78
	MHA	0.84/0.83	0.82/0.81	0.78/0.73	0.79/0.71	0.77/0.75
Trajectory	Uni	0.82/0.83	0.83/0.84	0.79/0.74	0.79/0.71	0.79/0.78
	DI	0.81/0.81	0.83/0.81	0.79/0.72	0.80/0.69	0.77/0.76
	SHA	0.82/0.85	0.87/0.86	0.80/0.76	0.81/0.73	0.78/0.79
	MHA	0.86/0.83	0.83/0.82	0.82/0.75	0.83/0.72	0.80/0.77
Appearance	Uni	0.83/0.86	0.85/0.84	0.79/0.76	0.81/0.73	0.77/0.78
	DI	0.83/0.85	0.84/0.83	0.80/0.75	0.82/0.74	0.78/0.76
	SHA	0.85/0.86	0.85/0.85	0.81/0.77	0.81/0.74	0.80/0.79
	MHA	0.87/0.85	0.87/0.83	0.79/0.77	0.79/0.76	0.79/0.77
Context	Uni	0.88/0.86	0.85/0.83	0.84/0.76	0.82/0.74	0.86/0.78
	DI	0.87/0.85	0.87/0.83	0.83/0.74	0.80/0.72	0.85/0.76
	SHA	0.91/0.87	0.88/0.85	0.84/0.78	0.81/0.75	0.87/0.81
	MHA	**0.92/0.89**	**0.90/0.89**	**0.87/0.81**	**0.85/0.79**	**0.88/0.84**

the PIE and JAAD datasets. It is observed that the proposed adaptive fusion with the BiLSTM + Attention layer achieves the best performance overall. The ROC Curve visualisations for different fusion approaches on the PIE dataset are shown in Fig. 6. (b) This study also explores four encoder variations (GRU, LSTM, BiLSTM, BiLSTM + Attention), highlighting the apparent increasing performance trend of the BiLSTM followed by the attention layer across different fusion architectures.

ME Ablation: The ME utilises the Conv-1D layer to extract significant motion features from the pedestrian's pose, trajectory and ego speed. The impact of the number of filters and the kernel size of the Conv-1D layer on the training performance is shown in Fig. 7, where kernel size of 3 and 32 filters yield the highest validation accuracy of ~92 % on the PIE dataset after 100 epochs. This is attributed to the smaller kernel size's ability to generate fine-grained features, though it lacks neighbouring context. Conversely, larger kernel sizes like 5,7, or 9, may overlook intricate details. The plot indicates that a higher number of filters are required for the ME to capture the complex and stochastic pedestrian motion. A significant drop of ~ 10-20% in accuracy is observed if the number of filters is increased or decreased from 32 during training for a fixed kernel size. However, the number of parameters also grows as we increase the kernel size or the number of filters. Therefore, a Conv-1D layer with 32 filters and a kernel size of 3 is employed in this work to minimise the trade-off between accuracy and the number of parameters.

IE Ablation: The graph node features and the adjacency matrix significantly enhance the graph's efficiency to model subtle interactions of the target pedestrians within a traffic scene. Table IV shows the impact of diverse node features on

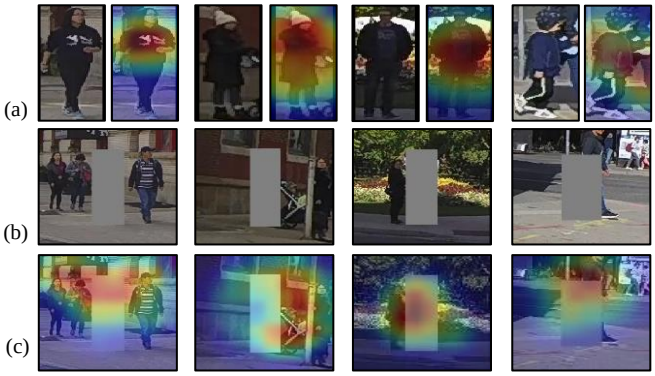


Fig. 8. GradCAM [33] visualizations illustrating the significant areas of focus captured by the Appearance and Context Encoders. (a) Pedestrian cropped image with highlighted regions of interest, (b) Pedestrian image integrated with surrounding context, and (c) contextual cues highlighted by using GradCAM output that contribute to the pedestrian intent prediction.

performance metrics. It is observed that the context features secure the highest metric values, followed by appearance. The ablation study also explores distinct adjacency matrix computation methods: *Uniform*(Uni) which involves random initialisation of the adjacency matrix with values ranging from 0 to 1, *Distance-Inverse*(DI) that employs inverse Euclidean distance between pedestrians[12], *Single-head Attention*(SHA) that utilises a self-attention module inspired by the work [2] and lastly the adjacency computation using *Multi-head attention* (MHA) described in this work[26] as reported in Table IV. Notably, the MHA method outperforms other adjacency matrix computation methods in terms of classification performance. This is attributed to SHA's limited human behavioural learning range and Uni's inability to adequately represent intrinsic human interaction randomness. The DI method too restricts the interaction to sheer separation based limiting its performance.

VE Ablation: The GradCAM [33] visualisations for the CBAM feature maps in the VE are obtained by superimposing on the real cropped images enhancing model explainability as depicted in Fig. 8. It is inferred from both the PIE and JAAD dataset images that the torso region is given higher weights in comparison to other body parts. This affirms that the head, gaze, shoulder and posture forming the torso region of the human body, play an integral part in assessing road-crossing intentions. Additionally, the GradCAM visualisations for context show that the VE pays higher attention towards co-pedestrians in the vicinity that are influencing the pedestrian's crossing intent.

V. Conclusion

This paper presents a multimodal pedestrian intention prediction framework that adaptively fuses rich visual, motion and interaction features. The application of attention in spatial, channel and temporal dimensions and the novel ($MHA - AdjMat$) based GCN in IE have enriched pedestrian features. This is validated by the superior performance of the proposed framework over other SOTA models in predicting pedestrian crossing intent on JAAD and PIE datasets, 2.5 seconds ahead of the crossing event.

While promising, the study has limitations in effectively capturing high-frequency temporal relations with GCNs. Rapid and fine-grained actions may challenge the framework's ability

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 08:30:04 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2023.3317426

to perceive swift changes in temporal relationships among pedestrian actions. Future work could focus on extending the framework for long-term intention prediction by exploring more robust discriminant features.

REFERENCES

[1] "GLOBAL STATUS REPORT ON ROAD SAFETY 2018 SUMMARY," 2018.

[2] Y. Zhou *et al.*, "Social graph convolutional LSTM for pedestrian trajectory prediction," *IET Intell. Transp. Syst.*, vol. 15, no. 3, pp. 396–405, Mar. 2021.

[3] N. Sharma, C. Dhiman, and S. Indu, "Pedestrian Intention Prediction for Autonomous Vehicles: A Comprehensive Survey," *Neurocomputing*, vol. 508, pp. 120–152, Oct. 2022.

[4] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Pedestrian Action Anticipation using Contextual Feature Fusion in Stacked RNNs," *30th Br. Mach. Vis. Conf. 2019, BMVC 2019*, pp. 1–13, May 2020.

[5] D. Yang, H. Zhang, E. Yurtsever, K. A. Redmill, and Ü. Özguner, "Predicting Pedestrian Crossing Intention with Feature Fusion and Spatio-Temporal Attention," *IEEE Trans. Intell. Veh.*, vol. 7, no. 2, pp. 221–230, Apr. 2022.

[6] A. Rasouli, M. Rohani, and J. Luo, "Bifold and Semantic Reasoning for Pedestrian Behavior Prediction," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 15580–15590.

[7] P. R. G. Cadena, Y. Qian, C. Wang, and M. Yang, "Pedestrian Graph +: A Fast Pedestrian Crossing Prediction Model Based on Graph Convolutional Networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 21050–21061, 2022.

[8] Z. Fang and A. M. López, "Is the Pedestrian going to Cross? Answering by 2D Pose Estimation," *IEEE Intell. Veh. Symp. Proc.*, vol. 2018-June, pp. 1271–1276, 2018.

[9] Y. Yao, E. Atkins, M. Johnson-Roberson, R. Vasudevan, and X. Du, "Coupling Intent and Action for Pedestrian Crossing Behavior Prediction," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, Aug. 2021, pp. 1238–1244.

[10] P. Zhang, W. Ouyang, P. Zhang, J. Xue, and N. Zheng, "SR-LSTM: State Refinement for LSTM towards Pedestrian Trajectory Prediction," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 12077–12086, Mar. 2019.

[11] A. Rasouli, T. Yau, M. Rohani, and J. Luo, "Multi-Modal Hybrid Architecture for Pedestrian Action Prediction," Nov. 2020.

[12] J. Lian, W. Ren, L. Li, Y. Zhou, and B. Zhou, "PTP-STGCN: Pedestrian Trajectory Prediction Based on a Spatio-temporal Graph Convolutional Neural Network," *Appl. Intell.*, vol. 53, no. 3, pp. 2862–2878, May 2023.

[13] T. Chen, R. Tian, Z. Ding, and C. Engineering, "Visual Reasoning using Graph Convolutional Networks for Predicting Pedestrian Crossing Intention," in *ICCV Workshop*, 2021, pp. 3103–3109.

[14] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proceedings of the European Conference On Computer Vision (ECCV)*, pp. 3-19. 2018.

[15] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep High-Resolution Representation Learning for Human Pose Estimation."

[16] S. Lu, Z. Zhu, J. M. Gorriz, S. H. Wang, and Y. D. Zhang, "NAGNN: Classification of COVID-19 based on neighboring aware representation from deep graph neural network," *Int. J. Intell. Syst.*, vol. 37, no. 2, pp. 1572–1598, 2022.

[17] Y. Zhang *et al.*, "Deep learning in food category recognition," *Inf. Fusion*, vol. 98, 2023, doi: 10.1016/j.inffus.2023.101859.

[18] S. Lu, S. H. Wang, and Y. D. Zhang, "Detection of abnormal brain in MRI via improved AlexNet and ELM optimized by chaotic bat algorithm," *Neural Computing and Applications*, vol. 33, no. 17. pp. 10799–10811, 2021.

[19] M. Tan and Q. V Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," May 2019.

[20] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 248–255.

[21] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," pp. 11966–11976, 2022.

[22] T. Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8693 LNCS, no. PART 5, pp. 740–755, 2014.

[23] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, 2021.

[24] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1D convolutional neural networks and applications: A survey," *Mech. Syst. Signal Process.*, vol. 151, pp. 1–20, 2021.

[25] B. Xu and H. Yin, "Graph Convolutional Networks in Feature Space for Image Deblurring and Super-resolution," *Proc. Int. Jt. Conf. Neural Networks*, vol. 2021-July, 2021.

[26] Z. Tao, C. Ouyang, Y. Liu, T. Chung, and Y. Cao, "Multi-head attention graph convolutional network model: End-to-end entity and relation joint extraction based on multi-head attention graph convolutional network," *CAAI Trans. Intell. Technol.*, 2022.

[27] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," *5th Int. Conf. Learn. Represent. ICLR 2017 - Conf. Track Proc.*, Sep. 2016.GAUSSIAN ERROR LINEAR UNITS (GELUS)."

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2015 Inter, pp. 1026–1034, 2015.

[30] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Are They Going to Cross? A Benchmark Dataset and Baseline for Pedestrian Crosswalk Behavior," *Proc. - 2017 IEEE Int. Conf. Comput. Vis. Work. ICCVW 2017*, vol. 2018-Janua, pp. 206–213, 2017.

[31] A. Rasouli, I. Kotseruba, T. Kunic, and J. Tsotsos, "PIE: A Large-Scale Dataset and Models for Pedestrian Intention Estimation and Trajectory Prediction," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, vol. 2019-Octob, pp. 6261–6270.

[32] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "It's not all about size: On the role of data properties in pedestrian detection," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11129 LNCS, pp. 210–225, 2019.

[33] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Oct. 2016.

Neha Sharma received her B.Tech. degree from Indira Gandhi Delhi Technical University for Women (IGDTUW), Delhi, India, in 2017 and M.Tech. degree from Delhi Technological University (DTU), Delhi, India, in 2019. She is currently pursuing a P.hD. from DTU, Delhi, India. Her research interest includes Computer Vision and Pedestrian intention analysis. She has been awarded the "Research Excellence Award" by DTU in 2021 and 2023.

Chhavi Dhiman received her B.Tech. degree from Indira Gandhi Delhi Technical University for Women (IGDTUW), Delhi, India, in 2011, and M.Tech. and PhD degrees from Delhi Technological University (DTU), Delhi, India, in 2014 and 2019 respectively. She is currently working as Assistant Professor, in the Department of Electronics and Communication Engineering, DTU, Delhi, India. Her research interest includes Pattern Recognition, Video Analytics, Human Action Recognition, Visual Captioning, Face Anti-Spoofing. She is a reviewer of various Journals/Transactions of IEEE, Elsevier, Springer. She has been awarded the "Premier Research Award" by DTU in 2021.

S.Indu received her B.Tech degree from TKM College of Engineering, University of Kerala, India, MTech degree from College of Engineering Trivandrum, University of Kerala, and PhD from University of Delhi, India. She joined the Electronics and Communication Engineering Department of Delhi College of Engineering in 1999. Currently, she is working as Dean (student Welfare) and Professor of ECE Department of Delhi Technological University (DTU) Her research interests include Computer Vision, Sensor Networks and Image Processing. She also received Commendable Research Award in 2018 and 2019 from DTU.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on November 03,2023 at 08:30:04 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.